ENGINEERING HANDBOOK

Model Refactoring

for Software Engineers

Profiling, pruning, distillation, quantization,
and mobile deployment

VISION / LANGUAGE / AUDIO
IMAGE DIFFUSION / VIDEO DIFFUSION

COMPLETE MANUSCRIPT · PDF EDITION

24 September 2026
Includes technical glossary + HDemucs experiment plan

ABOUT THIS EDITION

Scope and evidence

Complete manuscript | PDF edition | 24 September 2026

Review status: This edition preserves the working manuscript, adds a technical glossary with internal lookup links, and includes an HDemucs experiment plan. The included mechanism tests and synthetic learning experiment are not production-model validation. The HDemucs appendix is proposed work, not an executed benchmark.

From an existing trained model to a smaller, measurable, deployable system.

This handbook treats a model change as an engineering change: define the problem, inspect the implementation, make one controlled modification, test the result, and preserve a rollback path.

The worked examples use Python and PyTorch. The methods apply to experiments with vision, language, audio, image diffusion, and video diffusion. Android deployment is one target, not the definition of success. A desktop GPU experiment and an on-device product use the same evidence discipline but different resource budgets.

> **Reading this PDF**
>
> The contents page and PDF bookmarks navigate the full manuscript. The first occurrence of a glossary term in each chapter links to its definition in Appendix C. Source markers link to the reference directory. Companion code and recorded experiment outputs remain in the separately supplied ZIP package.

PDF production preserves the manuscript's evidence distinctions. Formatting and source inspection are not substitutes for independent reproduction, production-model evaluation, or device benchmarks.

NAVIGATION

Contents

How to use this handbook

Read Chapters 1 through 3 before changing a checkpoint. Chapters 4 through 11 form a catalog of transformations. Chapter 12 addresses image and video diffusion. Chapters 13 and 14 address export, runtimes, and device integration. Chapter 15 reports the executed labs. Chapters 16 and 17 provide experiment plans and review criteria. Appendix A explains how to run the companion code. Appendix C provides the technical glossary used by the linked terminology throughout the handbook. Appendix D applies the method to an HDemucs experiment plan.

You should be comfortable reading Python, using version control, writing automated tests, and interpreting performance measurements. You do not need to derive backpropagation to use the examples. You do need to understand tensor shapes and distinguish a trained parameter from a value computed during inference.

Model refactoring is not always behavior-preserving. Removing an unused graph node can preserve the intended function. Deleting a learned block usually changes it. In the latter case, your regression test is a distribution of outcomes, not just equality on a few sample inputs. Keep these two kinds of change separate in code review. ONNX Runtime explicitly distinguishes semantics-preserving graph rewrites from other optimizations, including optional approximations. **[R15]**

Evidence labels

Documented means the behavior is described in official documentation or the referenced implementation. **Reported** means an external study describes a result; this handbook has not reproduced that study. **Executed** means the companion code was run in the preparation environment and its outputs are included. **Proposed** means the procedure is an engineering recommendation or a template that still needs testing in your environment.

References such as **[R14]** identify external primary sources in the reference directory. References [E01], [E02], and [E03] identify the included experiment records. A citation establishes the specific claim associated with it. It does not validate an entire application or guarantee a particular compression ratio.

What has and has not been tested

The executed lab uses **RefactorNet 1.0**, an original, small residual classifier trained on generated numerical data. It demonstrates block ablation, physical feed-forward pruning, recovery training, checkpoint reconstruction, and CPU timing. A second executed example demonstrates chunked evaluation of one causal convolution. A third tests convolution-channel pruning, attention-head pruning, tokenwise feed-forward chunking, image tiling, and a deliberately incorrect normalization change. The supplied records include software versions and raw measurements. [E01, E02, E03]

No Android device, NPU, pretrained diffusion checkpoint, production language or speech model, or task-specific media dataset was used in those experiments. The ONNX export and quantization template was not executed because the required packages were unavailable and package installation could not reach the network. Device deployment and production-model recipes are therefore proposed work, not benchmark results. This is a practical draft with executed mechanism tests, not an independently reviewed claim of universal compression performance.

The laboratory runtime was PyTorch 2.10.0+cpu on Python 3.13.5. Online documentation can describe newer releases. Version-specific PyTorch references are used where available, and mutable documentation is identified by its access date. Do not assume that a newly installed package is equivalent to the tested environment. PyTorch also cautions that reproducibility is not guaranteed across releases and platforms. **[R10]**

1. Establish a reproducible baseline

Problem

You have a trained model that appears too large or slow. The temptation is to delete layers immediately. Without a reproducible baseline, you cannot distinguish a useful optimization from a different input, a changed preprocessing step, a lucky timing result, or an export error.

Your first deliverable is not a smaller model. It is an experiment that another engineer can run and interpret.

Define the product contract

Describe the task in observable terms. A detector may need to identify small objects. An image generator may need to preserve requested composition. A video generator may need to maintain identity and motion across a specified duration. A speech model may need intelligibility and speaker consistency. These are proposed product requirements, not properties established by a parameter count.

Define the operating envelope separately: supported devices, maximum input duration, supported resolutions, sequence lengths, channels, and sample rates, whether network access is allowed, and whether the job may pause. A batch application and a live translator have different deadlines even when they use the same model.

Avoid a requirement such as "the model must fit in 900 MB." Specify what the number measures: model download, mapped parameter data, peak process footprint, or one activation buffer. Android's managed heap limit varies by device and is not a universal allowance for native ML tensors. **[R21]**

Freeze the full implementation

A model is more than its checkpoint. Record the architecture source, configuration, tokenizer or phonemizer, preprocessing, checkpoint digest, inference arguments, dependency versions, and decoding settings. Two applications can load identical weights and produce different output because they interpret the input differently.

Use a manifest that your program can populate rather than a handwritten description that drifts from the implementation:

JSON

```
{
  "architecture": "RefactorNet",
  "architecture_version": "1.0",
  "checkpoint_sha256": "computed_by_the_run",
  "input_contract": {"shape": [1, 16], "dtype": "float32"},
  "runtime": "PyTorch 2.10.0+cpu",
  "device_class": "Linux x86_64 CPU",
  "threads": 1,
  "seed": 20260924
}
```

This is a schema example. The companion run writes the actual digest and configuration into `evidence/run/report.json`; it does not retain the illustrative digest above. [E01]

Establish two baselines

The **quality baseline** defines how the unmodified checkpoint behaves on the evaluation suite. The **performance baseline** defines how the deployed implementation behaves with a fixed workload and device configuration. Keep both, because a valid PyTorch model is not automatically an equivalent exported model.

Measure initialization separately from warm inference. Report the input shape, batch size, number of warmup calls, number of measured calls, thread configuration, and whether accelerator work was synchronized. For an audio stage, define real-time factor as processing seconds divided by input-audio seconds. For a multi-stage application, measure the complete job as well as individual stages.

Do not replace measured latency with multiply-accumulate counts. NetAdapt specifically studies adaptation using direct platform measurements because indirect quantities such as operation count do not reliably determine latency and energy. Its mobile vision experiments are evidence for a hardware-aware procedure, not a performance forecast for speech models. **[R08]**

Define acceptance before selecting candidates

Write down which quality regressions are unacceptable, which performance objective matters, and which inputs must continue to work. Your thresholds are product decisions. There is no universal acceptable percentage loss for detection, semantic correctness, image fidelity, motion consistency, or voice identity.

For example, a release gate can require that a critical object category retain recall, that an image-editing mask remain respected, and that a chosen memory measurement fall. State how each condition will be measured and who reviews ambiguous outputs. Do not adopt numerical thresholds from an unrelated application.

Verification: rerun the unchanged baseline before beginning surgery. Confirm that the quality result is stable enough to detect the changes you care about. Treat inconsistent runs as an experiment problem first.

Decision and rollback: preserve the original checkpoint, environment manifest, reference outputs, and preprocessing. A candidate is not the new baseline until it passes the agreed evaluation.

2. Find the actual bottleneck

Problem

A model file contains the parameters needed to run the network, but the process also needs intermediate values, persistent state, runtime workspaces, and application buffers. A parameter reduction can leave the real memory bottleneck untouched.

Separate the quantities

For a dense tensor, its logical payload is the product of its dimensions multiplied by bytes per element. This arithmetic excludes allocator padding, views that share storage, compressed representations, and workspace. For example:

EXAMPLE

```
Shape: [1, 2000, 1024]
FP16 element size: 2 bytes
Payload: 1 * 2000 * 1024 * 2 = 4,096,000 bytes
Decimal MB: 4.096
Binary MiB: 3.90625
```

The distinction between decimal MB and binary MiB matters when comparing logs. State the unit and conversion rather than silently mixing them.

A useful conceptual accounting model is:

EXAMPLE

```
At time t:
working memory = resident parameters + live intermediate tensors
               + persistent state + temporary workspace
               + runtime and application allocations

Peak memory = maximum of that quantity over the execution
```

This is an accounting model, not a formula for adding RSS to parameter bytes. RSS already includes resident allocations. Some buffers share pages or storage; count those relationships explicitly when you need an accurate total. Android's PSS accounts proportionally for shared memory and is a different measurement from a tensor's logical payload. [R21, R22]

Tensor lifetimes matter

Suppose a layer reads a 300 MB tensor and writes a separate 250 MB tensor. Those buffers can overlap in lifetime. A following layer reads the 250 MB tensor and writes 350 MB. With no in-place operation, no additional state, and no scratch space, the two steps need approximately 550 MB and 600 MB respectively.

The required storage is not automatically 900 MB, because earlier buffers may be reused. It is also not automatically 350 MB, because input and output often coexist. This is a worked lifetime example, not a measurement of a particular runtime.

Deleting some sequential layers may reduce parameters and execution time while barely changing the largest live allocation. Removing a long-lived skip connection or shortening the sequence may affect the peak more directly. Confirm this against the actual execution schedule, rather than adding per-layer output sizes.

Profile the model without turning the profiler into the bottleneck

PyTorch's profiler can record operator time, input shapes, and tensor allocation activity. Its documentation warns that collecting shapes and stacks adds overhead and can retain references. Use profiling to locate expensive work, then benchmark without the profiler. [R09]

PYTHON

```python
# Integration pattern. Run against your own model and inputs.
model.eval()
with torch.inference_mode():
    with torch.profiler.profile(
        activities=[torch.profiler.ProfilerActivity.CPU],
        record_shapes=True,
        profile_memory=True,
    ) as prof:
        model(example_input)
print(prof.key_averages().table(
    sort_by="self_cpu_time_total", row_limit=15
))
```

The lab includes a metadata-only forward-hook pass that writes module output shapes. Those records locate large outputs; they are not a peak-activation measurement. Hooks on modules can miss functional operations, and two reported outputs may share underlying storage. The lab deliberately does not publish a fabricated peak from their sum. [E01]

Distinguish training from inference

Use `model.eval()` to select evaluation behavior where modules have separate training behavior. Disable gradient recording for inference as appropriate. PyTorch's `inference_mode` removes additional autograd overhead but does not itself call `eval()`. Do not apply it to gradient-based importance estimation or to tools that need autograd to analyze the graph. [R12]

Activation checkpointing is principally a training tradeoff: it saves selected activations and recomputes others for backward. It is not an automatic remedy for a forward-only mobile memory problem. [R36]

Choose the next experiment

If parameters dominate, investigate quantization, width reduction, low-rank factors, or fewer blocks. If long feature sequences dominate, investigate chunking, state management, or temporal resolution. If runtime copies dominate, inspect layout changes and execution-provider boundaries. If application buffers dominate, fix the surrounding code before modifying the model.

Verification: identify the exact operation or lifetime responsible for the chosen bottleneck. **Decision:** proceed only when a proposed change can plausibly affect that bottleneck. **Rollback:** preserve the trace and input that exposed it so the same condition can be reproduced.

3. Build an evaluation suite that can reject a bad change

Problem

A model can perform well on a demonstration and still fail on a small object, a long prompt, an unusual camera movement, a quiet speaker, or a rare output class. Your suite must expose the failures relevant to the experiment rather than reward the examples you happened to inspect.

Use data for distinct purposes

The **training set** updates parameters. A **calibration subset** estimates quantization ranges or importance statistics. The **validation set** guides choices such as which block to remove. The **test set** evaluates the selected procedure after those choices are fixed.

Calibration data can come from training data. It does not have to be held-out test data. Any procedure that fits statistics must not quietly fit them on the final test set. This separation is also emphasized in scikit-learn's guidance on data leakage. **[R37]**

A frozen validation set makes development comparisons consistent. It does not remain unbiased after repeated tuning. Reserve a separate final test, record its use, and do not revise the candidate after inspecting it while still calling it unseen.

Where the examples come from

Start with the original model's documented evaluation data when it is accessible and suitable. Add your own legally usable, representative inputs and reviewed expected outcomes. You do not need to reproduce the original training corpus to begin a regression experiment. You do need enough coverage to justify the claim you plan to make.

A classifier needs inputs and class labels. A segmentation model needs masks or another appropriate reference. A generator may use a fixed prompt suite, condition images, masks, random seeds, and human review rather than one uniquely correct output image. A teacher's generated outputs can be recovery targets, but agreement with the teacher does not independently establish quality.

Do not choose a universal sample count. Begin with a small diagnostic suite to catch implementation errors, then expand the held-out evaluation to cover the operating envelope and obtain useful uncertainty estimates. Ten attractive outputs cannot establish a broad quality claim.

Split at the unit that can leak

Do not scatter adjacent frames of the same video across training and testing. Decide whether the claim concerns new clips, new scenes, new subjects, or new source collections, then group the split accordingly. Apply the same reasoning to near-duplicate images, lines from the same speaker, and documents from the same source. This is a proposed dataset-design procedure, not a prescribed split ratio.

For prompt-based generation, separate prompt templates as well as individual strings when template generalization matters. A different noun inserted into a repeatedly tuned prompt is not necessarily a new kind of test.

Choose metrics that describe the task

Model family	Useful evaluation dimensions	Failure an average may conceal
Classification or detection	Correctness, calibration, per-class recall, small-object performance	A rare but important class is lost.
Segmentation or restoration	Boundary quality, structural fidelity, reference agreement	Thin structures disappear or edges shift.
Language	Task success, factual consistency where relevant, long-context behavior	The model passes short prompts but fails structured output.
Speech or audio	Intelligibility, identity, temporal alignment, audible artifacts	Words remain clear while a speaker's identity changes.
Image generation	Prompt following, composition, diversity, visual defects	A model produces pleasing images while ignoring requested relations.
Video generation	Identity, motion, temporal flicker, spatial quality, prompt following	Individual frames look good but the sequence is unstable.

The table is a proposed engineering checklist, not a definition of any named benchmark. VBench is a primary research example of evaluating video generation along distinct dimensions instead of collapsing all behavior into one visual score. ITU-T P.808 is a source for speech listening-study methodology, not a universal speaker-similarity measure. [R43, R33]

Make generative comparisons reproducible

Use the same prompt and conditioning case across baseline and candidate. For closely related generators, reset the random generator for each case and retain the initial latent tensor when the implementation permits. Reusing a mutable generator object without resetting it advances its state, so the same nominal seed recorded once is insufficient. Cross-device and cross-version equality is not guaranteed. [R44, R10]

When changing resolution, the initial noise tensor changes shape. The same integer seed no longer creates a directly aligned pixel-level experiment. Compare task outcomes across a fixed seed list rather than interpreting pixel differences as pure model error.

Record the evaluator's preprocessing. Image resizing and compression can change FID measurements, as demonstrated by the Clean-FID work. Report the metric implementation, feature extractor, reference set, sample count, and preprocessing. Do not attach a published FID label to an informal handful of pictures. **[R45]**

Make evaluation actionable

A record should connect an input to its conditions, candidate, and failure categories:

JSON

```
{
  "case_id": "heldout_video_014",
  "condition_groups": ["camera_pan", "occlusion", "two_subjects"],
  "reference_version": "reviewed_cases_v1",
  "candidate_digest": "computed_by_the_evaluator",
  "checks": ["identity", "motion", "prompt_following", "artifacts"]
}
```

This is a schema example, not an executed video result. Store failed outputs alongside aggregates. An accuracy change from 90% to 89% is one percentage point. State units rather than reporting an ambiguous "1% loss."

Interpret uncertainty

Use paired comparisons where possible. Repeat recovery seeds when a decision depends on a small difference. Estimate uncertainty at the independent sampling unit, such as a video or source group, rather than treating correlated frames as independent evidence.

For human comparisons, conceal candidate identities, randomize presentation order, and ask a specific question. Fidelity, naturalness, identity, and preference are different outcomes. Keep difficult cases in the suite after discovering them, but track that they have become development examples.

Verification: version the examples, split rules, and scoring code. **Decision:** reject critical subgroup regressions even when the mean improves. **Rollback:** retain failure outputs and the configuration that generated them.

4. Remove a residual block

Problem and preconditions

The model contains repeated blocks whose input and output contracts match. You want to remove one to reduce depth. A valid candidate is a block that can be bypassed without breaking tensor dimensions, required state, or the meaning of the next interface.

A residual block of the form `y = x + F(x)` has a natural bypass: return `x`. An encoder stage that also changes time resolution or channel count does not necessarily have that property. A block that produces a decoder cache or a second output cannot safely be replaced by a one-output identity module without an adapter.

LayerDrop is a reported training method designed to make Transformer depth adjustable. Its results do not establish that arbitrary layers in an ordinary pretrained checkpoint can be removed without recovery. **[R06]**

Rank candidates with ablation

Ablation means changing one candidate while holding the rest of the experiment fixed. For each removable block, build a separate model, run the validation set, and record the difference from the unmodified model. The companion lab uses validation cross-entropy as the selection criterion. It does not select a block using final test accuracy. [E01]

PYTHON

```python
# Executed implementation is in lab.py.
student = copy.deepcopy(model)
student.blocks = torch.nn.ModuleList([
    block for i, block in enumerate(student.blocks)
    if i != remove_index
])
```

The code is valid for RefactorNet because every block preserves a 32-value feature dimension and the forward method simply iterates over `blocks`. It is not a general Transformer-pruning function. A production architecture can also require changes to layer counts, cache indices, relative-position modules, serialization metadata, and generation helpers.

Separate ranking from acceptance

A layer can have the smallest loss increase and still be too important to remove. Ranking tells you which experiment to investigate first. Acceptance tells you whether the result meets the product contract.

Do not assume that ablation losses add. Two individually tolerable removals can damage complementary functions. Remove a small set, recover if appropriate, and rerun importance tests on the modified network.

Low residual magnitude, activation similarity, and gradient-based sensitivity can help shortlist candidates. They remain heuristics unless checked against task outcomes. First-order pruning criteria have research support in specific settings, including Molchanov and colleagues' CNN experiments. That support does not make small gradients a universal test of irrelevance. **[R07]**

What the executed lab found

The lab evaluated all six blocks. Removing block 5 gave the lowest validation loss. Some removals actually improved validation loss. That is possible in a finite, imperfectly trained model and is not proof that the removed layer was universally useless. The selected removal slightly reduced final test accuracy before recovery. The complete ablation table appears in Chapter 15. [E01]

This contrast is important: the validation set chooses a candidate; the test set measures whether the selected procedure generalizes. It is normal for the two not to move in exactly the same direction.

Verification and recovery

Run shape and finite-value checks first. Rebuild the optimizer after changing the parameter structure. Save the edited architecture configuration with its weights, then reconstruct it in a clean process and require strict loading. The lab tests exact output equality for that save-and-reload operation, not equality between the original and pruned models. [E01]

Recover the candidate with supervised fine-tuning or distillation. Compare recovery methods under an explicit training budget rather than allowing one candidate substantially more optimization without saying so.

Decision: keep the deletion only when quality and the target-device resource objective pass. **Rollback:** restore the original checkpoint and the original block ordering. Do not attempt to reverse a sequence of undocumented in-place edits.

5. Shrink feed-forward channels without changing the external interface

Problem and preconditions

A block contains a feed-forward network that expands from width `d` to width `h`, applies an elementwise activation, and projects back to `d`. You want to reduce `h` while leaving the residual interface unchanged.

This is a useful first structural edit because it is narrower in scope than changing the network's global hidden width. The worked implementation applies to an ordinary dense `Linear -> GELU -> Linear` path with no normalization across the intermediate feature dimension, no gated parallel path, and no tied parameters. [E01]

Remove the connected dimensions together

For a PyTorch linear layer, weight rows correspond to output features. If you retain hidden indices `K`, the connected changes are:

EXAMPLE

```
Original:
up.weight   [h, d]     up.bias   [h]
down.weight [d, h]     down.bias [d]

After retaining k hidden channels:
up.weight   [k, d]     up.bias   [k]
down.weight [d, k]     down.bias [d]
```

Copy rows `K` from the first weight and bias. Copy columns `K` from the second weight. Preserve the second bias. The lab's `shrink_ffn` function validates indices, creates smaller modules on the original device and dtype, and copies the coupled parameters. [E01]

PYTHON

```python
# Core of the executed transformation; guards are in lab.py.
new.up.weight.copy_(old.up.weight[keep])
new.up.bias.copy_(old.up.bias[keep])
new.down.weight.copy_(old.down.weight[:, keep])
new.down.bias.copy_(old.down.bias)
```

Do not write directly into `.data` to hide shape inconsistencies. Construct the correct modules and preserve their configuration explicitly.

Prove the surgery did what you intended

There are two different comparisons. The structurally pruned FFN should match the original FFN with the removed intermediate channels masked to zero, within numerical tolerance. It does not generally match the unmasked original FFN.

The companion self-test checks the first equivalence. It also checks that the parameter count falls and that duplicate indices are rejected. These are implementation tests. They do not establish acceptable task quality. [E01]

PyTorch's standard pruning utilities demonstrate masking. Removing the pruning reparameterization makes that sparsity permanent in the values; it does not by itself produce smaller dense tensor dimensions. Torch-Pruning addresses structural removal and coupled dependencies. [R01, R02]

Choose channels, then test the choice

The lab uses a simple heuristic: average absolute hidden activation multiplied by the norm of the corresponding output-weight column. It computes this on 256 training examples, not on the final test set. This is original illustrative scoring code, not an implementation claimed to reproduce a published optimal pruning criterion. [E01]

Other candidate-selection strategies can use validation ablation, learned gates, or gradient-based importance. The important engineering property is that the score, data, and retained indices are recorded. Avoid saying "unimportant channels" when you have only measured low values under one criterion.

For a gated FFN, such as one with separate gate and value projections multiplied elementwise, corresponding hidden channels must be removed from both branches and the downstream projection. For convolutions, channel changes can propagate through batch normalization, residual additions, concatenations, and grouped operators. DepGraph formalizes this dependency problem, but custom modules still require inspection and tests. **[R03]**

Verification, acceptance, and rollback

Benchmark widths that the target kernel handles well, not only the mathematically smallest width. The number of parameters changes predictably from the shapes; latency remains a measurement. The lab's width reduction removed substantially more parameters than the observed percentage improvement in warm CPU latency. [E01]

Retain the original indices in the experiment record, reconstruct the student from its new configuration, run recovery, and evaluate subgroup failures. Revert if task quality is unacceptable, if export fails, or if the changed shape does not improve the resource objective that justified the work.

6. Prune convolution channels without breaking the graph

Problem and preconditions

A convolutional stage produces more feature channels than your resource budget permits. Unlike deleting scattered weight values, reducing channel dimensions can create smaller dense tensors for the next stage. The engineering problem is identifying every consumer of those channels.

The executed `ConvPair` example has a deliberately narrow contract:

EXAMPLE

```
Input image [B, 3, H, W]
  -> Conv2d: 3 channels to 12
  -> ReLU
  -> Conv2d: 12 channels to 3
  -> Output image [B, 3, H, W]
```

It has no residual branch, concatenation, normalization, grouped convolution, or shared parameters. This makes the hidden channel axis easy to inspect. It is a mechanism test with random weights, not a trained image model. [E03]

Make one physical edit

Retain a set of hidden channels `K`. Copy the first convolution's output filters and biases at those indices. Copy the corresponding input-channel slices from the second convolution. Keep the second bias unchanged.

PYTHON

```python
# Executed for the exact ConvPair architecture in the companion lab.
new.first.weight.copy_(old.first.weight[keep])
new.first.bias.copy_(old.first.bias[keep])
new.second.weight.copy_(old.second.weight[:, keep])
new.second.bias.copy_(old.second.bias)
```

The full function validates the indices, creates the new modules on the same device and dtype, and preserves evaluation state. Its self-test compares the smaller model with the original after masking the removed hidden activations. Both produce the same result in the recorded test. That does not mean the smaller model equals the unmasked original. [E03]

Expand from a chain to a dependency group

A real image model is usually more connected than the example. Before pruning a channel, draw its consumers or inspect them in the captured graph. A residual addition requires compatible shapes on both branches. A concatenation changes the input offsets of a downstream operator. A normalization layer can own channel-indexed parameters and state. Dependency-aware pruning groups these connected edits rather than treating each module independently. **[R03]**

Consider this proposed inspection:

EXAMPLE

```
A produces channels 0..63
  -> normalization
  -> residual branch
  -> concatenation with B
  -> projection
```

Your edit record should identify the retained channel order, the modified normalization state, the residual adapter if any, and the concatenation offsets seen by the projection. "Pruned 25% of layer A" is not enough to reconstruct the graph.

Torch-Pruning is useful for generating structural dependency groups. Its graph construction uses autograd, so do not wrap that analysis in inference mode. Use the package's documented workflow for the pinned commit, then inspect the proposed group before applying it. A graph-analysis tool does not know your application-specific quality contract. **[R02]**

Grouped operators require a different edit

For grouped convolution, channels are divided into groups rather than fully connected. Input and output channel counts must respect the group configuration. Depthwise convolution is a special grouping pattern, not merely an ordinary convolution with many zero weights. PyTorch documents these constraints for `Conv2d`. **[R47]**

Do not feed arbitrary retained indices into the `ConvPair` function and expect it to support a depthwise layer. Either retain valid groups and update the group metadata, or implement a tested transformation specifically for that operator. Removing only one member of a coupled group can invalidate both the shape and channel mapping.

Normalization deserves the same care. Shrinking the set of values over which a normalizer computes statistics changes the function. The fact that all shapes match does not prove equivalence. The tiling experiment in Chapter 9 makes this issue visible for GroupNorm. [R28, E03]

Replacing convolution is a new model, not a rename

Depthwise separable convolution applies spatial filtering per input channel and then mixes channels with a pointwise projection. MobileNet is a primary example of this architecture. It is an architectural choice with a different parameterization, not a general exact replacement for an arbitrary dense convolution. **[R29]**

Treat a proposed replacement as student design. Specify how it is initialized and trained. Compare it with simpler alternatives such as channel pruning, lower resolution, or a better-supported kernel before committing to a retraining effort.

Verification and decision

Run channel-mapping tests on a small controlled graph first. Then require strict checkpoint reload, finite outputs, task evaluation, and target-backend execution. Inspect whether the new shapes still select an efficient kernel.

Accept when the complete dependency group is valid, quality passes, and the measured bottleneck improves. **Revert** when the modification requires fragile runtime exceptions, loses critical image features, or produces no useful device-level benefit. Preserve the original group and retained indices so the transformation is reproducible.

7. Prune attention heads and distinguish them from model width

Problem

An attention module is expensive, and its configuration includes several numbers that appear interchangeable: model width, head count, head dimension, and key/value head count. They describe different tensor dimensions. Changing one field in a configuration file does not necessarily remove computation.

In a conventional multi-head projection, the internal attention width is:

EXAMPLE

```
internal_width = number_of_heads * head_dimension
```

If you halve the head count but double the dimension of each head, the internal width remains unchanged. The projection matrices may retain the same number of values. That edit is not the same as physically pruning half the heads.

Preserve the public interface

The executed attention lab keeps the external width at 32 and each head dimension at 8. It changes four internal heads to two. The query, key, and value projections shrink from 32 output features to 16; the output projection maps those 16 features back to the original 32-feature interface. [E03]

EXAMPLE

```
Original:
input 32 -> Q/K/V width 32 -> four 8-value heads -> output 32

Pruned:
input 32 -> Q/K/V width 16 -> two 8-value heads -> output 32
```

This preserves the surrounding residual shape without pretending that the unmasked attention function is unchanged.

Remove the corresponding slices

For separate dense Q, K, and V projections, a retained head corresponds to a contiguous group of projected feature indices. Retain those output rows in all three projections and the matching input columns in the output projection.

PYTHON

```python
# Core of the executed SplitSelfAttention transformation.
offsets = torch.arange(old.head_dim, device=keep.device)
indices = (keep[:, None] * old.head_dim + offsets).reshape(-1)

for name in ("q", "k", "v"):
    source = getattr(old, name)
    target = getattr(new, name)
    target.weight.copy_(source.weight[indices])
    target.bias.copy_(source.bias[indices])

new.out.weight.copy_(old.out.weight[:, indices])
new.out.bias.copy_(old.out.bias)
```

The full implementation rejects invalid head indices and tests equivalence to the original with the removed head outputs masked. It has no dropout, rotary encoding, cache, fused QKV storage, or shared key/value heads. Do not substitute it into a pretrained Transformer without adapting those contracts. [E03]

Choose heads by task impact

The paper *Are Sixteen Heads Really Better than One?* studies head pruning and shows that importance differs across attention components in the tasks evaluated. It supports investigating head redundancy, not assuming every attention layer can lose the same fraction of heads. **[R25]**

For your model, begin with validation ablation or another explicitly recorded importance criterion. Re-evaluate the entire task after recovery. A head that appears dispensable on short inputs may matter on longer contexts or different conditioning.

For diffusion, include multiple noise levels and conditioning modes in the evaluation. For a language decoder, include both prompt processing and token generation. These are proposed coverage requirements derived from the execution paths you intend to support.

Do not confuse a better kernel with a different attention rule

An IO-aware implementation can compute the same mathematical attention without storing all of its intermediate matrices at once. FlashAttention is a primary example. Windowed attention instead changes which positions can interact. One is an implementation strategy; the other can change the model's behavior. **[R26]**

PyTorch's scaled-dot-product attention API can select among implementations subject to input and backend support. It is not a promise that a particular Android runtime has the same kernels. It also applies dropout according to the supplied probability, so explicitly use zero dropout for an inference-only call. **[R27]**

Before pruning, test whether a supported memory-efficient attention implementation addresses the actual bottleneck. That experiment may avoid a structural change, but still needs numerical and task checks.

Global width is a larger migration

Changing model width from 1024 to 512 affects embeddings, residual paths, normalizers, attention projections, feed-forward layers, and often output heads. Shared or tied parameters make the dependency wider. Treat it as an architecture migration with a new configuration, not a local edit.

A practical sequence is to try internal FFN width and internal attention width independently before shrinking the public hidden dimension. This is a proposed risk-reduction strategy, not a universal optimal pruning order.

Inspect cache costs separately

For a simple decoder with the same cache layout in every layer, the logical key/value payload is approximately:

EXAMPLE

```
2 * layer_count * batch_size * cached_tokens
  * kv_head_count * head_dimension * bytes_per_value
```

The factor of two accounts for keys and values. This accounting excludes padding, temporary tensors, allocator overhead, and alternative cache layouts. Query-head pruning alone need not reduce the key/value cache in an architecture with shared KV heads. Hugging Face documents distinct cache strategies and their memory/latency tradeoffs. **[R46]**

Verification: test attention masks, position handling, cache construction, and checkpoint reload. **Decision:** accept only measured gains on the relevant inference path. **Rollback:** restore projection slices and cache configuration together.

8. Replace a dense matrix with low-rank factors

Problem and contract

A large dense linear layer contributes heavily to parameter storage or matrix multiplication. You want a smaller representation while preserving its input and output dimensions.

For a layer with weight shape `[m, n]`, replace one mapping from `n` to `m` with two mappings:

EXAMPLE

```
n -> rank r -> m
```

The dense weight contains `m*n` values. The factors contain `r*(m+n)` values. Ignoring a retained output bias, the factors are smaller when:

EXAMPLE

```
r < (m*n) / (m+n)
```

For a 4096-by-4096 matrix and rank 512, the weight count changes from 16,777,216 to 4,194,304. This is arithmetic, not a measured fourfold speedup.

Initialize with SVD

Singular-value decomposition expresses the weight using orthogonal factors and singular values. The companion function uses `torch.linalg.svd` to construct two linear modules from the retained components. The first has no bias; the second retains the original bias. [R11, E01]

PYTHON

```python
# Executed implementation, simplified to show the factorization.
u, s, vh = torch.linalg.svd(layer.weight.float(), full_matrices=False)
first.weight.copy_(vh[:rank].to(layer.weight))
second.weight.copy_((u[:, :rank] * s[:rank]).to(layer.weight))
if layer.bias is not None:
    second.bias.copy_(layer.bias)
```

There must be no extra activation between the two factors if the objective is a low-rank approximation to the original linear operation. Adding a ReLU changes the hypothesis and requires a different test.

Test the implementation before the approximation

At full rank, the reconstructed operation should match the original within numerical tolerance. This validates factor order, transpose conventions, and bias placement. Full-rank reconstruction is not itself a useful compression result.

At a lower rank, output differences are expected. The lab checks the output shape but does not claim task quality for its low-rank candidate. A small weight-reconstruction error is not a guarantee that the errors occur in task-irrelevant directions. Evaluate the actual model and recover it if necessary. [E01]

Rank selection is an experiment

A singular-value spectrum can help shortlist ranks. Choose several candidates that satisfy backend shape constraints, evaluate them under the same protocol, and compare against other ways of spending the same parameter budget.

Two small matrix multiplications can be slower than one optimized large multiplication because the new graph adds an intermediate tensor and another operator invocation. This follows from the changed execution schedule; determine the net result by measurement rather than parameter arithmetic.

LoRA is not this transformation

LoRA learns low-rank updates while retaining the pretrained base weight. It reduces the number of parameters trained for adaptation. It does not, by itself, replace the base model with a small low-rank inference model. Merging a low-rank update into the base weight preserves the base matrix dimensions. **[R30]**

Low-rank **replacement** and low-rank **adaptation** can both be useful, but record which one the experiment performs. A small adapter file does not establish a small total model footprint.

Verification: require full-rank numerical agreement, valid low-rank shapes, strict reload, and task evaluation. **Decision:** retain a rank only when the deployment metric improves under the quality contract. **Rollback:** preserve the original dense weight and factorization configuration.

9. Reduce resolution and bound intermediate memory

Problem

The weights fit, but long sequences, high-resolution feature maps, or video frames create a large working set. Pruning a few layers may not address the largest simultaneously live tensors. Your next experiment should change the amount of data processed at once or the dimensions produced by the network.

Name the dimension you intend to reduce

An image tensor may have shape `[batch, channels, height, width]`. A video tensor adds a frame or time dimension, whose position depends on the implementation. An attention module may flatten image patches or video patches into a token sequence. Write down the actual layout before changing a shape.

For a fixed number of channels, halving both image dimensions produces one quarter as many feature values. For a dense attention score matrix over flattened tokens, quartering the token count reduces the number of score entries by a factor of sixteen. These are shape calculations. Fused attention may avoid materializing that score matrix, and end-to-end latency includes other operations.

Halving generated frames reduces temporal coverage unless playback rate or another part of the product changes. It is not a free memory optimization when the required output is a fixed-duration video.

Separate four different interventions

Lower input resolution changes what information reaches the model. **Lower internal resolution** changes the architecture or representation. **Tiling** changes the execution region while attempting to preserve an output resolution. **Chunking** changes how a sequence is scheduled, sometimes with carried state.

These interventions have different risks. Record them as separate candidates instead of labeling all four "activation pruning."

Begin with transformations that preserve local computation

The companion image test evaluates one stride-one convolution with an odd square kernel. For each output tile, it reads enough neighboring input pixels to cover the kernel radius. That neighborhood is the tile's **halo**. It applies padding at the original image boundary, not at every internal tile boundary. [E03]

EXAMPLE

```
Desired output tile
  + required surrounding input halo
  -> convolution
  -> write only that tile's output region
```

For the tested single convolution, tiled and full outputs match. The test retains the complete input and output to compare them, so it does not demonstrate bounded total process memory. It establishes the numerical boundary rule for that operator. [E03]

In a deeper network, determine the complete receptive field and any stride alignment. A halo sufficient for one convolution is not automatically sufficient for ten layers. A spatially global operation can make a finite local halo insufficient.

A failing test is part of the lesson

The image lab deliberately adds GroupNorm independently inside each tile. This fails to reproduce GroupNorm applied to the complete convolution output. Each tile supplies different statistics. Calling `eval()` does not turn GroupNorm into fixed whole-image statistics. [R28, E03]

The recorded maximum absolute difference is about 0.678 for that random test. The number is not an image-quality score. It is a concrete counterexample to the claim that any convolutional model can be tiled exactly with a small overlap. [E03]

Likewise, global pooling, attention, and coordinate-dependent operations need explicit analysis. Blending overlapping outputs may conceal seams without restoring the original mathematical function.

Use model-supported tiling where available

Diffusers documents VAE tiling and slicing as distinct mechanisms. Tiling splits spatial work; slicing addresses batches. Its documentation notes possible tile-to-tile tone variation. Therefore, treat a documented tiling switch as a supported approximation or implementation feature that still needs visual validation, not an automatic equivalence certificate. **[R34]**

Do not assume enabling VAE tiling changes the denoiser's largest attention allocation. Profile the stage responsible for the peak and select the matching intervention.

Chunk operations that are independent across positions

A tokenwise feed-forward module processes each token independently while sharing weights. Splitting the sequence for that module and concatenating the outputs can preserve its function when no cross-token operation is introduced. The companion test verifies this for `Linear -> GELU -> Linear` at inference. [E03]

PYTHON

```python
# Executed for a tokenwise FFN with no cross-token operations.
parts = [ffn(part) for part in tokens.split(5, dim=1)]
chunked = torch.cat(parts, dim=1)
```

This test retains all output pieces for comparison. A deployment implementation should avoid retaining unnecessary intermediates. It cannot apply the same independent split to full self-attention without changing which tokens see each other or introducing an appropriate attention algorithm.

Carry state for causal processing

The causal-convolution lab retains only the required input history between chunks. For a stride-one convolution with kernel size `k` and dilation `d`, that history has `(k-1)*d` positions. It pads only the beginning of the stream. It tests ordinary kernels, dilation, and a kernel size of one. [E02]

EXAMPLE

```
previous state + current chunk
  -> causal convolution
  -> output chunk
  -> retain only the required tail as next state
```

The implementation clones the retained tail. A view would share storage with the larger joined input, potentially keeping that allocation alive. PyTorch documents that views share their base storage. [R50]

This is not a streaming adapter for an arbitrary audio model or video Transformer. Bidirectional context, lookahead, normalization, striding, and encoder-decoder skips can require a different contract. Decide whether the target is exact equivalence, a bounded-context approximation, or a newly trained streaming student.

Resolution changes can require recovery

Changing a sample rate is not equivalent to relabeling an audio file. Changing patch size can alter projection shapes. Changing spectrogram bins can alter both encoder and vocoder interfaces. Changing a latent channel count can affect the encoder, denoiser, and decoder together.

For every proposed representation change, name the producer and all consumers. Build adapters or retrain the affected components explicitly. A shape-compatible tensor is not necessarily semantically compatible.

Verification and acceptance

For exact scheduling changes, compare full and chunked outputs at boundaries, on short inputs, on partial final chunks, and across supported shapes. For approximate changes, evaluate the final task and boundary-specific failures separately.

Accept when the specific peak falls and quality remains within the contract. **Revert** when tiling creates tone shifts, chunking creates temporal discontinuities, or lower resolution removes information the product must retain. Never report savings from shortened or downscaled output as same-task acceleration without disclosing the changed output contract.

10. Recover a changed model with fine-tuning and distillation

Problem

A structural change executes correctly but loses task quality. You now have two separate engineering tasks: confirm that the transformation is implemented correctly, and train the modified architecture to use its remaining capacity. Training cannot reliably compensate for an unnoticed channel-mapping bug.

Fine-tuning updates the student using the task objective. **Distillation** adds learning signals from a teacher. The teacher may be the original checkpoint or another suitable model. The original distillation paper and PyTorch's tutorial provide primary descriptions and implementation examples. [R05, R04]

Define what the student is learning

For classification, the student can learn from labels and the teacher's class probabilities. A typical objective combines ordinary cross-entropy with a temperature-scaled divergence:

PYTHON

```python
# This calculation is executed inside fit() in lab.py.
logits = student(x)
with torch.no_grad():
    teacher_logits = teacher(x)

temperature = 2.0
supervised = F.cross_entropy(logits, labels)
soft_targets = F.kl_div(
    F.log_softmax(logits / temperature, dim=-1),
    F.softmax(teacher_logits / temperature, dim=-1),
    reduction="batchmean",
) * temperature**2
loss = 0.5 * supervised + 0.5 * soft_targets
```

The temperature and weights above are fixed choices for the synthetic lab, not recommended defaults for every model. The teacher is held in evaluation mode and receives no gradient updates. The student receives a fresh optimizer after surgery. [E01]

Use `no_grad()` for the teacher branch when its outputs participate in a student loss. Do not indiscriminately wrap the student training step in inference mode. That would prevent the gradient computation the recovery step needs.

Distillation targets depend on the task

For a regression model, matching class probabilities may make no sense. You might compare a teacher's continuous outputs, selected representations, or task-specific features. When teacher and student widths differ, an explicit projection can align intermediate dimensions. Remove training-only adapters from the deployment graph unless they are deliberately part of the student.

For detection or segmentation, preserve the meaning and alignment of spatial outputs. For sequence models, align valid positions and mask padding consistently. For audio synthesis, waveform alignment and perceptual criteria require more care than applying mean squared error to arbitrary samples.

These are proposed design checks. The presence of a loss function in a paper does not establish that it is appropriate for a different representation or application.

Diffusion recovery needs aligned noise and conditioning

A denoiser receives a noisy representation, a noise level or timestep, and possibly text, image, or other conditioning. For a teacher-student comparison, feed both the same noisy state and conditioning. Match the same prediction parameterization, such as noise, clean sample, or velocity, rather than comparing tensors with different meanings. Scheduler documentation exposes these prediction-type distinctions. **[R42]**

A generic proposal is:

EXAMPLE

```
sample a training example and noise level
construct the same noisy input for teacher and student
run both with the same conditioning
compute a compatible prediction or feature loss
update only the student
periodically evaluate complete generated outputs
```

This is not a complete reproduction of progressive distillation or consistency training. Those methods define specific sampling and training objectives. Use their exact method when claiming to implement them. [R38, R39]

Matching one denoising call can support recovery experiments, but it does not prove that errors remain acceptable across the complete generation trajectory. Inspect completed samples as well as the training loss.

Compare recovery against fair controls

At minimum, compare the original checkpoint, the changed model before recovery, and the changed model after recovery. To understand the contribution of distillation, compare it with ordinary fine-tuning under a documented budget. Additional training of the original model is also informative.

The lab includes both cross-entropy and distillation recovery for the depth-pruned model. It also trains the original architecture for additional epochs. This prevents an improvement after pruning from being automatically attributed to pruning itself. [E01]

An equal number of epochs is not necessarily equal compute. Distillation evaluates the teacher and student, while ordinary fine-tuning may evaluate only one model. State whether the comparison equalizes steps, examples, wall time, or compute allocation.

Preserve independent evaluation

Use training data for recovery. Use validation for stopping or candidate selection if that is the recorded protocol. Keep the final test independent. Do not fine-tune on failure examples from a previously reported test and continue calling the result held out.

When using teacher-generated data, record its origin and sampling procedure. Keep independently reviewed task examples in the final suite. A student that reproduces teacher errors is successful at imitation but not necessarily at the application.

Stop when the evidence says to stop

A student can lack sufficient capacity for the required operating envelope. A teacher can lack the capability you need to preserve. Distillation does not guarantee recovery of every removed function or create a fixed percentage of retained quality.

Accept recovery only when the complete candidate passes the same task contract used for the baseline. **Revert** or choose a less aggressive architecture when improvements require excessive training, critical behavior remains missing, or the deployed implementation fails to deliver the resource gain.

11. Quantize deliberately, not by file extension

Problem

You need fewer parameter bytes, less memory traffic, or a backend that requires quantized inputs. Numeric precision is a separate experiment from architecture pruning. A structurally smaller FP32 model and an equally shaped INT8 model save different resources.

Distinguish storage, activations, and arithmetic

A label such as W4A16 means four-bit weights and sixteen-bit activations in the stated scheme. It does not say that every operator uses that representation or that accumulation occurs at four bits. Some components can remain at higher precision.

For dense weight payload alone, 100 million values occupy approximately 400 MB at FP32, 200 MB at FP16, 100 MB at INT8, or 50 MB when packed at four bits. This ignores scales, zero points, alignment, metadata, unquantized tensors, and runtime repacking. It is not a total-memory estimate.

FP16 and BF16 are floating-point formats. INT16 is an integer format. An audio file using sixteen-bit PCM is a representation of the input or output signal, not evidence that a neural model computes using INT16.

Understand what calibration estimates

In an affine quantizer, an integer value is interpreted using a scale and a zero point. Values outside the represented range can be clipped. The choice of range therefore matters as well as the bit count. ONNX Runtime documents this mapping and its static and dynamic quantization workflows. **[R14]**

For static activation quantization, build a calibration subset that represents the inputs and intermediate conditions expected in deployment. Keep it separate from the final test. Calibration is not recovery training, although both can consume training examples.

For image models, include supported resolutions, contrast ranges, and image conditions. For video, include motion and temporal lengths. For language, include representative context lengths and token distributions. For audio, include the intended signal conditions. These are proposed coverage dimensions, not a guarantee that a small curated set is sufficient.

For diffusion, calibrating only clean images or a single noise level can miss the distribution actually seen during denoising. Q-Diffusion specifically studies diffusion quantization and timestep-dependent behavior. This supports using diffusion-aware calibration, not treating a generic image-classification recipe as established for a denoiser. **[R40]**

Choose the workflow according to the backend

Post-training quantization starts from trained weights and estimates a lower-precision representation. **Quantization-aware training** includes simulated quantization effects during training before conversion to a deployable representation. The correct operators and quantization scheme depend on the target toolchain.

For ONNX Runtime, examine the supported static, dynamic, and weight-only paths for the exact operators and execution provider. INT4 support in its documented tooling is operator-specific, including suitable constant-weight matrix multiplications. An INT4 model type in the interchange format is not a universal INT4 convolution kernel. [R14, R49]

Do not combine a quantizer intended for one backend with another backend and assume the resulting graph will be accelerated. Shape, layout, precision, and operator support must agree throughout the deployment path.

Test selective precision before changing everything

Create a floating-point reference export first. Then quantize a supported component and compare it with that reference. Add more components only when the first change is understood. This staged approach is a proposed debugging strategy.

A mixed-precision candidate may retain sensitive operations at higher precision while compressing large projections. The relevant question is not whether 100% of the model is INT8. It is whether the executed graph meets quality and resource requirements.

Repeated dequantization and requantization can add overhead. Weight-only compression may reduce storage without reducing all activation allocations. Measure the runtime rather than assuming that a smaller checkpoint implies faster execution. ONNX Runtime explicitly notes that quantization performance depends on hardware and model behavior. **[R14]**

Debug the first divergence

Compare tensors at a small number of meaningful boundaries. Find where the candidate first differs substantially from the floating-point reference, then inspect saturation, range estimation, precision exclusions, and unsupported operations around that boundary.

Do not retain every activation from a large model just to debug it. Sample a small diagnostic batch or capture one boundary at a time. Record the comparison metric and tolerance. Relative error near zero can be misleading, so include absolute error or another meaningful normalization.

For a generator, passing local tensor tolerances does not replace final-sample evaluation. Small recurrent deviations can produce different outputs. A numerical difference can also be harmless if the task contract allows it. Separate implementation fidelity from product quality.

Recalibrate after structural surgery

Pruning or recovery training can change activation distributions. Reusing quantization statistics from the unmodified checkpoint is an additional hypothesis that must be tested. The conservative workflow is to recalibrate the final recovered structure, then rerun the complete evaluation.

Verification: inspect converted operators, precision boundaries, runtime logs, and final task outputs. **Decision:** keep a quantized candidate only when the actual backend improves the chosen metric. **Rollback:** preserve the recovered floating-point checkpoint and calibration manifest alongside the quantized artifact.

12. Refactor image and video diffusion systems

Problem

A diffusion system can consume substantial resources in several different places: condition encoding, repeated denoising, latent decoding, temporal processing, and output assembly. "Make the diffusion model smaller" is too broad to specify a useful experiment.

Latent diffusion uses an encoded representation rather than operating entirely in image pixels. A Diffusion Transformer can operate on latent patches instead of a U-Net backbone. These architectures have different structural dependencies, even when both are part of an image-generation pipeline. [R31, R32]

Draw the complete pipeline first

A representative pipeline is:

EXAMPLE

```
prompt and conditioning inputs
  -> tokenizer / image preprocessing
  -> condition encoder
  -> initial latent or noisy input
  -> repeated denoiser and scheduler updates
  -> latent decoder
  -> image or video postprocessing
```

Not every pipeline contains every component. Some use additional encoders, adapters, or temporal modules. Inspect the implementation you have rather than treating this diagram as a universal interface.

Measure condition encoding, one denoiser evaluation, the complete denoising loop, decoding, and final encoding separately. Record model-loading time as well. A memory peak in the decoder needs a different intervention from a peak in the denoiser's attention.

Sampling steps are not network layers

A model can contain 24 blocks and be evaluated 30 times during sampling. Removing a block changes one network evaluation. Reducing sampling steps changes how many evaluations are requested and where they occur along the trajectory.

Do not report "30 layers reduced to 10" when you changed `num_inference_steps`. Preserve the distinction in configuration, benchmark names, and review comments.

A simplified cost model is:

EXAMPLE

```
total time = condition encoding
           + sum of actual denoiser evaluation times
           + scheduler work
           + decoding and postprocessing
```

The number of user-visible sampling steps need not equal the number of denoiser calls. Solver behavior and guidance implementation can affect the executed work. Count actual calls or effective evaluated batches instead of inferring them from one setting. Diffusers exposes scheduler configurations and inference-performance options for this purpose. [R42, R48]

Fewer steps usually change total repeated work more directly than the largest single-step activation. If the same denoiser and shape remain resident, a step-count reduction need not lower that peak.

First experiment: a compatible sampler configuration

Start with an unchanged checkpoint and the sampler configuration recommended for it. Record prediction type, timestep or sigma schedule, solver settings, guidance behavior, seed list, output shape, and actual denoiser evaluations.

Test a small number of compatible step counts. Do not switch scheduler, precision, resolution, and checkpoint in the same candidate. Evaluate prompt following, details, and failures on the same suite. Scheduler documentation is explicit about prediction types and solver configuration; arbitrary combinations are not interchangeable. **[R42]**

Do not turn an ordinary many-step model into a one-step model by setting the loop count to one and calling it distilled. Progressive distillation trains a model to reproduce a longer sampler with fewer learned steps. Consistency models also use a specific training construction. These are learned changes, not merely loop optimizations. [R38, R39]

Second experiment: isolate precision and attention implementation

For a supported device, test the precision and attention kernel independently while retaining the same sampling configuration. Measure numerical differences on a fixed noisy input, then evaluate complete outputs across the suite.

A kernel that avoids materializing large attention intermediates can address memory without changing the attention neighborhood. Replacing global attention with local attention changes the interaction pattern and is a different model hypothesis. **[R26]**

Treat compilation as another separate candidate. Report compilation and warmup cost rather than excluding them from a one-image user workflow without explanation. The benefit of a compiled steady-state loop may matter for a batch tool and be less useful for a short session. **[R48]**

Third experiment: decoder and feed-forward scheduling

Try model-supported VAE tiling when the spatial decoder is the bottleneck. Try slicing when a supported decoder processes multiple images as a batch. Neither is a blanket solution to denoiser memory, and tiling can change local tone. **[R34]**

For video, distinguish decoding frames in smaller groups from changing the temporal model itself. Diffusers' Stable Video Diffusion documentation describes feed-forward chunking and `decode_chunk_size`, and warns that very small decode chunks can produce flickering. This is an implementation-specific tradeoff, not evidence that all video models can decode frames independently. **[R35]**

Use the normalization counterexample in the companion lab as a reminder to test the mathematical boundary of any custom tiler. Production decoder behavior requires its own tests.

Fourth experiment: structural pruning of the denoiser

Choose one structurally valid target: repeated residual blocks, internal FFN channels, compatible attention heads, or a coupled convolution group. Establish importance under representative noise levels and conditioning before removing it.

A diffusion denoiser is reused across the generation trajectory. A block that appears inactive at one noise level can matter at another. The Diff-Pruning paper proposes a timestep-aware structural-pruning method, providing primary evidence that diffusion pruning requires attention to this reuse. Its experiments do not validate arbitrary block deletion in a different video generator. **[R41]**

Use the following proposed protocol:

EXAMPLE

```
freeze checkpoint and sampler
  -> collect representative noisy states and conditions
  -> rank a small set of structurally valid candidates
  -> remove one candidate group
  -> run shape and prediction checks
  -> recover with a compatible objective
  -> evaluate complete samples and performance
```

For a U-Net, inspect resolution transitions and skip connections before deleting a block. For a DiT, inspect adaptive conditioning, position handling, token shapes, and the final projection. A matching hidden width is necessary in many edits but not sufficient to establish a valid bypass.

Fifth experiment: lower spatial or temporal resolution

Reducing height and width can substantially reduce the number of latent positions. Increasing patch size can also reduce token count, but changes patch projection and reconstruction interfaces. Treat it as an architecture change rather than a convenient inference flag. The DiT paper explicitly studies depth, width, and token count as different scaling axes. **[R32]**

For video, record generated frame count, conditioning frame rate, playback frame rate, and temporal compression separately. Lowering playback frame rate does not retroactively reduce the model's generation work. Generating fewer frames and interpolating them adds another model and a different artifact profile.

Review fast movement, camera motion, occlusion, object persistence, and scene transitions. A framewise visual score cannot alone establish temporal quality. VBench provides a primary reference for separating these dimensions. **[R43]**

Caching changes the validity conditions

Caching a condition embedding is exact only when every input and model state determining that embedding is unchanged. Include tokenizer version, truncation, encoder weights, prompt text, and relevant conditioning in the cache key.

Reusing a denoiser activation at a different timestep is usually an approximation unless the method establishes a special invariant. Do not confuse this with caching an unchanged text embedding. Treat cross-step activation reuse as its own proposed experiment with an explicit refresh policy and quality checks.

Offloading is not the same as reducing total RAM

A desktop pipeline can move inactive weights between GPU and CPU memory. Diffusers documents several offloading strategies and their transfer costs. This can reduce device-local memory without eliminating the host copy or reducing the size of the model itself. [R34]

For a mobile system, identify the actual memory domains and copy behavior rather than transplanting a desktop CUDA recipe. An app that no longer exceeds a GPU allocation can still exceed the phone's tolerable total footprint. Measure both the accelerator and process-level behavior where those measurements are available.

What a successful report contains

Publish the exact checkpoint and component revisions, complete generation settings, prompt and conditioning suite, seed procedure, output shapes, task-specific evaluation, cold and warm timing, resource metrics, and failed examples.

Accept only a candidate that meets the stated generation contract. **Revert** a candidate that improves static images while introducing video flicker, reduces fidelity outside the tuned prompts, or claims same-task speed by silently shortening the output.

This chapter provides experiment designs supported by primary methods and documentation. No production diffusion checkpoint was compressed or benchmarked in the preparation environment.

13. Export the candidate and inspect the executed graph

Problem

The changed model works in eager PyTorch. You now need an artifact that another runtime can load, execute, and optimize. Export is an interface migration with its own tests, not the final proof that a candidate is mobile-compatible.

ONNX describes a model representation. ONNX Runtime executes supported graphs using its available kernels and execution providers. The interchange format and the runtime are different layers of the system. [R49, R17]

Export after reconstructing the architecture

Load the changed architecture from its saved configuration and require strict checkpoint loading. Do not export a notebook object whose class definitions or in-place mutations cannot be reconstructed.

Define input names, shapes, dtypes, output names, and any state tensors. For a recurrent decoder, include the cache contract. For a denoiser, expose the noisy state, timestep, and conditioning consistently. For a multi-stage system, consider exporting stable component boundaries rather than capturing a Python-heavy orchestration layer.

The PyTorch 2.10 exporter supports a `torch.export`-based path, shape constraints, a diagnostic report, and optional ONNX Runtime verification. These facilities help detect export problems; they do not evaluate the product's task quality. **[R13]**

PYTHON

```python
# Integration template. ONNX execution was not available in this environment.
torch.onnx.export(
    model.eval(),
    (example_input,),
    "candidate.onnx",
    input_names=["input"],
    output_names=["output"],
    opset_version=target_opset,
    dynamo=True,
    report=True,
    verify=True,
)
```

Set `target_opset` from the supported export/runtime combination. Do not copy the newest opset into an older deployment runtime without checking compatibility. Store any external weight files with the graph and include all of them in artifact integrity checks.

Test three boundaries separately

First compare the reconstructed PyTorch model with the original edited object. Then compare the exported floating-point graph with reconstructed PyTorch. Finally compare the quantized or backend-optimized graph with the exported reference.

This isolates structural errors, export errors, and conversion errors. If all three changes happen at once, a bad output does not identify which stage introduced it.

For numerical checks, use representative inputs and the supported shape envelope. Random input can catch shape errors, but it does not establish task fidelity. For generators, compare one step on identical state and also compare completed outputs.

Dynamic shapes are not universally free

A dynamic dimension makes the model more flexible, but the backend may impose restrictions or compile shape-specific variants. ONNX Runtime's QNN documentation states that its documented path requires fixed shapes and supports only a subset of operators. Verify those constraints against the pinned toolchain. **[R18]**

A proposed deployment strategy is to define a small set of shape buckets. Each bucket needs correct preprocessing, padding or cropping rules, and its own tests. Do not improve benchmark results by padding away difficult cases or silently truncating user input.

Read the execution-provider allocation

An accelerator can execute only the graph portions it supports. The runtime may partition supported subgraphs and leave other work on another provider. Cross-provider transitions can introduce transfers and synchronization. ONNX Runtime documents this execution-provider model and recommends testing the actual mobile combination. [R17, R16]

Inspect the execution trace rather than assuming that requesting an NPU means every operation ran on it. Record unsupported operators, their attributes and shapes, and the cost of crossing each boundary.

Sometimes changing an export decomposition is enough. Sometimes a CPU-only graph is preferable to a fragmented accelerated graph. Sometimes the architecture needs to change. These are measurements to perform, not a fixed ranking of CPU, GPU, and NPU.

Verify graph optimizations independently

Constant folding, dead-node removal, and supported operator fusions can reduce execution overhead without the same learning problem as pruning. ONNX Runtime documents graph-optimization levels. Some optimizations are hardware-dependent, so retain the target platform in the artifact metadata. **[R15]**

Do not equate an operator disappearing from a graph viewer with a learned capability disappearing. It may have been fused into another kernel. Conversely, a smaller serialized graph can still require a large temporary workspace.

Runtime size is a different budget

ONNX Runtime can use a reduced operator configuration to build a runtime containing only the required operator and type support. This reduces the runtime package, not necessarily the model's activation peak. Generate the configuration from all model variants the application must support. **[R19]**

ExecuTorch is another edge-deployment stack with its own export, lowering, and backend support. It is not simply a different reader for every ONNX artifact. Evaluate toolchains against the graph and devices you actually need, rather than selecting one by a general popularity claim. **[R20]**

Verification: strict reload, export equivalence, operator coverage, provider trace, and final task tests. **Decision:** keep the candidate only after it runs through the intended deployment path. **Rollback:** retain the recovered checkpoint, export settings, runtime build, and previous deployable artifact.

14. Integrate the model into a device application

Problem

A model benchmark succeeds in isolation, but the application also loads media, allocates buffers, displays previews, writes files, and responds to lifecycle events. Device compatibility is a system property, not just a tensor-shape property.

The procedures below are proposed integration tests. No Android device was available for the accompanying experiments.

Define an application memory envelope

Measure process-level behavior separately from tensor payloads. Keep managed heap, native allocations, mapped files, accelerator memory, and shared accounting distinct. Android documents memory-pressure behavior and device-dependent heap constraints; it does not define a universal 900 MB CPU-activation allowance. **[R21]**

A useful initial diagnostic is:

```shell
# Replace the package name with the application under test.
adb shell dumpsys meminfo com.example.modelapp
```

This produces a snapshot, not a guaranteed record of the peak between snapshots. Use a trace or a suitable profiler for short-lived spikes, and keep the measurement method in the report. The meaning of the fields is documented in Android's `dumpsys` reference. **[R22]**

Do not add parameter bytes to RSS and call the result total RAM. The resident weights may already be included in RSS. Do not add independently measured stage peaks and claim they are simultaneously live.

Schedule stages deliberately

When components run sequentially, consider loading only the active stage and retaining minimal intermediate state. For diffusion, the denoiser is reused many times, so repeatedly unloading it between every step can add avoidable overhead. For a multi-model media pipeline, persistent outputs can connect stages without keeping every model resident.

Unloading an object does not guarantee an immediate fall in the measured process footprint. Allocator caches, mapped storage, shared references, and runtime workspaces can remain. Test repeated jobs in a clean process and in a long-lived process. Record whether the working set stabilizes or grows.

Bound application queues. Decoding an entire video to uncompressed frames can dominate memory before the model runs. Limit pending inputs and outputs, and make the producer wait when the consumer cannot keep up. This is a proposed application-level memory policy.

Benchmark sustained work

Measure cold initialization, first inference, warm calls, and complete jobs. Include the supported workload lengths, not only a single short input. Record device model, SoC, OS build, runtime, thread count, charging state, and relevant thermal conditions.

Android's Thermal API provides status and headroom signals for adapting workload. It does not guarantee the same sustainable performance on all devices. Define a pause, reduced-concurrency, or quality-tier policy and test that it preserves a coherent job state. [R23]

Do not infer sustained latency from a desktop CPU result or a phone's advertised peak accelerator throughput. Your application executes a particular graph with particular memory movement and scheduling.

Test concurrency rather than maximizing threads

Use a documented range of thread and concurrency configurations. Test the entire application, since a codec, a rendering thread, and an inference runtime may compete for the same cores and bandwidth.

Separate throughput goals from response-time goals. Running two model instances can improve completed jobs per unit time while worsening per-job latency and memory. The correct configuration is the one that satisfies the product contract.

Treat cancellation and process death as normal cases

An offline media job should record its completed stage, input digest, model versions, and resumable artifacts. Write final output atomically where the platform permits so a partial file is not presented as a successful export.

Use the appropriate platform mechanism for user-visible or deferred long-running work. Android's long-running worker guidance includes version-specific constraints; it does not authorize an unlimited background computation simply because the task uses ML. Check the rules for the application's actual target SDK. [R24]

Test cancellation during model loading, inference, and output writing. Test insufficient storage and damaged model files. These failures belong in the deployment suite even when all tensor calculations are correct.

Ship a complete artifact contract

Package the architecture variant, checkpoint digest, preprocessing version, runtime compatibility, input envelope, and output format. Validate downloads before loading. Retain a known-good model version and a way to select it when a candidate fails a device-specific check.

Accept a deployment only after complete jobs pass quality, lifecycle, storage, memory, and sustained-performance tests on the supported device matrix. **Revert** a model update independently from an application update when the packaging and compatibility design permit it.

15. Inspect the executed experiments

What the experiments establish

The companion programs are deliberately small. They let you inspect the complete transformation, test the shape contract, and separate a programming error from an approximation error. They are not miniature evidence that a particular pretrained video generator can be compressed successfully.

There are three evidence records. RefactorNet is a trained synthetic classifier. The causal-convolution example tests stateful chunking. The vision and attention example tests structural equivalence on randomly initialized modules. Their scope and failure conditions differ, so their results must not be combined into one compression claim. [E01, E02, E03]

Experiment A: block removal and feed-forward pruning

The classifier maps 16 input values to four classes. Its stem produces 32 features. Six residual blocks preserve that width and each expands internally to 96 features. The labels come from a fixed nonlinear function implemented in `make_data`, not a downloaded media dataset.

The run used 4,096 training examples, 1,024 validation examples, and 1,024 independently generated test examples. The first 256 training inputs supplied channel-importance statistics. The original model trained for 24 epochs. Recovery variants received eight additional epochs with fresh optimizers. The seed was 20260924. The saved protocol records the exact learning rate, batch size, and distillation settings. [E01]

Each removable block was tested separately against validation cross-entropy. Indices below are zero-based. A negative delta means the ablated model had lower validation loss than the original in this experiment.

Removed block	Validation loss	Loss delta	Validation accuracy
0	0.6585	+0.1801	80.96%
1	0.6566	+0.1781	82.03%
2	0.5262	+0.0478	83.59%
3	0.5585	+0.0800	83.20%
4	0.4447	-0.0338	85.55%
5	0.3978	-0.0807	87.01%

Block 5 was selected by the recorded loss criterion. The selection did not use test accuracy. The run then compared depth reduction, internal width reduction from 96 to 64, their combination, and recovery variants. The classifier's public hidden width remained 32. [E01]

Recorded quality and timing

In the table, **raw** means no recovery after surgery. **CE** means recovery with supervised cross-entropy. **KD** means the executed mixture of supervised loss and teacher-student distillation. These names describe the training recipe, not a claim that one method is always superior.

Variant	Parameters	Validation accuracy	Test accuracy	Median ms	p95 ms
`baseline`	38,756	86.82%	85.25%	0.1705	0.2966
`depth_raw`	32,420	87.01%	84.96%	0.1456	0.2353
`depth_ce`	32,420	85.74%	85.16%	0.1456	0.2985
`depth_kd`	32,420	87.01%	86.13%	0.1450	0.2255
`width_raw`	26,276	84.96%	86.52%	0.1581	0.2394
`width_kd`	26,276	87.01%	87.40%	0.1680	0.2731
`combined_raw`	22,020	84.47%	85.55%	0.1433	0.2578
`combined_kd`	22,020	86.23%	87.40%	0.1434	0.2583
`baseline_extra_ce`	38,756	85.55%	87.21%	0.1699	0.2218

Timing was measured on a Linux x86_64 host reporting an AMD EPYC 9V74 CPU, using PyTorch 2.10.0+cpu and one intra-op thread. The workload was one input of shape `[1, 16]`. Each model received 30 warmup calls, followed by five rounds of 200 measured calls. Model order was shuffled between rounds. All models remained resident. These are warm Python/PyTorch call timings, not initialization, peak-memory, GPU, or Android measurements. The raw samples and round ordering are included. [E01]

Interpretation without overstating the result

The combined KD model has 22,020 parameters compared with 38,756 in the original, a reduction of 43.2%. Its recorded median call time was 15.9% lower. These quantities did not fall by the same proportion. This is one small-model CPU observation, not a predicted speedup for a larger model or another runtime. [E01]

The original model's recorded test accuracy was 85.25%. The combined KD variant recorded 87.40%, but the unpruned model with the extra supervised training budget recorded 87.21%. The difference between the latter two is two examples out of 1,024. One seed and one test split do not establish that the compressed model is generally more accurate. [E01]

The ablation selected using validation loss slightly reduced test accuracy before recovery. Some recovery runs improved test accuracy while their validation results did not improve in the same way. Preserve these results instead of selecting only a favorable metric. They illustrate why the selection criterion, final test, and training controls must remain explicit.

The run does not include matched-compute recovery, repeated training seeds, confidence intervals, or a student trained from scratch. It therefore does not establish that distillation was necessary, that it was more compute-efficient, or that this architecture is optimally compressed. It also did not measure peak process memory or peak live activations. Parameter bytes and module-output sizes are not substitutes for those measurements.

Experiment B: chunked causal convolution

`chunking.py` tests a single causal one-dimensional convolution with persistent left context. The three tested configurations include a kernel of size five, dilation one or two, and a kernel of size one. The largest recorded absolute error was approximately `2.38e-7`, within the test tolerance. [E02]

This supports the implementation of the state buffer for the tested operation. It does not establish that an arbitrary noncausal audio encoder can be made causal or streamed without changing its behavior. The test collects outputs for comparison and does not measure end-to-end peak memory.

Experiment C: image, attention, and chunking contracts

The convolution-channel test reduces a two-convolution module from 663 to 333 parameters. The attention-head test reduces its custom module from 4,224 to 2,128 parameters while preserving its external 32-feature interface. Both match their specifically masked reference functions within the recorded checks. Neither is asserted to match the original unmasked function. [E03]

The image tiler reproduces a single convolution for three tested tile sizes when it includes the required input halo. Tokenwise feed-forward chunking also matches the unchunked calculation for the tested module. These tests use random weights and inputs, so they establish implementation behavior, not learned task quality. [E03]

The deliberately incorrect per-tile GroupNorm calculation produced a maximum absolute difference of approximately 0.678 from normalization over the complete feature map. The test passes by confirming the expected mismatch. It is a regression test against an invalid assumption, not a successful approximation to deploy. [E03]

What to reproduce next

Run the self-tests first. Then reproduce the synthetic training experiment in a new output directory. Do not overwrite the supplied evidence. Once you understand the tests, replace the toy architecture and data with a specific checkpoint and task suite, retaining the same distinction between structural checks, task evaluation, and device measurements.

The first useful production experiment is a small, reversible change with a clearly measured bottleneck. There is no requirement to begin with the most aggressive compression technique.

16. Plan experiments across model families

A reusable experiment contract

Before coding, write one sentence that connects an observed cost to a proposed change. For example: "The decoder's spatial activation peak exceeds our budget; test its supported tiling path without modifying the denoiser." This is narrower and more testable than "optimize the model."

Use the following template. Empty measurement fields are intentional. They must be populated by execution, not estimated from the desired outcome.

YAML

```yaml
experiment_id: exp_001
status: proposed
baseline:
  architecture_revision: REQUIRED
  checkpoint_sha256: REQUIRED
  runtime_and_backend: REQUIRED
  input_contract: REQUIRED
hypothesis: REQUIRED
one_primary_change: REQUIRED
held_constant:
  - evaluation_cases
  - preprocessing
  - unrelated_inference_settings
structural_checks: []
quality_gates: []
performance_objective: REQUIRED
recovery_budget: REQUIRED_OR_NONE
measurements: {}
known_limitations: []
decision: not_evaluated
rollback_artifact: REQUIRED
```

Keep an experiment directory containing this contract, the patch, the generated architecture configuration, outputs, environment details, and the final decision. A failed experiment is useful when the failure is reproducible.

Image classification, detection, and segmentation

A proposed first structural experiment is to prune one compatible convolution-channel group or one internal FFN width while preserving the prediction interface. Use dependency analysis for branches and coupled parameters. DepGraph provides a concrete implementation and research basis for identifying such dependencies. It does not choose your product's acceptable recall loss. [R02, R03]

Keep input resolution unchanged for that experiment. Measure per-class outcomes and boundary or small-object behavior where relevant. Then test resolution separately. Combining channel pruning and reduced resolution in the first run makes it difficult to identify which change removed fine detail.

For a segmentation model, a matching output shape is only a structural check. Examine the output alignment, thin structures, and task-specific errors. For a detector, preserve label mappings, box coordinate conventions, and postprocessing settings.

Image diffusion

Begin by locating the expensive stage. A condition encoder, denoiser, and latent decoder need different interventions. Try a compatible inference configuration or supported memory-scheduling option before committing to a new architecture. The diffusion chapter describes these as separate hypotheses. [R34, R42, R48]

For structural pruning, capture representative noisy states and conditioning, identify a valid dependency group, test a local prediction boundary, and recover using a compatible objective. Then evaluate completed samples. Noise prediction error alone does not establish composition, prompt adherence, or visual diversity.

Keep a fixed prompt and seed suite for comparison. Record failures such as omitted subjects, broken spatial relations, mask violations, or unwanted changes outside an edit region. Do not evaluate only whether a generated image looks attractive.

Video diffusion

First record temporal length, spatial dimensions, conditioning frame rate, playback frame rate, and decoder grouping. Keep these quantities separate. A proposal to generate fewer frames is a workload change. A proposal to decode the same latent sequence in groups is a scheduling change whose validity depends on the decoder.

Test identity, motion, flicker, occlusion, and scene changes independently. VBench is a reference for this multidimensional evaluation. A framewise score cannot establish all of those properties. **[R43]**

Temporal chunking deserves a dedicated test. State how context crosses chunk boundaries, whether normalization depends on the complete sequence, and whether a frame can attend to information outside the chunk. The image tiling counterexample is a useful warning, not a direct proof about your video architecture.

Language models

Separate prompt processing from autoregressive generation. Profile weights, intermediate allocations, and cache growth rather than publishing one memory number for all context lengths. A head-pruning change can affect query projections without shrinking a shared key/value cache. **[R46]**

Begin with a compatible internal FFN or block ablation. Verify attention masks, position handling, cache indices, and tied parameters. Evaluate the actual tasks, including structured output and longer contexts when they are supported. Short-prompt fluency is not evidence that the full operating envelope survived.

Keep decoding settings fixed when comparing model changes. A different sampling temperature can conceal or imitate a behavioral change in the network.

Audio models

State whether the model operates on waveforms, spectrograms, tokens, or another latent representation. Sample rate, frame hop, channel count, and causal context are part of the contract. Changing them can alter the task rather than merely making the same calculation cheaper.

Test boundary behavior when chunking. Preserve required context and inspect timing or phase behavior as appropriate. For generated speech, intelligibility, speaker identity, and prosody are separate acceptance dimensions. For separation, inspect leakage and artifacts rather than assuming clearer speech means better source recovery.

The supplied causal-convolution test is a starting point for reasoning about state. It is not an exportable replacement for a production separator or speech model.

Compare candidates without hiding tradeoffs

Maintain a table of quality, latency, memory, energy when measured, and recovery cost. Mark every unmeasured field as unmeasured. Do not fill a memory column with a parameter-count estimate simply because the table looks incomplete.

Prefer reporting tradeoffs to inventing one universal score. A candidate can be useful for batch generation and unsuitable for live interaction. A smaller checkpoint can be valuable for distribution even when execution is not faster. Keep the deployment objective attached to the decision.

17. Review, accept, or revert

Review the claim before the code

An optimization review starts with a claim such as "reduces warm batch-one latency for these input shapes on this backend." Check that the evidence actually measures that claim. A desktop MAC count does not prove phone latency. A masked-equivalence test does not prove unchanged task quality.

Use three review questions: Is the transformation implemented correctly? Does the task still meet its requirements? Does the intended deployment benefit exist? Evidence for one question does not answer the other two.

Review the transformation

Inspect dependent dimensions, residual interfaces, output contracts, state, and serialization. Confirm that the pruned architecture can be reconstructed from a clean process. New parameters require a deliberate optimizer policy during recovery. Reusing an optimizer that still references removed tensors is not a valid migration.

Check that code snippets marked as illustrative are not silently promoted to executed recipes. A custom module in this handbook is not a drop-in patch for a pretrained implementation merely because both contain a `Linear` layer.

Review the evaluation

Confirm that candidate selection and final testing are separated. Look for duplicate or correlated inputs across splits. Verify that model, metric, and preprocessing versions are recorded. Inspect examples where the candidate fails, not only aggregate improvements.

For generative outputs, verify the prompt list, conditioning, seed handling, and review procedure. For recovery, compare against an appropriate extra-training control. State when repeated seeds or independent review were not performed.

Review deployment evidence

Read the executed graph and backend trace, then inspect full-application behavior. Include model loading, repeated jobs, cancellation, and long inputs. Confirm that timing excludes or includes compilation and initialization intentionally rather than accidentally.

Treat missing measurements as open work. "Peak memory not measured" is an acceptable limitation in a prototype report. "Peak memory reduced" without a measurement is not.

Write the decision record

A useful final record names the candidate, identifies the measured improvement, states the accepted quality tradeoff, lists known limitations, and links the rollback artifact. It also identifies the supported environment. Avoid declaring a general winner when the experiment tested one device and one input shape.

A release decision may be **accept**, **accept for a restricted operating envelope**, **investigate**, or **revert**. These are engineering decisions against a written contract, not universal judgments about an architecture.

The endpoint is not the smallest possible network. It is a deployable model whose behavior and costs are understood well enough for the intended application.

Appendix A. Run the companion code

Package structure

The complete draft package contains the manuscript, these scripts, and the recorded evidence. The original laboratory checkpoints are generated examples, not third-party pretrained weights.

EXAMPLE

```
Model_Refactoring_Draft/
  Model_Refactoring_for_Software_Engineers_Draft.md
  README.md
  requirements-lab.txt
  code/
    lab.py
    chunking.py
    vision_attention_lab.py
  evidence/
    sources.json
    run/
      report.json
      timing.json
      module_outputs.json
      baseline.pt
      ... candidate checkpoints ...
    chunking_results.json
    vision_attention.json
  SHA256SUMS.txt
```

The manuscript's `evidence/` references are relative to this package. Downloading only the Markdown manuscript provides the text and bibliography, not the referenced local experiment files.

Environment and installation

The recorded environment used Python 3.13.5 and PyTorch 2.10.0+cpu on Linux x86_64. The scripts use the Python standard library and PyTorch. Install an appropriate PyTorch build for your platform. The requirement file identifies the tested base package version; it is not a complete platform lockfile and does not guarantee an identical wheel on another operating system.

SHELL

```
python -m venv .venv
# Activate the environment using your platform's command.
python -m pip install -r requirements-lab.txt
python -c "import torch; print(torch.__version__)"
```

Package installation requires a suitable package source and network or cached wheels. It is separate from running the examples. The recorded environment could not install the additional ONNX packages, so no ONNX or quantization result is included.

Run the mechanism tests

From the extracted package directory:

```shell
python code/lab.py --self-test
python code/chunking.py
python code/vision_attention_lab.py --out vision_attention_rerun.json
```

A test confirming that naive normalization differs is expected to pass by detecting the mismatch. Do not change that test into an equivalence assertion to make an incorrect implementation appear acceptable.

Run the learning experiment

```shell
python code/lab.py --out new_run
```

The command refuses an existing output directory. Choose a new name for every run. It generates its own numerical dataset, trains the baseline, selects the block using validation loss, constructs the candidates, runs the fixed recovery recipes, and writes results and checkpoints.

Compare the new report with the supplied one. Expect hardware-dependent timing and possible numerical differences across releases or platforms. The supplied evidence is a record of a particular run, not a promise of bitwise equality on every machine. [R10]

Load a saved example

```python
from pathlib import Path
import sys
import torch

sys.path.insert(0, "code")
from lab import load_model

model = load_model(Path("evidence/run/combined_kd.pt"))
with torch.inference_mode():
    logits = model(torch.zeros(1, 16))
assert logits.shape == (1, 4)
```

The loader reconstructs the architecture from its versioned configuration and uses strict state-dictionary loading. Only load checkpoints whose origin you trust. The supplied loader is intended for these generated examples, not arbitrary objects or architecture files.

Extend rather than overwrite

Create a new experiment directory for a pretrained model. Retain its architecture revision, checkpoint digest, licenses, preprocessing, and evaluator configuration. Add model-specific tests before adapting a transformation. Keep this handbook's recorded results unchanged so the difference between the example and your experiment remains visible.

Appendix B. Evidence directory

E01. RefactorNet 1.0 CPU learning lab

Origin: Original code in `code/lab.py`, executed in the preparation environment. **Primary records:** `evidence/run/report.json` and `evidence/run/timing.json`. **Supporting artifacts:** checkpoint files and `module_outputs.json`.

The report records the source digest, seed, environment, split sizes, training protocol, ablation ranking, retained channel indices, parameter counts, checkpoint digests, quality values, and limitations. Timing samples are included rather than only the aggregate values. This is one synthetic classification experiment, not a pretrained-model benchmark.

E02. Causal-convolution chunking checks

Origin: Original code in `code/chunking.py`. **Record:** `evidence/chunking_results.json`.

Three configurations test a single causal convolution with carried input context. The record contains kernel, dilation, chunk size, and maximum absolute error. No process-memory or production-audio measurement is claimed.

E03. Vision and attention mechanism tests

Origin: Original code in `code/vision_attention_lab.py`. **Record:** `evidence/vision_attention.json`.

The record includes the environment, source digest, parameter changes, numerical equivalence checks, and the intentionally nonequivalent normalization case. Randomly initialized modules and random inputs were used. The tests do not measure image-generation quality, temporal coherence, or device performance.

Appendix C. Technical glossary and lookup index

This appendix is an alphabetical lookup for terminology used in model refactoring, compression, deployment, and performance work. It also includes common terms that a software engineer is likely to encounter while applying the techniques in this handbook. Definitions are intentionally implementation-oriented. When behavior depends on a specific framework, runtime, hardware backend, or model revision, verify it against the pinned implementation used by the experiment.

Quick abbreviation index

Abbreviation	Meaning
BF16	Brain floating point, 16-bit floating-point format
CPU	Central processing unit
CUDA	NVIDIA parallel computing platform and programming model
EP	Execution Provider in ONNX Runtime terminology
FFN	Feed-forward network
FLOPs	Floating-point operations, usually used as an approximate compute measure
FP16	IEEE half-precision floating point
FP32	IEEE single-precision floating point
GELU	Gaussian Error Linear Unit
GEMM	General matrix multiplication
GPU	Graphics processing unit
INT4	4-bit integer representation
INT8	8-bit integer representation
KV cache	Key-value cache used by autoregressive attention models
MAC	Multiply-accumulate operation
NPU	Neural processing unit
ONNX	Open Neural Network Exchange
PTQ	Post-training quantization
QAT	Quantization-aware training

Abbreviation	Meaning
QNN	Qualcomm AI software stack and ONNX Runtime execution provider
Q/K/V	Query, key, and value projections in attention
RAM	Random-access memory
RSS	Resident Set Size
SIMD	Single Instruction, Multiple Data
SVD	Singular Value Decomposition
TFLite	TensorFlow Lite
VAE	Variational Autoencoder

A

Activation

A tensor produced while data passes through a model layer. Activations are temporary inference values rather than persistent learned parameters. Their size and lifetime can dominate peak memory even when model weights are small. See Chapters 2 and 9.

Activation function

A nonlinear function applied to a layer output. Common examples include ReLU, GELU, SiLU, and tanh. Activation choice can affect accuracy, numerical range, quantization behavior, and runtime operator support.

Activation memory

Memory occupied by live intermediate tensors during inference or training. It depends on tensor shape, dtype, execution order, retained skip connections, attention state, and runtime memory planning. See Chapters 2 and 9.

Adam and AdamW

Adaptive gradient-based optimizers widely used for training neural networks. AdamW separates weight decay from the gradient update. Optimizer state can require substantially more training memory than inference because moment estimates are stored for parameters.

Attention

A mechanism that computes interactions between sequence elements. Standard self-attention forms query, key, and value tensors and combines values according to similarity between queries and keys. See Chapter 7.

Attention head

One parallel attention subspace inside multi-head attention. Removing a head only saves meaningful compute when the associated projection dimensions or operations are physically reduced. Disabling a head while preserving the same dense matrices may provide little speedup. See Chapter 7.

Autoregressive model

A model that generates each new token, frame, or sample conditioned on previous outputs. Autoregressive decoding often has sequential latency and may use persistent state such as a KV cache.

B

Batch size

The number of samples processed together. Larger batches can improve accelerator utilization but increase activation memory. Mobile inference commonly uses batch size 1.

BF16

A 16-bit floating-point format with the same exponent width as FP32 but fewer mantissa bits. It provides a wide dynamic range and is often useful for training or inference on hardware with native BF16 support.

Bottleneck block

A block that reduces a representation to a smaller width for expensive computation and then expands it again. Bottlenecks are common in efficient convolutional and transformer architectures.

Buffer reuse

A runtime optimization that reuses memory after an intermediate tensor is no longer live. Effective reuse can reduce peak memory without changing model mathematics. See Chapter 2.

C

Calibration data

Representative inputs used to estimate numerical ranges for post-training quantization. Poor calibration data can produce inaccurate scales and larger quality regressions. See Chapter 11.

Channel

A feature dimension in convolutional or other tensor representations. Structured channel pruning physically removes selected channels and dependent weights. See Chapter 6.

Checkpoint

A saved representation of learned model state, usually parameters and sometimes optimizer state, scheduler state, configuration, or metadata. A checkpoint alone may not fully specify inference behavior. See Chapter 1.

Chunking

Processing a long input as smaller windows while preserving required context. Chunking can bound activation memory and enable streaming, but correctness depends on overlap, receptive field, state handling, and boundary treatment. See Chapter 9.

Compute graph

The network of operations and tensor dependencies executed by a framework or runtime. Graph structure determines operator compatibility, scheduling opportunities, and tensor lifetimes.

Convolution

An operation that applies learned filters across spatial, temporal, or other structured dimensions. Convolutional cost depends on channels, kernel size, input resolution, groups, stride, and output size.

CPU

A general-purpose processor. Mobile CPUs usually contain heterogeneous cores with different performance and power characteristics. CPU inference performance depends heavily on vectorized kernels, memory bandwidth, thread scheduling, and sustained thermals.

CUDA

NVIDIA's platform for executing parallel workloads on NVIDIA GPUs. A model that relies on custom CUDA operators usually requires modification before it can run on non-NVIDIA mobile hardware.

D

Decoder

The part of a model that transforms an internal representation into an output representation. Depending on the architecture, a decoder may generate tokens, spectrograms, images, latent tensors, masks, or other outputs.

Dequantization

Conversion of a quantized integer representation back to a floating-point representation or to another numerical domain used for computation. Frequent quantize-dequantize transitions can reduce expected performance gains.

Depthwise convolution

A grouped convolution in which each input channel is filtered independently, commonly followed by a pointwise convolution. It is a core building block of many mobile-friendly vision architectures.

Diffusion model

A generative model that learns to reverse a noise process, typically through repeated denoising steps. Image and video diffusion systems often contain a denoiser, text encoder, scheduler, and VAE. See Chapter 12.

Distillation

Training a smaller or otherwise constrained student model using signals from a teacher model. The student may imitate logits, hidden representations, features, predicted noise, audio features, or other task-specific targets. See Chapter 10.

Dtype

The numerical data type of a tensor, such as FP32, FP16, BF16, INT8, or INT4. Dtype affects storage, bandwidth, numerical behavior, supported kernels, and sometimes activation memory.

Dynamic shape

A tensor dimension that can vary at runtime. Dynamic shapes improve flexibility but can restrict some graph optimizations, static memory planning, or accelerator compilation strategies.

E

Embedding

A learned vector representation for an item such as a token, speaker, class, timestep, or condition. Embedding dimensions contribute to parameter size and sometimes to persistent state.

Encoder

The portion of a model that converts an input into an internal representation. Encoders are used in language, vision, audio, and multimodal systems.

Execution Provider (EP)

An ONNX Runtime backend that implements execution on a specific class of hardware or software stack, such as CPU, CUDA, or Qualcomm QNN. Unsupported operators may cause graph partitioning across providers. See Chapter 13.

F

Feed-forward network (FFN)

A per-position multilayer network used inside transformer blocks, commonly expanding the hidden dimension and then projecting it back down. FFNs can account for a large fraction of transformer parameters and compute. See Chapter 5.

Fine-tuning

Additional training of an existing model on selected data or objectives. Fine-tuning is often used after structural pruning or other surgery to recover quality.

FLOPs

An estimate of floating-point operation count. FLOPs can help compare arithmetic workload, but they do not directly predict latency, energy use, memory traffic, or accelerator utilization. See Chapter 2.

FP16

A 16-bit floating-point format. Compared with FP32, FP16 usually halves raw tensor storage and memory traffic. Whether it is faster depends on hardware and kernel support.

FP32

A 32-bit floating-point format commonly used as a reference precision for training and inference.

Fused operator

A runtime or compiler operation that combines multiple graph operations into one kernel or optimized execution region. Fusion can reduce memory traffic and launch overhead.

G

GELU

Gaussian Error Linear Unit. A smooth activation function common in transformer architectures. In conceptual form, GELU multiplies the input by the probability that a standard normal variable is less than that input. Implementations often use an exact or approximate formulation. When refactoring a model, preserve the formulation expected by the checkpoint unless equivalence has been validated.

GEMM

General matrix multiplication. Dense neural network layers, transformer projections, and many convolutions ultimately map to GEMM-like kernels. GEMM efficiency is strongly influenced by tensor dimensions, dtype, memory layout, and hardware support.

GPU

A processor optimized for highly parallel workloads. GPUs can accelerate many neural operations but are sensitive to kernel launch overhead, memory transfer, tensor layout, supported precision, and sustained thermal limits on mobile devices.

Gradient

The derivative of a training objective with respect to model parameters. Gradients are used by optimizers to update parameters. They are normally not required during inference.

Graph optimization

A transformation applied to the computational graph to improve execution while preserving intended semantics, unless explicitly documented as approximate. Examples include constant folding, operator fusion, and eliminating redundant nodes. See Chapter 13.

GroupNorm

Group Normalization. A normalization layer that divides channels into groups and normalizes within each group. It is widely used in diffusion and vision models because it does not depend on batch statistics in the same way as BatchNorm.

H

Head dimension

The dimensionality assigned to one attention head. In many implementations, hidden size equals the number of heads multiplied by head dimension.

Hidden size

The width of the main internal representation of a model. Reducing hidden size can reduce parameters, compute, and activations, but it usually requires coordinated changes throughout the architecture.

I

Inference

Running a trained model to produce outputs without performing parameter updates. Inference usually requires much less memory than training because gradients and optimizer state are absent.

INT4

A 4-bit integer representation used for aggressive quantization, most commonly for weights. INT4 cuts raw weight storage to one eighth of FP32, excluding scales and metadata. Runtime speedup depends on native kernel support and whether unpacking or dequantization is required.

INT8

An 8-bit integer representation commonly used for mobile and edge quantization. INT8 can reduce weight and activation bandwidth and may use integer matrix kernels when supported. See Chapter 11.

Intermediate tensor

A temporary value produced between model operations. Intermediate tensors are often what engineers mean when discussing activation memory.

K

Kernel

A low-level implementation of an operation executed by a CPU, GPU, NPU, or other backend. Two runtimes can execute the same model graph with very different performance because their kernels differ.

Knowledge distillation

See Distillation.

KV cache

Key-value cache. Persistent attention state stored during autoregressive generation so earlier tokens do not need to be recomputed at every decoding step. KV cache memory grows with sequence length, layer count, head configuration, batch size, and cache dtype.

L

Latency

Elapsed time required to complete an operation or inference request. Report latency with input shape, batch size, device, thread configuration, warmup procedure, and percentile or averaging method. See Chapters 1 and 2.

LayerNorm

Layer Normalization. A normalization operation over selected feature dimensions, widely used in transformer architectures.

LoRA

Low-Rank Adaptation. A parameter-efficient fine-tuning method that trains low-rank update matrices while leaving the original dense weights frozen. LoRA is primarily an adaptation technique, although low-rank ideas are also used for compression. See Chapter 8 and reference R30.

Low-rank factorization

Approximating a large matrix with the product of smaller matrices. It can reduce parameters and sometimes compute when a sufficiently small rank preserves acceptable behavior. See Chapter 8.

M

MAC

Multiply-accumulate operation. A common unit for describing neural network arithmetic. MAC counts do not capture memory traffic, control flow, or device-specific efficiency.

Memory bandwidth

The rate at which data can be moved between memory and compute units. Neural inference can be bandwidth-bound even when the processor has unused arithmetic capacity.

Memory-mapped weights

Model parameters accessed through operating-system virtual memory mapping rather than copied into a separately allocated buffer. Mapped file size and resident memory are different measurements.

Mixed precision

Using more than one numerical precision within a model, such as INT8 for robust encoder layers and FP16 for numerically sensitive output layers.

Model surgery

An engineering term for changing a trained model's structure, such as removing blocks, pruning channels, replacing layers, or changing internal dimensions, followed by validation and often recovery training.

N

NPU

Neural processing unit. A specialized accelerator designed for machine-learning workloads. Actual model acceleration depends on supported operators, tensor shapes, precision, compiler behavior, and graph partitioning.

Normalization

An operation that rescales or recenters activations according to defined statistics. Examples include LayerNorm, GroupNorm, BatchNorm, and RMSNorm. Replacing one normalization method with another is generally not behavior-preserving without retraining or validation.

O

ONNX

Open Neural Network Exchange. A model representation format used to describe graphs of operators and tensors across frameworks and runtimes. ONNX itself is not the execution engine. See Chapter 13.

ONNX Runtime

An inference runtime that executes ONNX models through one or more Execution Providers. It performs graph optimization, memory planning, kernel dispatch, and backend partitioning.

Operator

A graph-level computation such as MatMul, Conv, Add, Softmax, LayerNormalization, or Reshape. Mobile deployment often fails or slows down when required operators are unsupported by the intended accelerator.

Operator fusion

Combining adjacent operations into a single optimized execution unit. Fusion can reduce intermediate memory traffic and dispatch overhead.

P

Parameter

A learned value stored in the model, such as a weight or bias. Parameters persist across inference calls and differ from temporary activations.

Parameter count

The number of learned scalar values in a model. Parameter count is useful for estimating dense weight storage but is not a complete measure of runtime memory or latency.

Peak memory

The maximum measured memory usage during a defined workload. Always state what is measured, such as process RSS, accelerator allocation, framework allocator peak, or tensor payload estimate.

Pointwise convolution

A convolution with a 1 x 1 spatial kernel, commonly used to mix channels after depthwise convolution.

Post-training quantization (PTQ)

Quantizing a trained model without fully retraining it. PTQ often uses calibration data to estimate activation ranges. See Chapter 11.

Pruning

Removing or suppressing model parameters or structures judged unnecessary for the target objective. See Chapters 4 through 7.

Prosody

Speech properties such as rhythm, stress, intonation, timing, and pitch contour. For speech models, prosody may be a separate quality dimension from intelligibility or speaker identity.

Q

QAT

Quantization-aware training. Training or fine-tuning a model while simulating quantized numerical behavior so the model can adapt to expected quantization error.

QNN

Qualcomm's AI software stack and the name used by ONNX Runtime for its Qualcomm QNN Execution Provider. Supported behavior depends on device, SDK, operator set, and model configuration. See Chapter 13.

Quantization

Representing weights or activations with lower-precision numerical formats. Quantization can reduce storage and bandwidth, but quality and speed effects depend on calibration, operator support, hardware kernels, and where conversions occur. See Chapter 11.

Quantization scale

A numerical factor used to map between real-valued and integer representations in quantized inference.

Query, key, value (Q/K/V)

The three projected representations used by attention. Queries are compared with keys to produce attention weights that combine values.

R

Receptive field

The region of an input that can influence a given output. In temporal or spatial chunking, the required overlap depends partly on the model's effective receptive field.

ReLU

Rectified Linear Unit. An activation function defined as the maximum of zero and the input. It is inexpensive and widely supported by inference runtimes.

Residual connection

A path that adds or otherwise combines an earlier representation with the output of a later block. Residual connections improve trainability but can extend activation lifetimes because earlier tensors must remain available until the merge operation.

RMSNorm

Root Mean Square Normalization. A normalization method that scales activations using their root mean square without subtracting the mean. It is used in several modern transformer families.

RSS

Resident Set Size. An operating-system measure of the physical memory currently resident for a process. RSS is useful but does not directly equal model activation memory.

Runtime

Software that loads and executes an exported model, such as ONNX Runtime, TensorFlow Lite, ExecuTorch, Core ML, or ncnn.

S

Scheduler

In diffusion systems, the algorithm that defines how denoising timesteps and updates are traversed during sampling. Changing the scheduler or number of steps can alter quality and runtime without modifying model weights. See Chapter 12.

Self-attention

Attention in which queries, keys, and values are derived from the same sequence.

SiLU

Sigmoid Linear Unit, also called the swish activation in a common parameterization. It computes the input multiplied by its sigmoid and is widely used in convolutional and diffusion architectures.

SIMD

Single Instruction, Multiple Data. A CPU execution model in which one instruction operates on multiple values in parallel. Efficient mobile inference libraries depend heavily on SIMD vectorization.

Softmax

A function that converts a vector of values into normalized positive weights that sum to one. It is commonly used to normalize attention scores or class logits.

Sparsity

The fraction or pattern of zero or removed values in a tensor. Unstructured sparsity does not guarantee speedup unless the runtime and hardware use sparse kernels effectively.

Speaker embedding

A compact vector intended to represent speaker identity or voice characteristics. In multi-speaker speech generation, the same acoustic model can often be conditioned on different speaker embeddings.

Static shape

A tensor shape fully known at export or compile time. Static shapes can enable stronger memory planning or accelerator compilation, at the cost of flexibility.

Structured pruning

Pruning that physically removes complete structures such as channels, heads, neurons, or blocks so tensor dimensions become smaller. This is often more useful for mobile deployment than merely inserting zeros. See Chapters 4 through 7.

SVD

Singular Value Decomposition. A matrix decomposition that can be used to construct low-rank approximations of dense matrices. See Chapter 8.

T

Tensor

A multidimensional array with a shape and dtype. Model parameters, inputs, outputs, activations, caches, and intermediate values are represented as tensors.

Tensor shape

The ordered dimensions of a tensor, such as `[batch, sequence, hidden]` or `[batch, channels, height, width]`. Shape is one of the strongest determinants of activation memory and operator cost.

TFLite

TensorFlow Lite, now commonly associated with Google's LiteRT ecosystem, is a mobile and edge inference stack for TensorFlow-compatible models and related deployment workflows.

Throughput

The amount of work completed per unit time, such as images per second, tokens per second, or audio seconds processed per second. Throughput and single-request latency are different metrics.

Token

A discrete unit processed by many language and multimodal models. A token may represent a word fragment, character, code unit, image patch index, or another learned symbol depending on the tokenizer and model.

Tokenizer

Software that maps raw input text or another discrete representation to token IDs expected by a model. A checkpoint and tokenizer must be kept compatible.

U

Unstructured pruning

Pruning individual scalar weights without changing the physical tensor shape. It can create many zeros but may not reduce latency or memory unless sparse storage and sparse kernels are used.

Upsampling

Increasing spatial, temporal, or latent resolution. Upsampling stages can create large activations, especially in image, video, and audio decoders.

V

VAE

Variational Autoencoder. In latent diffusion systems, a VAE commonly encodes images or video frames into lower-dimensional latent tensors and decodes generated latents back to pixel space.

Validation set

A fixed collection of representative examples used to measure whether a model change preserves required behavior. It should not be optimized against so aggressively that it stops representing unseen production cases. See Chapters 1 and 3.

Vectorization

Organizing computation so a processor can operate on several values per instruction. Tensor dimensions aligned with efficient vector widths can materially affect CPU performance.

Vocoder

A model or signal-processing component that converts an acoustic representation such as a mel spectrogram into a waveform. In TTS pipelines, the vocoder can be a major source of latency and quality sensitivity.

W

Weight

A learned parameter used by an operation such as a linear projection or convolution. Weight storage is usually easy to estimate from parameter count and dtype, but runtime memory includes more than weights.

Weight-only quantization

Quantization in which model weights use a reduced precision such as INT8 or INT4 while activations remain in floating point or another format. It is common for large language models and some matrix-heavy architectures.

Working set

The memory actively required during a workload, including resident parameters, live activations, caches, temporary workspace, runtime allocations, and relevant application buffers.

X

XNNPACK

A library of optimized neural network operators for CPU execution, especially on ARM and mobile-class processors. Performance depends on operator, shape, dtype, threading, and integration in the chosen runtime.

Practical lookup by engineering problem

If you are investigating...	Start with these terms
Model file is too large	Parameter count, dtype, quantization, weight-only quantization, low-rank factorization
Peak RAM is too high	Activation memory, tensor shape, buffer reuse, chunking, receptive field, working set

If you are investigating...	Start with these terms
Transformer is too slow	FFN, attention head, hidden size, GEMM, Q/K/V, KV cache, structured pruning
CNN is too slow	Channel, convolution, depthwise convolution, pointwise convolution, structured pruning
Diffusion is too slow	Diffusion model, scheduler, VAE, resolution, quantization, distillation
Android NPU is not accelerating	Execution Provider, operator, graph partitioning, QNN, dynamic shape, fused operator
INT8 quality is poor	Calibration data, quantization scale, PTQ, QAT, mixed precision
TTS quality changed after compression	Prosody, speaker embedding, vocoder, validation set, mixed precision
Long input causes OOM	Chunking, receptive field, activation memory, dynamic shape, KV cache
FLOPs fell but latency did not	Kernel, memory bandwidth, operator fusion, vectorization, execution provider

Appendix D. HDemucs: a concrete refactoring experiment

Evidence status: proposed experiment plan. This appendix applies the handbook to the original hybrid Demucs family. No HDemucs checkpoint was pruned, retrained, exported, or benchmarked while preparing this PDF. The observations below come from the referenced implementation. The proposed measurements and acceptance decisions are work for the experimenter.

D.1 Identify the model before changing it

HDemucs and HTDemucs are different architectures. The original hybrid design combines waveform and spectrogram processing. The repository also distributes a later hybrid Transformer family. A command that loads the default checkpoint is therefore not an adequate specification of an HDemucs experiment. The published model list identifies `hdemucs_mmi` as a hybrid baseline, while the inspected loader defaults to `htdemucs`. [R51, R52, R53]

Record the repository commit, installed package versions, checkpoint digest, model class, source names, sample rate, channel count, and every member of a model ensemble. The following inspection template is not a tested environment or a pinned installation recipe. Run it only after installing your chosen, trusted revision and recording its dependencies.

PYTHON

```python
# Proposed inspection template; not executed for this appendix.
from demucs.apply import BagOfModels
from demucs.hdemucs import HDemucs
from demucs.pretrained import get_model

loaded = get_model("hdemucs_mmi")
members = list(loaded.models) if isinstance(
    loaded, BagOfModels
) else [loaded]

for index, model in enumerate(members):
    if not isinstance(model, HDemucs):
        raise TypeError(f"Unexpected model: {type(model).__name__}")
    print({
        "member": index,
        "class": type(model).__name__,
        "sources": model.sources,
        "sample_rate": model.samplerate,
        "audio_channels": model.audio_channels,
        "parameters": sum(p.numel() for p in model.parameters()),
    })
```

A successful inspection verifies identity and metadata. It does not demonstrate separation quality, runtime compatibility, or mobile feasibility. Inspect any ensemble rather than silently replacing it with its first member. The loader and inference wrapper explicitly support model bags. [R53, R54]

D.2 Do not apply the residual-block recipe to an encoder stage

The inspected HDemucs implementation has frequency and time encoders, corresponding decoders, skip connections, and a point where the branches meet. Stages change dimensions and resolution. Deleting `encoder[2]` is not equivalent to removing a shape-preserving residual block. The decoder and branch alignment can become invalid. The source also warns that changing the FFT size requires coordinated shape calculations. **[R55]**

For an initial experiment, leave the input contract, stage count, branch alignment, FFT settings, source ordering, and output reconstruction unchanged. Profile the actual checkpoint before choosing a transformation.

Inside supported stages, the implementation can contain `DConv` modules. Their internal residual steps preserve the external channel count and are a more localized place to investigate ablation. The inspected forward loop adds each step's output to its input. Replacing the step with `Identity` would therefore double that input, not remove the residual contribution. Removing a step requires rebuilding the step list or returning a zero residual, followed by verification. **[R56]**

Proposed selection procedure: inspect which residual modules exist, measure their costs, ablate one compatible step on a copied model, and evaluate that candidate before recovery training. Keep the original checkpoint immutable. Reconstruction must encode the changed structure, not just save a state dictionary against the original architecture.

D.3 Separate inference settings from architecture changes

Start with the existing inference wrapper. It exposes splitting, segment duration, overlap, random shifts, and worker configuration. In the inspected code, `shifts=0` disables random-shift augmentation. Reducing repeated evaluations and reducing segment length are different experiments. Neither operation changes the stored parameter count. **[R54]**

The following order is a proposed investigation sequence, not a measured ranking. Start from your pinned baseline and change one variable per comparison.

Experiment	Change under test	Keep fixed	Required check
A. Segment length	Test a shorter supported segment	Weights, overlap, shifts, threads	Peak memory, boundary artifacts, separation quality
B. Shift count	Reduce augmentation passes when enabled	Segment length, weights, overlap	Total runtime and quality
C. Residual step	Remove one compatible internal step	Stage interfaces and preprocessing	Shape validity and quality before and after recovery
D. Internal width	Rebuild one dependency group with fewer channels	Source contract and evaluation data	Coupled dimensions, quality, deployed latency
E. Precision	Quantize a supported subgraph	Accepted architecture and inputs	Numerical parity, actual kernels, audio artifacts

Do not interpret the `--two-stems` output option as a two-source architecture. In the inspected command-line implementation, source selection occurs after separation. Similarly, `--int24` controls the saved audio representation, not neural-network weight precision. These flags do not establish a smaller inference model. **[R57]**

D.4 Account for memory outside the network

The splitting implementation allocates an output tensor for the full requested duration. Shorter segments can reduce the size of an individual model evaluation without making the complete application constant-memory. [R54]

For a hypothetical four-source, two-channel output lasting 20 minutes at 44,100 samples per second, the FP32 payload alone is:

EXAMPLE

```
4 sources * 2 channels * 1,200 seconds * 44,100 samples/second
          * 4 bytes/sample
= 1,693,440,000 bytes
= approximately 1.69 GB, or 1.58 GiB
```

This is tensor-size arithmetic, not a measured HDemucs allocation. It excludes the mixture, weights, activations, workspaces, overlap buffers, and application memory.

For a proposed bounded-memory application, read bounded input windows, keep only the overlap region still needed for blending, and write finalized output progressively. Validate its waveform output against the reference wrapper. Preserve normalization, padding, weighting, and boundary conventions. Chunked processing is not automatically equivalent to causal streaming or full-track inference.

D.5 Treat resolution changes as new model experiments

Do not relabel 44.1 kHz samples as 16 kHz input. Do not reduce `nfft` and assume the existing checkpoint still represents the same learned operation. A lower-rate student requires an explicit resampling policy, compatible architecture, recovery or new training, and a fresh evaluation contract. Read the actual checkpoint metadata rather than assuming constructor defaults describe every release. [R55]

For a mobile export experiment, test the spectral frontend and reconstruction separately. The repository's spectral wrapper uses a Hann window, centered transforms, normalization, and complex-valued STFT output. A native DSP replacement must match the selected implementation's conventions before it replaces those operations in a larger pipeline. [R58]

Proposed verification: compare frontend tensors on silence, an impulse, tones, short clips, and representative mixtures. Compare reconstruction lengths and numerical error before evaluating end-to-end audio quality. Use declared tolerances and preserve the reference implementation for rollback.

D.6 Design an evaluation that can reject the optimization

The original research concerns music-source separation. Its evidence does not establish accurate dialogue, music, and sound-effect separation for cartoons or films. Define the intended sources and evaluate that domain independently. [R51]

For this proposed study, partition evaluation data by complete recording or source identity before extracting segments. Keep recovery-training data separate from the final test set. Where isolated references exist, report a specified separation metric per source and describe its implementation, aggregation, and handling of silent references. Without clean references, do not fabricate an objective reference-based quality score.

Add blinded listening comparisons for leakage, missing content, transients, stereo changes, and boundary clicks. Record failures as well as averages. Pair the quality report with the target-device configuration, complete-job time, warm inference latency, and a precisely named peak memory measurement. Compare candidates under the same workload and operating conditions.

Acceptance decision: retain a change only when it satisfies the previously declared quality and resource requirements. A successful export, a smaller checkpoint, or a passing shape test is insufficient. No compression ratio or Android speedup is asserted by this appendix.

References

Primary sources are listed below with the claims they support. Official documentation can change after the recorded access date. Pin the actual dependency and architecture revisions before reproducing a production experiment. A research reference describes the authors' evaluated conditions, not a guarantee for another checkpoint.

R01. Pruning Tutorial

PyTorch contributors. Official tutorial. Accessed 2026-09-24.

Supports: Mask-based pruning and removal of reparameterization; mutable tutorial.

Primary source

R02. Torch-Pruning

Gongfan Fang and Torch-Pruning contributors. Primary implementation. Accessed 2026-09-24.

Supports: Dependency-aware structural pruning; pin a commit before using the package.

Primary source

R03. DepGraph: Towards Any Structural Pruning

Fang et al.. Research paper, 2023. Accessed 2026-09-24.

Supports: Coupled parameter groups across neural network architectures.

Primary source

R04. Knowledge Distillation Tutorial

PyTorch contributors. Official tutorial. Accessed 2026-09-24.

Supports: Teacher-student training examples; tutorial is not a guarantee of recovery.

Primary source

R05. Distilling the Knowledge in a Neural Network

Hinton, Vinyals and Dean. Research paper, 2015. Accessed 2026-09-24.

Supports: Soft targets and temperature-based distillation.

Primary source

R06. Reducing Transformer Depth on Demand with Structured Dropout

Fan, Grave and Joulin. Research paper, 2019. Accessed 2026-09-24.

Supports: LayerDrop is a training method, not evidence for arbitrary layer deletion.

Primary source

R07. Pruning Convolutional Neural Networks for Resource Efficient Inference

Molchanov et al.. Research paper, ICLR 2017. Accessed 2026-09-24.

Supports: First-order sensitivity and iterative pruning in studied CNN tasks.

Primary source

R08. NetAdapt: Platform-Aware Neural Network Adaptation for Mobile Applications

Yang et al.. Research paper, ECCV 2018. Accessed 2026-09-24.

Supports: Direct platform measurements instead of relying only on operation counts.

Primary source

R09. torch.profiler

PyTorch contributors. Official documentation, 2.10. Accessed 2026-09-24.

Supports: Operator profiling, shapes, memory events and instrumentation overhead.

Primary source

R10. Reproducibility

PyTorch contributors. Official documentation, 2.10. Accessed 2026-09-24.

Supports: Determinism controls and limits across platforms and versions.

Primary source

R11. torch.linalg.svd

PyTorch contributors. Official documentation, 2.10. Accessed 2026-09-24.

Supports: Singular-value decomposition API and numerical caveats.

Primary source

R12. inference_mode

PyTorch contributors. Official documentation, 2.10. Accessed 2026-09-24.

Supports: Inference-mode behavior; evaluation mode remains separate.

Primary source

R13. torch.onnx

PyTorch contributors. Official documentation, 2.10. Accessed 2026-09-24.

Supports: Export options, shape constraints, verification and external weights.

Primary source

R14. Quantize ONNX models

ONNX Runtime contributors. Official documentation. Accessed 2026-09-24.

Supports: Static/dynamic quantization, QDQ, calibration and operator-specific INT4 support.

Primary source

R15. Graph optimizations

ONNX Runtime contributors. Official documentation. Accessed 2026-09-24.

Supports: Graph optimization levels and semantics-preserving rewrites.

Primary source

R16. Deploy on mobile

ONNX Runtime contributors. Official documentation. Accessed 2026-09-24.

Supports: Mobile execution-provider selection and device benchmarking.

Primary source

R17. Execution Providers

ONNX Runtime contributors. Official documentation. Accessed 2026-09-24.

Supports: Backend selection and allocation of supported subgraphs.

Primary source

R18. Qualcomm QNN Execution Provider

ONNX Runtime contributors. Official documentation. Accessed 2026-09-24.

Supports: Supported operators, static shapes and backend-specific quantization requirements.

Primary source

R19. Reduced operator config file

ONNX Runtime contributors. Official documentation. Accessed 2026-09-24.

Supports: Selective runtime builds using required operator/type lists.

Primary source

R20. ExecuTorch documentation

PyTorch contributors. Official documentation. Accessed 2026-09-24.

Supports: Edge deployment and supported hardware backends; mutable stable documentation.

Primary source

R21. Overview of memory management

Android Developers. Official documentation. Accessed 2026-09-24.

Supports: Managed heap, memory pressure and proportional set size.

Primary source

R22. dumpsys

Android Developers. Official documentation. Accessed 2026-09-24.

Supports: Inspecting Android process memory; snapshots are not continuous peak traces.

Primary source

R23. Thermal API

Android Developers. Official documentation. Accessed 2026-09-24.

Supports: Thermal status and headroom signals for adapting workload.

Primary source

R24. Support for long-running workers

Android Developers. Official documentation. Accessed 2026-09-24.

Supports: Long-running work, foreground execution and platform-version constraints.

Primary source

R25. Are Sixteen Heads Really Better than One?

Michel, Levy and Neubig. Research paper, 2019. Accessed 2026-09-24.

Supports: Attention-head importance and pruning in evaluated Transformer tasks.

Primary source

R26. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness

Dao et al.. Research paper, 2022. Accessed 2026-09-24.

Supports: Exact attention with an IO-aware implementation; not an Android guarantee.

Primary source

R27. scaled_dot_product_attention

PyTorch contributors. Official documentation, 2.10. Accessed 2026-09-24.

Supports: Kernel dispatch, dropout behavior and supported attention implementations.

Primary source

R28. GroupNorm

PyTorch contributors. Official documentation, 2.10. Accessed 2026-09-24.

Supports: Groupwise input statistics, including evaluation-time behavior.

Primary source

R29. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications

Howard et al.. Research paper, 2017. Accessed 2026-09-24.

Supports: Depthwise separable convolution and width/resolution tradeoffs.

Primary source

R30. LoRA: Low-Rank Adaptation of Large Language Models

Hu et al.. Research paper, 2021. Accessed 2026-09-24.

Supports: Low-rank updates to a retained pretrained weight matrix.

Primary source

R31. High-Resolution Image Synthesis with Latent Diffusion Models

Rombach et al.. Research paper, CVPR 2022. Accessed 2026-09-24.

Supports: Latent autoencoder and denoising pipeline architecture.

Primary source

R32. Scalable Diffusion Models with Transformers

Peebles and Xie. Research paper, 2023 version. Accessed 2026-09-24.

Supports: Diffusion Transformer depth, width and latent patch-token architecture.

Primary source

R33. P.808: Subjective evaluation of speech quality with a crowdsourcing approach

ITU-T. Recommendation. Accessed 2026-09-24.

Supports: Speech listening-study methodology; not a universal voice-identity metric.

Primary source

R34. Reduce memory usage

Hugging Face Diffusers contributors. Official documentation. Accessed 2026-09-24.

Supports: VAE tiling/slicing, offloading, limitations and tradeoffs.

Primary source

R35. Stable Video Diffusion

Hugging Face Diffusers contributors. Official documentation. Accessed 2026-09-24.

Supports: Feed-forward chunking, frame decoding and possible flicker.

Primary source

R36. torch.utils.checkpoint

PyTorch contributors. Official documentation, 2.10. Accessed 2026-09-24.

Supports: Activation checkpointing exchanges training memory for recomputation.

Primary source

R37. Common pitfalls and recommended practices

scikit-learn contributors. Official documentation. Accessed 2026-09-24.

Supports: Data leakage and separation of fitting from evaluation.

Primary source

R38. Progressive Distillation for Fast Sampling of Diffusion Models

Salimans and Ho. Research paper, ICLR 2022. Accessed 2026-09-24.

Supports: Learning to replace a multi-step sampler with fewer learned steps.

Primary source

R39. Consistency Models

Song et al.. Research paper, 2023. Accessed 2026-09-24.

Supports: Consistency training and distillation for few-step generation.

Primary source

R40. Q-Diffusion: Quantizing Diffusion Models

Li et al.. Research paper, 2023. Accessed 2026-09-24.

Supports: Diffusion-specific calibration and quantization of studied denoisers.

Primary source

R41. Structural Pruning for Diffusion Models

Fang, Ma and Wang. Research paper, NeurIPS 2023. Accessed 2026-09-24.

Supports: Time-aware importance selection and structural diffusion pruning.

Primary source

R42. DPMSolverMultistepScheduler

Hugging Face Diffusers contributors. Official documentation. Accessed 2026-09-24.

Supports: Prediction types, schedules and solver configuration.

Primary source

R43. VBench: Comprehensive Benchmark Suite for Video Generative Models

Huang et al.. Research paper. Accessed 2026-09-24.

Supports: Separate temporal, spatial and semantic video-evaluation dimensions.

Primary source

R44. Reproducibility

Hugging Face Diffusers contributors. Official documentation. Accessed 2026-09-24.

Supports: Random-generator state and limits of repeated seeded generation.

Primary source

R45. On Aliased Resizing and Surprising Subtleties in GAN Evaluation

Parmar, Zhang and Zhu. Research paper, CVPR 2022. Accessed 2026-09-24.

Supports: Resizing and compression can alter FID comparisons.

Primary source

R46. Cache strategies

Hugging Face Transformers contributors. Official documentation. Accessed 2026-09-24.

Supports: Static, dynamic, offloaded and quantized cache tradeoffs.

Primary source

R47. Conv2d

PyTorch contributors. Official documentation, 2.10. Accessed 2026-09-24.

Supports: Convolution shape and group constraints.

Primary source

R48. Accelerate inference

Hugging Face Diffusers contributors. Official documentation. Accessed 2026-09-24.

Supports: Precision, compilation and memory-efficient attention guidance.

Primary source

R49. ONNX Intermediate Representation Specification

ONNX contributors. Official specification. Accessed 2026-09-24.

Supports: Graph representation, operators and tensor types; not runtime kernel guarantees.

Primary source

R50. Tensor Views

PyTorch contributors. Official documentation, 2.10. Accessed 2026-09-24.

Supports: Storage sharing and non-contiguous views.

Primary source

R51. Hybrid Spectrogram and Waveform Source Separation

Alexandre Defossez. Original research paper, arXiv:2111.03600. Accessed 2026-09-24.

Supports: The original hybrid waveform and spectrogram architecture and the music-separation scope of its evaluation. Results are not reproduced in this appendix.

Primary source

R52. Demucs model descriptions

Meta / Facebook Research. Official repository README. Accessed 2026-09-24.

Supports: Distinguishing `hdemucs_mmi` from the `htdemucs` family in the published model list.

Primary source

R53. Demucs pretrained model loader

Meta / Facebook Research. Official `demucs/pretrained.py` implementation, inspected `main` snapshot. Accessed 2026-09-24.

Supports: Model loading, model-bag resolution, source labels, and the inspected default model name. Pin the selected revision before execution.

Primary source

R54. Demucs inference wrapper

Meta / Facebook Research. Official `demucs/apply.py` implementation, inspected `main` snapshot. Accessed 2026-09-24.

Supports: Segment splitting, overlap weighting, shift augmentation, ensemble evaluation, and full-duration output allocation. Pin the selected revision before execution.

Primary source

R55. HDemucs architecture implementation

Meta / Facebook Research. Official `demucs/hdemucs.py` implementation, inspected `main` snapshot. Accessed 2026-09-24.

Supports: Coupled time and frequency branches, encoder and decoder interfaces, FFT-dependent shapes, and reconstruction behavior. Constructor defaults are not a substitute for checkpoint metadata.

Primary source

R56. DConv residual implementation

Meta / Facebook Research. Official `demucs/demucs.py` implementation, inspected `main` snapshot. Accessed 2026-09-24.

Supports: Shape-preserving internal residual steps and additive forward behavior. No pruning-quality result is claimed.

Primary source

R57. Demucs command-line separation implementation

Meta / Facebook Research. Official `demucs/separate.py` implementation, inspected `main` snapshot. Accessed 2026-09-24.

Supports: Source-output selection after separation and the distinction between saved-audio precision and model precision.

Primary source

R58. Demucs spectral transform wrapper

Meta / Facebook Research. Official `demucs/spec.py` implementation, inspected `main` snapshot. Accessed 2026-09-24.

Supports: STFT and inverse-STFT windowing, normalization, centering, and complex representation conventions.

Primary source
