# Control, Not Content: Input-Dependent Weight Generation as a Neural Control Plane

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Hypernetworks are usually described by how many parameters generate an update; we argue that the decisive quantity is what the generated update can make its host do. For fixed low-rank factors, diagonal gains span at most an $r$-dimensional weight family regardless of controller size. Across language control, audio, vision, video, affect, and audiovisual reasoning, this bound predicts a repeatable division: generated gains reliably carry persistent policy, but instance evidence requires a structure-preserving signal, a direct objective, and a compatible actuator. A scalar length dial obtains an $11.13\times$ conditioning-to-prompt-variance ratio ($0.09\times$ shuffled), while structured audio and a counterfactual loss raise order accuracy from $49.75\%$ to $92.25\%$. On official AVHBench, repairing dense-core initialization improves a fully trainable hNET by $8.66$ points at identical size; explicit cross-modal contrasts then raise macro balanced accuracy from $78.11 \pm 0.67\%$ to $82.85 \pm 1.43\%$ across five paired seeds (cluster interval $[+2.50, +6.95]$). A synthetic recurrent/attention hybrid sharpens the mechanism: generated rank-8 transforms reach $83.84 \pm 4.65\%$ on unseen ordered programs versus $12.68 \pm 2.68\%$ for parameter-matched FiLM. We then transfer the comparison to WikiText-103. In an exact eight-layer topology, the hNET beats matched Gated DeltaNet (GDN), FiLM, and static low-rank controls across three seeds at two million tokens while using $3.6\%$ fewer parameters than GDN. At 20 million tokens it remains better than FiLM but GDN leads by $0.041$ NLL. Increasing the hNET's generated recurrent core from 64 to 576 coefficients makes controller shuffling ten times more destructive yet does not improve NLL. The intervention separates two resources that parameter count conflates: generated weights are a high-leverage *control plane*; recurrent associative state is a *content memory*. Competitive adaptive architectures should compose them, not ask either to impersonate the other.

## 1 Introduction

Hypernetworks turn one neural artifact into another: a condition is mapped to weights, adapter coefficients, or an update applied to a host network [Ha et al., 2017]. This makes them attractive as compact "all-rounders." Instead of storing separate audio, visual, audiovisual, style, and control experts, one shared controller might synthesize the appropriate adaptation for each input. The premise matters beyond storage. Per-example parameter generation can place heterogeneous conditions in one backbone batch, and a weight-space intervention persists without consuming context tokens.

There is, however, an ambiguity in the phrase *generated weights*. A large controller can emit only a few scalars. Those scalars may select among behaviors already latent in a frozen host, or they may be expected to transmit instance-level sensory evidence. These are different functions, but conventional

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

parameter counts conflate them. The distinction is especially important for gain-conditioned LoRA, where an input-dependent vector rescales fixed low-rank directions. Such a mechanism may be a strong control interface while remaining a poor information interface.

This paper asks a mechanism question: *what kind of channel is created when an input controls neural-network weights?* We answer it using a formal reachable-set observation, controlled interventions across modalities, a preregistered actuator-by-demand study, prospective repairs, and a matched comparison with Gated DeltaNet (GDN). We do not pool heterogeneous tasks into a leaderboard or claim universal architectural superiority. Instead, we identify when generated weights behave as policy, evidence transport, programmable computation, or an inadequate substitute for content memory.

Our contributions are:

1. We formalize the reachable update family of diagonal gain conditioning. For fixed LoRA factors at rank $r$, no controller can make the family exceed $r$ linear degrees of freedom; dense and rotational cores expand this family in specific, bounded ways.
2. We introduce a *signal–objective–actuator* diagnosis. Direct probes and targeted repairs localize failures that controller scaling cannot resolve.
3. We provide a corrected cross-domain intervention record. Low-dimensional persistent controls succeed, whereas instance and relational information require structured signals, direct objectives, or a better-aligned host interface.
4. We run a preregistered, parameter-matched five-actuator study on AVHBench [Kim et al., 2025]. Its predicted monotone interaction is falsified, but it exposes a robust dissociation: every actuator carries a one-bit policy, while token-space prefixes transport unimodal evidence more reliably than diagonal gains, dense cores, or positional attention bias.
5. We causally intervene on expert-weight geometry. Raw directions are seed-unstable and layer permutation is harmless. We then repair a zero-gradient dense-core initialization: the repaired fully trainable system gains 8.66 points over the standard parameterization at identical size, exposing optimization geometry as part of channel capacity.
6. We repair the remaining AVHBench signal bottleneck with explicit audio–visual contrast features. The same dense hNET improves from $78.11\%$ to $82.85\%$, wins all five paired seeds, and retains large shuffled-condition effects.
7. We give a constructive actuator separation in a sparse-attention architecture. Generated dense cores solve and transfer ordered programs across length, while parameter-matched scalar gains do not; correct-condition shuffling removes the effect.
8. We transfer this mechanism to WikiText-103 and compare exact-topology GDN, hNET, FiLM, and static controls. The hNET wins the matched two-million-token study, remains causally condition-dependent at 20 million tokens, and reveals why GDN later leads: strengthening generated control greatly increases intervention sensitivity without adding the associative content state that language modeling rewards.

"Low bandwidth" is operational rather than Shannon-theoretic. It denotes an interface that reliably carries a small number of behaviorally useful control variables but loses instance-level or relational information unless the representation, loss, and actuator are aligned with that information. The title is therefore a default and a diagnostic, not a universal impossibility claim.

# 2 Conditional adaptation as a channel

## 2.1 Reachable weight updates

Consider a frozen linear map with a condition-dependent low-rank update

$$\Delta W(c) = B\,\mathrm{diag}(g(c))\,A = \sum_{i=1}^{r} g_i(c) b_i a_i^\top, \tag{1}$$

where $A \in \mathbb{R}^{r \times d_{\mathrm{in}}}$ and $B \in \mathbb{R}^{d_{\mathrm{out}} \times r}$ are fixed after training, $a_i^\top$ and $b_i$ are their rows and columns, and a hypernetwork maps condition $c$ to $g(c) \in \mathbb{R}^r$.

Table 1: Actuator families. Degrees are condition-dependent coordinates per adapted module (prefix coordinates are global).

| Interface | Degrees | Host intervention | Natural role / limitation |
|---|---|---|---|
| Diagonal gains | $r$ | Rescale fixed rank-one atoms | Persistent policy; no rank-channel mixing |
| Dense core | $r^2$ | Mix fixed $A/B$ channels | More directions; alignment is not guaranteed |
| Cayley core | $r(r-1)/2$ | Rotate rank channels | Direction without scale explosion; task-sensitive |
| Soft prefix | $kd$ | Add token values to attention | Evidence transport; uses context/KV positions |
| ARAM-style bias | layers $\times$ bases | Alter pre-softmax routing | Temporal/routing control; no new token values |

**Proposition 1** (Reachable-set bound)**.** *For fixed $A$ and $B$, the family $\{B\,\mathrm{diag}(g)A : g \in \mathbb{R}^r\}$ lies in a linear subspace of dimension at most $r$, independently of the architecture or parameter count of the controller that produces $g$. Replacing the diagonal with an unconstrained core $C \in \mathbb{R}^{r\times r}$ gives dimension at most $r^2$. Restricting $C$ to a Cayley-parameterized orthogonal transformation supplies at most $r(r-1)/2$ local degrees of freedom.*

The proof is immediate from the fixed outer-product expansion in Eq. 1; a complete proof is in Appendix A. The statement is modest but consequential. At rank one, every nonzero update is collinear. Making the controller deeper changes the partition of condition space and the coefficient assigned to the one direction; it cannot create another direction. Conversely, a small off-diagonal or rotational coordinate can matter when the missing behavior requires reorienting existing rank channels.

This is a bound on an *actuator family*, not on the information present in $c$, not on mutual information, and not on the nonlinear function computed by the full host. Fixed atoms can have broad effects after composition through many transformer layers. The proposition instead rules out the intuition that controller parameter count alone measures the diversity of conditional weight interventions.

### 2.2 Signal, objective, and actuator

We separate three maps that are often collapsed into the word "hypernetwork":

$$x \xrightarrow{\text{signal } z} z(x) \xrightarrow{\text{controller } h} u \xrightarrow{\text{actuator } \mathcal{A}} f_{\theta,\mathcal{A}(u)}. \tag{2}$$

Training additionally supplies an objective $\mathcal{L}$ that decides which distinctions in $z$ survive into behavior. This yields three diagnoses.

**Signal bottleneck.** Pooling may erase order or binding before the controller sees it. If a direct probe cannot recover the target from $z$, increasing controller width or actuator rank is not a valid repair.

**Objective bottleneck.** A distinction may be present in $z$ yet irrelevant to next-token likelihood. A matched counterfactual loss can test whether the ordinary objective simply failed to reward condition dependence.

**Actuator bottleneck.** The signal and objective may be adequate while the host interface exposes the wrong transformations. Diagonal gains, dense cores, soft tokens, and attention-logit biases intervene at different places and should be compared with matched information and trainable budgets.

Table 1 summarizes the interfaces evaluated in this record. Static gains are broadcast across generation and therefore behave as persistent policy; per-token gains form a control schedule. Neither property implies faithful transmission of sensory content.

Table 2: Cross-domain mechanism record. "Repair" identifies the component changed after a failure. BAcc denotes balanced accuracy.

| Domain | Controlled result | Diagnosis | Targeted intervention / implication |
|---|---|---|---|
| Length | $11.13\times$ condition/prompt variance; shuffle $0.09\times$ | Scalar persistent policy | 2-D controller retains $11.29\times$ at $\sim 90\times$ fewer conditioning parameters |
| Audio identity | 87.5% random; 68.75% hard; shuffle 46.25% | Coarse identity passes | Rank 1 sufficient for easy selection |
| Audio order | 49.75% ordinary NLL | Signal and objective | Structured representation probes at 99%; counterfactual loss reaches 92.25% |
| Vision | 93.9% random; 62.6% nearest neighbor | Fine evidence plateau | Rank 1,4,16,64: 62.5,62.0,62.6,61.6% |
| Video | 97.2% ordered schedule; 72% generation redirect | Schedule succeeds; gains lossy | Boundary F1: gain 0.850 vs source embedding 0.947 |
| AV binding | Frozen map 50% | Relation absent from map | Direct probe 85.07%; controller+replay 82.21% |
| AV lag | Held-out 19.60%, chance 14.29% | Weak, boundary-biased signal | $>$ 2 s MAE; no composition claim |
| UCF-GZSL | Scalar HM 4.11%; prefix 19.54% | Actuator geometry | Cayley rank 4 HM 13.97%; improvement is task-specific |
| Affect | Warm/cold endpoints distinct; dynamic $+0.08$, CI $[-0.08, +0.25]$ | Latent-register sensitivity test | Endpoint tone is detectable; continuous control is below blind resolution while prompts span 3.45 points |

# 3 Experimental program

## 3.1 Evidence policy

The study is a sequence of controlled experiments rather than a single benchmark sweep. We include a result in the main narrative only when its current artifact supports the stated metric and scope. Each cited test is labeled primary, corrected, or diagnostic in an audited ledger (Appendix D); superseded pilots and metrics without raw support are outside the claim set. Null contrasts are retained only when they identify a bottleneck or distinguish competing mechanisms.

Across experiments, three controls recur: (i) hold the text prompt byte-identical while changing only the condition; (ii) shuffle, reverse, or replace the condition; and (iii) compare the host against a direct probe on the condition representation. Where feasible, protocols were written before the reported runs. Sample counts, preregistration status, seeds, and sources are listed in Appendix D.

## 3.2 Shared host and adapter pattern

The recent audio and audiovisual experiments freeze Qwen3-4B-Instruct-2507 [Yang et al., 2025] and adapt the query, key, value, and output projections in all transformer layers. Binary tasks are scored by the one-token likelihood difference between `Yes` and `No`. AVHBench conditions are frozen 1,024-dimensional PE-AV embeddings [Vyas et al., 2026]; earlier audio experiments use frozen Qwen3-ASR features and vision/video experiments use frozen image representations. Unless a specific intervention changes them, only the condition stem, controller, and adapter parameters train. This design deliberately tests the conditional interface rather than end-to-end sensory learning.

## 3.3 What counts as success

A raw task score is not credited to conditioning unless the correct condition outperforms a shuffled or counterfactual condition. A failed host score is not attributed to the actuator when the signal probe is at chance. A repair is most informative when it changes one component in Eq. 2 and passes its control. This discipline prevents a benchmark prior, a powerful frozen encoder, or a weak metric from being mistaken for a weight-space mechanism.

# 4 Cross-domain intervention map

Table 2 condenses the admissible record. Values from different tasks are not directly comparable as a leaderboard; the point is the intervention pattern.

### 4.1 Positive anchors: control and schedules

**Length.** The clearest success is a requested-length controller. Its output-length range is $11.13\times$ the prompt-induced variance, versus $0.09\times$ under a shuffled condition and $1.32\times$ for prompting. Targets of 30, 60, and 90 words yield 33.75, 60.12, and 88.00 words, with 1.96 words mean absolute bias. Mid-generation reconditioning causes no degeneration in the tested arms. A two-dimensional controller preserves the dial at $11.29\times$ with roughly $90\times$ fewer conditioning parameters. Length is a low-dimensional policy that applies throughout decoding and is directly rewarded, exactly the regime in which gains should work.

**Identity and schedules.** With text held fixed, audio-conditioned discrimination reaches 87.5% on random pairs and 68.75% on hard pairs; shuffling falls to 46.25%. Image-conditioned likelihood reaches 93.9% on random pairs and 62.6% on nearest-neighbor pairs, versus 50.5% when shuffled. Per-token visual schedules recover synthetic temporal order at 97.2%, and changing the condition during decoding redirects 72.0% of continuations versus 16.4% in the control. These are causal successes: a small condition selects or redirects host behavior. They do not show that gains preserve all instance information.

### 4.2 Signal and objective rescue temporal audio

Global pooling makes reversed audio event pairs almost indistinguishable (cosine similarity 0.9983) and order remains at chance. An early/late representation lowers similarity to 0.8890 and supports 99% direct order recovery. Yet ordinary next-token likelihood still yields 49.75% at rank one and 51.75% at rank sixteen. Adding a matched-versus-reversed counterfactual ranking loss raises order accuracy to 92.25%; shuffling restores 51.5%. The two-stage rescue localizes two independent faults: pooling first erased order, then ordinary likelihood ignored the repaired signal. Rank scaling addressed neither.

### 4.3 Information plateaus and relation failures

Fine image discrimination is flat across a 64-fold gain-count sweep: 62.5, 62.0, 62.6, and 61.6% for ranks 1, 4, 16, and 64. Similarly, a learned shot-boundary readout from video gains reaches F1 0.850, below a threshold on the source embeddings at 0.947. More gain coordinates do not recreate distinctions discarded or inaccessible at the interface.

Audiovisual composition separates signal from actuator more directly. A frozen controller remains at 50% AV binding while a classifier on paired representations reaches 85.07%. Updating the controller and fusion layer with replay reaches 82.21% and retains prior tasks within five points. In contrast, the harder seven-way lag probe obtains only 14.81% on seen and 19.60% on held-out clips against 14.29% chance, with more than two seconds mean absolute error and predictions biased toward boundary classes. We therefore count binding as a supervised repair and lag as a null, not as zero-shot relational composition.

### 4.4 Affect as a latent-conditioning stress test

Tone and register provide a deliberately difficult test of the conditioning channel because the relevant pattern is implicit, subjective, and weakly specified by surface text. We ask whether an hNET can identify this hidden stylistic variable and use it as a control signal while ordinary prompt content is held as the positive-control pathway. Fixed warm and cold endpoint adapters are behaviorally distinguishable, establishing that the endpoint signal exists. The corrected blind audit then measures the resolution of dynamic control using 110 individual 1–7 warmth ratings: interpolation through a nominal neutral target changes the rating by $+0.08$, with CI $[-0.08, +0.25]$, whereas prompt content spans 3.45 points. Expanding the prompt-dependent controller to 24,600 parameters does not strengthen the ordering. Thus the experiment bounds what the hNET can extract from a subtle latent register: it detects separated endpoints, but the tested representation and objective do not support a reliable continuous tone coordinate. This is a sensitivity boundary, not evidence that affect is intrinsically uncontrollable.

### 4.5 Geometry and persistence are conditional advantages

On the official UCF-GZSL main split with released SeLaVi features [Mercea et al., 2022], scalar gains reach 4.11% harmonic mean (HM) and 11.54% ZSL. A documented post-hoc checkpoint diagnostic finds effective gain rank 2.28, with 99.85% of variance in the top ten components; the upstream stem has effective rank 9.92 and a linear gain-to-stem map recovers 82.8% of variance. The raw diagnostic tensor was not packaged, so we treat these values as characterization rather than primary evidence. A rank-2 Cayley actuator reaches 8.01% HM/17.29% ZSL and rank 4 reaches 13.97%/13.41%; a shared soft prefix is strongest at 19.54%/20.66%. Thus changing weight-space geometry helps without changing the upstream features, but rotation remains weaker than token-space transport. Transfer back to AVHBench is not monotone: Cayley raises mean AV Matching from 81.85% to 83.23% but lowers a directional hallucination task and reduces macro BAcc from 72.46% to 70.05%, with larger seed variance.

Weight-space conditioning has one real deployment prior: persistence under unseen context length. At 1,024 distractor tokens, frozen gains retain 81.26% BAcc while a frozen prefix falls to 58.50%. One matched long-context training epoch raises the prefix to 80.69%. Gains therefore provide zero-shot persistence because they do not compete for token position; they do not provide an irreducible quality advantage once the prefix is trained for the shift.

## 5 S7: a matched actuator-by-demand test

The cross-domain record suggests that richer information should increasingly favor richer actuators, but it changes datasets and tasks at the same time. We therefore preregistered S7 before inspecting any S7 test label: a single-pipeline comparison holding the condition encoder, frozen host, split, optimizer exposure, and trainable budget fixed.

### 5.1 Design

The population is AVHBench represented by the completed PE-AV cache. A SHA-256 split with seed 424242 assigns videos, not rows, to 70/15/15 train/validation/test partitions. Each seed contains 5,159/1,071/1,164 rows. The official Video-driven Audio Hallucination and Audio-driven Video Hallucination tasks form the *unimodal evidence* stratum; official AV Matching forms the *relation* stratum. A balanced deterministic bit embedded along one fixed direction in condition space requests `Yes` or `No` under a byte-identical prompt and supplies the *scalar policy* stratum. It is a controlled assay, not official AVHBench performance.

Five systems train jointly across the three demands:

1. rank-2 diagonal LoRA gains;
2. a generated dense $2 \times 2$ core in the same rank factors;
3. stateless ARAM-style coefficients over eight fixed relative-position attention-bias bases per layer;
4. one generated soft prefix token shared across demands; and
5. three demand-specific prefix projectors with the same *aggregate* storage budget. This is a consolidation comparator, not a per-task-capacity upper bound.

Native trainable counts are 1.809–1.861M with no inactive padding. Every arm uses frozen Qwen3-4B, frozen 1,024-D PE-AV features, four epochs, batch size 8, AdamW at $2 \times 10^{-4}$, weight decay 0.01, OneCycle cosine scheduling with 8% warmup, and gradient clipping at 1.0. Seeds are 173205, 271828, and 314159. Checkpoints maximize validation macro AUROC across demands. Evaluation reports threshold-zero BAcc and AUROC, correct versus within-demand cyclic-shuffled conditions, wrong-modality conditions, regularized linear probes, and all per-video predictions.

The locked estimand for richer actuator $a$ is

$$I_a = [\mathrm{BAcc}_{\mathrm{relation},a} - \mathrm{BAcc}_{\mathrm{policy},a}] - [\mathrm{BAcc}_{\mathrm{relation,diag}} - \mathrm{BAcc}_{\mathrm{policy,diag}}]. \tag{3}$$

We predicted $I_a > 0$ and required sign agreement plus condition dependence; three seeds were not treated as sufficient for a p-value gate. We report sample standard deviations and video-cluster bootstrap intervals.

Table 3: S7 test BAcc (%, mean ± sample SD over three seeds). Δshuf is normal minus shuffled-condition BAcc. Bold marks the strongest shared all-rounder on each official stratum.

| Actuator | Params | Policy | Evidence | Relation | Δshuf evid. | Δshuf rel. |
|---|---|---|---|---|---|---|
| Diagonal gains | 1.824M | 100.0±0.0 | 63.00±8.39 | 81.19±1.36 | +1.64 | +28.77 |
| Dense core | 1.861M | 100.0±0.0 | 52.73±2.18 | 82.74±1.05 | +3.49 | +30.55 |
| Attention bias | 1.809M | 100.0±0.0 | 54.02±5.03 | 75.98±10.02 | +1.61 | +23.43 |
| Shared prefix | 1.820M | 100.0±0.0 | **72.33±1.09** | **82.95±0.91** | +5.68 | +31.77 |
| Specialist prefixes | 1.825M | 100.0±0.0 | 70.93±0.28 | 82.10±0.98 | +4.28 | +29.99 |

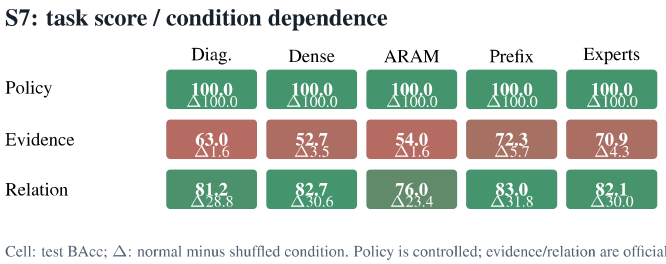


Figure 1: All actuators pass scalar control. Token-space prefix conditioning is the most stable evidence path; relation is already strong in the fused PE-AV signal and separates the arms less.

### 5.2 Results

Every interface learns perfect scalar policy and loses 100 points when the policy direction is reversed. Parameter count and nominal degrees therefore do not distinguish the low-demand task. The official strata do distinguish the interfaces. The shared prefix raises evidence BAcc by 9.33 points over diagonal gains, improves on every paired seed, and reduces seed SD from 8.39 to 1.09 points. It also achieves the strongest relation score, although only 1.76 points above diagonal gains. One shared prefix slightly exceeds three separated paths at the same aggregate storage budget.

The preregistered interaction is *not confirmed*. For the shared prefix, $I = +1.76$ points with paired-seed values $+3.85, +1.49, -0.08$ and video-cluster interval $[-0.66, +4.40]$. Dense cores yield $+1.55$ $[-1.06, +4.65]$; specialist prefixes $+0.90$ $[-1.31, +3.16]$; attention bias $-5.21$ $[-16.83, +2.27]$. No interval excludes zero.

### 5.3 Diagnostics and the falsified simple theory

The frozen input signal probe reaches 100% on policy, 55.68% on evidence, and 78.79% on relation. The weak evidence probe limits attribution: some label-relevant information may be nonlinear, but failure on this stratum cannot be assigned solely to actuator bandwidth. Controller-output linear probes are also diagnostic rather than capacity estimates. The shared-prefix output probe is only 51.18% on evidence while the host reaches 72.33%, indicating that the learned prefix–host interaction performs nonlinear work absent from a linear readout.

Shuffling is decisive for policy and relation but modest for evidence. The prefix has the largest evidence shuffle drop (+5.68), while diagonal gains drop only +1.64; therefore more of the prefix advantage is condition-dependent, but task and language priors remain. A secondary diagonal arm adding a matched-versus-shuffled margin does not rescue evidence (57.56% versus 63.00%) and leaves relation essentially unchanged (81.53% versus 81.19%).

S7 falsifies two convenient stories. First, more actuator degrees do not monotonically improve performance: a dense core selectively helps relation while damaging evidence, and a million-parameter attention-bias controller is unstable. Second, "relation" is not intrinsically the hardest demand. PE-AV already computes AV matching well, so official relation is easier than directional-hallucination evidence for every arm. The surviving variable is *alignment*: what relation the frozen encoder has already computed, and whether the actuator gives the host an appropriate way to use it.

Table 4: S8–S9 held-out results (%, five fresh seeds). Stored counts include fixed factors; trainable counts are optimized parameters. Intervals cluster test rows by video while resampling seeds.

| Weight dictionary | Trainable | Stored | Official macro | Evidence | Relation |
|---|---|---|---|---|---|
| Expert-aligned fixed | 0.387M | 1.861M | 77.41±0.46 | 72.02±1.07 | 82.80±0.46 |
| Expert-aligned trainable | 1.861M | 1.861M | 78.11±0.67 | 73.62±1.05 | 82.60±0.48 |
| Expert-permuted fixed | 0.387M | 1.861M | 77.56±0.63 | 72.33±0.86 | 82.78±0.51 |
| Random fixed | 0.387M | 1.861M | 75.97±0.31 | 69.30±1.48 | 82.65±1.27 |
| Jointly learned dense | 1.861M | 1.861M | 69.45±4.56 | 59.75±8.87 | 79.15±1.47 |

# 6 S8: are expert weights content or coordinates?

S7 shows that a dense core is expressive but hard to train. S8 asks whether trained task experts can supply a better weight-space basis. Before training, a retrospective gate tested the stronger claim that expert directions are reusable across seeds. It failed: a rank-2 row/column basis built from two seeds captures only $0.0117\%$ of a held-out seed's expert-update energy. Same-seed shared-basis coverage weakly predicts expert gaps (Pearson $-0.22$, Spearman $-0.33$). Raw coordinates are therefore not seed-invariant expert content.

We then locked a narrower prospective test. For each of 144 attention projections, three seed-173205 task experts define rank-2 row and column subspaces. Five fresh controller seeds use the S7 data, split, optimizer, and validation rule. Three fixed arms have identical stored and trainable counts: aligned expert factors; the same factors cyclically permuted across layers within projection type; and norm-matched random orthonormal factors. A fresh jointly learned dense arm is secondary. The primary endpoint is the mean BAcc of the two official AVHBench strata; the expert-alignment claim requires at least four of five wins, a three-point mean gain over both controls, intervals excluding zero, positive condition dependence, and at least $95\%$ policy BAcc.

The alignment claim fails. Aligned minus permuted is $-0.15$ points, interval $[-1.25, +0.89]$, with three of five wins. Aligned beats random in all five seeds by $+1.44$ points $[+0.39, +2.42]$, but misses the locked three-point materiality gate. Thus expert-factor statistics help modestly, while layer-specific semantic placement contributes no detectable value.

The secondary parameterization result is large. Aligned fixed beats learned dense by $+7.96$ points $[+4.58, +11.48]$; even random fixed wins by $+6.52$ $[+2.81, +10.50]$. Fixed arms optimize $79.2\%$ fewer adapter parameters at equal deployment storage and attain $100\%$ policy BAcc. But this comparison changes both trainability and the initial gradient path: standard dense LoRA starts with $B = 0$ and an identity core, so its controller initially has zero gradient; fixed arms use nonzero factors and a zero core, retaining exact no-op initialization while making the controller trainable immediately.

## 6.1 S9: isolating the optimization repair

We prospectively isolate these explanations by loading the identical aligned factors and zero core, then allowing the factors to train. This arm has exactly the same 1.861M trainable and stored parameters as standard dense. It reaches $78.11 \pm 0.67\%$, compared with $77.41 \pm 0.46\%$ fixed and $69.45 \pm 4.56\%$ standard dense. Fixed minus trainable is $-0.70$ points $[-1.95, +0.45]$, so the freezing gate fails. Repaired trainable minus standard dense is $+8.66$ points $[+4.71, +12.86]$, positive in all five seeds and passing every locked repair gate. All arms retain $100\%$ policy BAcc.

The repair is therefore the nondegenerate factor/zero-core initialization and its immediate controller gradient, not freezing. The factors can adapt after the controller becomes trainable from step one. Together S8–S9 show that a useful generated artifact can be a coordinate in a stable intervention family without containing transferable expert semantics, and that nominal reachability is insufficient when parameterization blocks the controller's optimization path.

# 7 S10: repairing the multimodal signal

S9 repairs the optimization path but still gives each AVHBench task only one selected PE-AV stream. S10 prospectively tests the signal diagnosis while retaining the rank-2 trainable aligned dense actu-

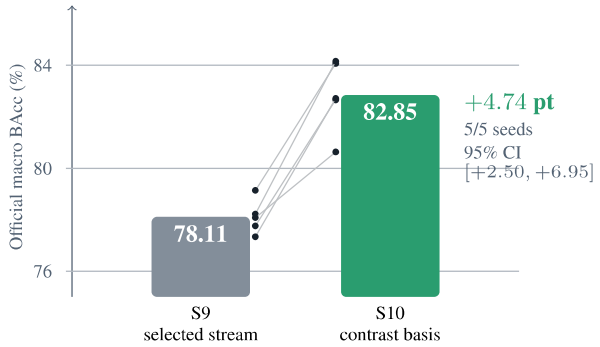


Figure 2: Explicit cross-modal contrasts repair the AVHBench signal supplied to the generated-weight controller. Bars show five-seed means; connected points are paired seeds.

Table 5: S13 ordered-program accuracy (%, 25,600 examples/cell). Dynamic arms show mean ± sample SD over three seeds.

| Arm | Parameters | Seen L6 | Unseen L6 | Seen L12 | Unseen L12 |
|---|---|---|---|---|---|
| Independent hybrid | 287,359 | 100.00 | 30.80 | 99.58 | 30.76 |
| Shared hybrid | 156,607 | 99.97 | 22.34 | 66.29 | 14.95 |
| Sequential FiLM | 231,295 | $23.05 \pm 0.24$ | $12.68 \pm 2.68$ | $13.97 \pm 0.05$ | $7.73 \pm 0.88$ |
| Sequential dense-core hNET | 229,727 | $\mathbf{100.00 \pm 0.00}$ | $\mathbf{83.84 \pm 4.65}$ | $\mathbf{99.21 \pm 0.62}$ | $\mathbf{75.27 \pm 5.21}$ |

ator. The controller instead receives the explicit cross-modal basis

$$z_{\text{contrast}} = [z_a, z_v, z_{av}, z_a - z_v, z_a \odot z_v], \tag{4}$$

which exposes identity, difference, and agreement before weight generation. A validation-only screen selected this basis without S10 test inference; the selected condition then ran on five fresh paired seeds.

Held-out official macro BAcc rises from $78.11 \pm 0.67\%$ to $82.85 \pm 1.43\%$ (Fig. 2), a paired gain of $+4.74 \pm 1.29$ points with five wins in five seeds and a video-cluster interval of $[+2.50, +6.95]$. Per-task BAcc is $82.38 \pm 1.06\%$ for video-driven audio hallucination, $77.42 \pm 3.89\%$ for audio-driven video hallucination, and $84.87 \pm 1.52\%$ for AV matching. Correct conditioning remains causal: within-demand shuffling removes $8.92 \pm 0.70$ points on evidence and $31.45 \pm 1.25$ points on relation. The preregistered stable-improvement gate passes; the deliberately stronger five-point materiality gate is missed by 0.26 points.

The contrast basis adds 1.049M trainable parameters (2.910M total; representative checkpoint 8.4 MiB versus 4.4 MiB), but controlled validation screens distinguish signal structure from width: the five-way contrast beats selected, mean, and three-stream alternatives, then beats the strongest audiovisual aggregate on all three confirmation validations ($82.18\%$ versus $80.92\%$ mean). S10 is therefore a targeted signal repair, not evidence that controller scaling alone solves multimodal conditioning.

# 8 From programmable transforms to real language modeling

## 8.1 S13: a constructive computation-plane separation

We first ask whether generated weights can be a model-level primitive rather than an adapter on a frozen host. S13 uses four causal layers with three gated recurrent mixers and one full-attention refresh. An ordered two-operation stream controls early and late depth; training withholds fixed operation pairs and evaluation doubles the memory length. All arms share the split, dimensions, optimizer, 20,000-step budget, and sinusoidal positions.

The hNET reaches the 95% learning threshold in every seed; FiLM never does. Shuffling its controller reduces seen accuracy from 100% to 5.61%. It also uses 1,568 fewer parameters than FiLM and 20.1% fewer than independent layer weights. Thus neither topology, sharing, nor scalar conditioning explains the effect: input-dependent off-diagonal transforms supply an ordered computation

Table 6: Real-language architecture transfer. Lower validation NLL is better. The 2M column is the mean of three paired seeds; the 20M column is one locked seed with batch 256. Shuffle is the 20M NLL increase after permuting controller outputs.

| Recurrent actuator | Parameters | 2M tokens | 20M tokens | Shuffle |
|---|---|---|---|---|
| Gated DeltaNet | 1.340M | 6.1857 | **5.6583** | — |
| Static rank-8 | 1.254M | 6.1744 | — | — |
| Activation FiLM | 1.300M | 6.1461 | 5.7395 | $\approx 0$ |
| Generated rank-8 hNET | 1.291M | **6.0953** | 5.6988 | +0.030 |
| Generated 4-head hNET | 1.347M | — | 5.7085 | +0.341 |
| Adaptive-clock hNET | 1.348M | — | 5.7065 | +0.0005 |

that fixed weights and channelwise gains do not. S13 establishes a mechanism, not yet a language-model advantage.

### 8.2 S23–S25: hNET versus Gated DeltaNet on WikiText-103

We transfer the exact mechanism to next-token prediction on WikiText-103 [Merity et al., 2017]. Every model has eight width-96 blocks arranged as two repetitions of one causal full-attention layer followed by three recurrent layers (25% attention), a 4,096-token BPE vocabulary, sequence length 128, and matched embedding/MLP components. The recurrent comparator is Gated DeltaNet (GDN) [Yang et al., 2024], which writes key–value associations into an error-corrected fast-weight state. The hNET instead uses the current token and recurrent state to generate a rank-8 transition in learned row/column subspaces. Static low-rank and activation-space FiLM arms preserve the same host interface without dynamic weight generation.

At two million tokens, the rank-8 hNET beats the matched GDN by 0.090 NLL, FiLM by 0.051, and the static low-rank mixer by 0.079, winning every paired seed against each arm while using 3.6% fewer parameters than GDN. Its normal-to-shuffled cost averages 0.051 NLL and its zero-controller cost 0.418; FiLM's corresponding controls are numerically zero. Dynamic weight generation is therefore used by the model and adds value beyond activation gains or a static adapter. The GDN contrast is topology-specific: a separately tested four-block, attention-last GDN obtains 5.9977 NLL at the same two-million-token budget. We consequently claim a matched-design advantage, not that hNET generally dominates GDN.

At 20 million tokens, the learning curves identify a second regime. The hNET leads early, GDN crosses it at roughly 11 million tokens, and final NLL is 5.6583 for GDN versus 5.6988 for hNET; the hNET still beats FiLM by 0.0407. This run controls tokens, topology, batch size, and parameter scale, although batch 256 yields only 610 optimizer updates and therefore does not settle update-matched scaling. In the reference implementation, peak training memory is 3.15 GiB for hNET versus 7.49 GiB for GDN; hNET fits batch 512 at 6.26 GiB and reaches 105.9k tokens/s, whereas GDN exceeds 15 GiB. These are unfused implementation measurements, not intrinsic complexity claims.

### 8.3 S26: stronger control is not more content memory

The architectural difference suggests a causal test. Each GDN layer maintains four $24 \times 24$ associative matrices—2,304 dynamic state elements—and updates key–value content at every token. The rank-8 hNET exposes only 64 generated core coefficients and a 96-dimensional vector state. We therefore enlarge the hNET within the GDN parameter and memory envelope: 192-dimensional recurrent state, four rank-12 generated cores (576 coefficients), four program slots, and a persistent controller. A fixed-clock arm forces the program to advance through depth; an adaptive-clock arm learns a monotone advance. Both remain below 4.0 GiB peak memory.

The intervention makes the controller dramatically more causal without improving language modeling. For the fixed-clock arm, shuffling costs 0.341 NLL and zeroing costs 4.05, compared with 0.030 and 0.136 for rank 8; yet its validation NLL is 5.7085 rather than 5.6988. The adaptive clock obtains 5.7065 but learns advances below $2 \times 10^{-4}$—effectively never advancing—and its shuffle cost is only 0.0005. The objective accepts stronger generated control when forced, but does not reward sequential routing on its own.

This is the paper's central causal boundary. Expanding the controller and reachable transform makes the hNET a stronger *control plane*; it does not create GDN's content-addressable *data plane*. Conversely, GDN's associative memory does not provide S13's input-programmed operator composition. The promising architecture is therefore not *hNET instead of GDN*, but a generated-weight controller over an associative recurrent substrate, punctuated by full attention. The present experiments establish the decomposition; testing that combined HyperGDN is the next prospective step.

# 9 Synthesis: control is a property of the interface

Six conclusions survive the heterogeneous record and prospective interventions.

**Condition-dependent gains have real causal effects.** Shuffling collapses length, audio, vision, policy, and AV relation controls; mid-generation changes redirect continuations. The mechanism is not merely a static adapter or prompt prior.

**Its natural role is persistent low-dimensional control.** Length, coarse selection, one-bit policy, and time-varying schedules are strong at low rank. These tasks ask the condition to choose among host-resident behaviors. Weight modulation applies globally and does not compete for a token position, which explains its zero-shot long-context robustness.

**It is not a general sensory information path.** Fine retrieval, boundary fidelity, frozen AV binding, and lag reasoning expose losses that controller width and gain rank do not repair. When instance evidence must be transported, a shared soft prefix is a stronger and more stable all-rounder in the matched S7 comparison. This does not make prefixes universally superior: they pay KV/context costs and require training for length shifts.

**Nominal capacity is the wrong abstraction.** Controller parameters measure how complicated the map $c \mapsto u$ may be, not which changes $\mathcal{A}(u)$ can make to the host. Dense and rotational cores expand the reachable set, yet their benefits depend on task and optimization geometry. S8–S9 show that a nondegenerate dictionary and zero-core initialization improve stability and accuracy even when expert-semantic placement is erased; freezing is unnecessary. The correct experimental unit is the complete signal–objective–actuator–optimization path.

**Richer weight-space actuation can become a computation plane.** S13 supplies the constructive boundary missing from the frozen-host record. Generated off-diagonal cores learn ordered programs, transfer to unseen compositions, and retain most performance at doubled length where parameter-matched scalar gains do not. Reachable geometry is therefore predictive under a host that repeatedly composes the generated transform. **A control plane is not a content memory.** On real text, dynamic weight generation beats FiLM and static low-rank controls, and at two million tokens it also beats an exact-topology GDN. Longer training reverses only the GDN comparison. Enlarging the generated core then increases condition dependence by an order of magnitude without improving NLL. Together these interventions isolate GDN's associative state as a complementary resource, not a stronger version of the same mechanism.

This mechanism view also clarifies the original consolidation goal. On three AVHBench tasks, one shared gain-conditioned adapter matches a bank of three task-specific hNET experts within the locked five-point band on every task while eliminating 66.7% of adapter parameters. A shared prefix likewise matches three prefix experts ($75.17 \pm 0.26\%$ versus $74.64 \pm 2.19\%$ macro BAcc). S9–S10 then turn consolidation into an hNET-specific quality gain: repairing the generated-core optimization path reaches $78.11\%$, and repairing the cross-modal signal reaches $82.85\%$ with five paired wins and a confidence interval excluding zero. Consolidation is therefore not only a storage effect: one condition-generated weight family can retain task breadth and improve when the controller receives the relation the task actually requires.

# 10 Implications for neural artifacts

Weight-space learning treats parameters, updates, gradients, and trajectories as structured data [Han et al., 2026]. Our results add a complementary question: *what can a generated artifact cause the host to do?* Two artifacts with the same parameter count can expose radically different functional

channels. Conversely, controllers with very different sizes can collapse onto the same reachable update span.

We recommend that evaluations of generated adapters report four objects:

1. the information available before generation (signal probes and structure-preserving controls);
2. the geometry and dimension of the reachable artifact family;
3. causal dependence on the generated artifact (shuffle, mean, reversal, and wrong-modality controls); and
4. task behavior after interaction with the host, including cross-task interference and context shift.

An embedding of weights is not validated by controller reconstruction alone, and a generated update is not validated by a raw task score alone. The artifact should be evaluated as an intervention.

One systems result reinforces this point without serving as behavioral evidence. A resizable $r \times r$ HyperLoRA core fits parameters, gradients, and optimizer state in a 20 MiB persisting-L2 window. Cache pinning restores $3.5\times$ synthetic bandwidth under contention but gives only $1.08$–$1.11\times$ end-to-end because backbone traffic dominates; heterogeneous batching instead gives $66.7\times$ per-sample throughput at batch 64. The practical systems value of generated artifacts is therefore conditional batching and compact control state, not adapter bytes alone.

# 11 Related work

**Conditional parameter generation and modulation.** HyperNetworks generate parameters for another network [Ha et al., 2017]; Dynamic Filter Networks generate input-specific filters [De Brabandere et al., 2016]; FiLM and $\text{IA}^3$ modulate activation or attention channels with learned scales [Perez et al., 2018, Liu et al., 2022]. LoRA supplies fixed low-rank update factors [Hu et al., 2022]. HyperFormer and HyperPELT generate parameter-efficient adaptations shared across tasks [Mahabadi et al., 2021, Zhang et al., 2023], while HyperLoRA directly generates constrained low-rank adapters [Lv et al., 2024]. Our contribution is not another controller architecture; it separates controller expressivity from the reachable host intervention and tests that distinction across controlled repairs.

**Token-space conditioning.** Prefix tuning, prompt tuning, and related methods insert learned vectors into the transformer computation [Li and Liang, 2021, Lester et al., 2021]. They are natural comparators because they transport token values rather than rescale fixed weight directions. Our matched study identifies a stability advantage for evidence transport, alongside the known context-position cost exposed by our long-context control.

**Merging and weight-space learning.** LoRAHub composes adapters for cross-task generalization [Huang et al., 2023]; TIES resolves interference in model merging [Yadav et al., 2023]; model soups and task arithmetic combine trained solutions in weight space [Wortsman et al., 2022, Ilharco et al., 2023]. Neural functional and weight-space learning methods explicitly model parameter symmetries [Navon et al., 2023, Ainsworth et al., 2023]. These works primarily ask how to represent, align, or combine existing artifacts. We ask how much condition information an input-dependent artifact can deliver through a particular interface before any merge.

**Recurrent fast weights.** Linear transformers and delta-rule models maintain recurrent key–value state as an alternative to quadratic attention; Gated DeltaNet combines gated decay with delta-rule updates [Yang et al., 2024]. This state stores token-dependent associations, whereas our hNET generates the transition applied to a smaller recurrent state. Our matched and strengthened-control experiments empirically separate these content-memory and control-plane roles. **Multimodal evaluation.** AVHBench evaluates cross-modal hallucination and audio–visual matching [Kim et al., 2025]; AVCA-GZSL provides official audiovisual generalized zero-shot splits and SeLaVi features [Mercea et al., 2022]. We use them as externally defined testbeds while adding controlled assays only when clearly separated from benchmark scores.

# 12 Limitations and broader impact

The evidence spans multiple backbones, datasets, and evaluation forms; this breadth motivates the mechanism but prevents pooled effect sizes. The cross-domain map contains several single-run

studies, while S7 and the two-million-token language comparison use only three seeds. The latter's hNET–GDN paired confidence interval includes zero, so its 3/3 wins establish a reproducible matched-setting result rather than population-level superiority. The 20-million-token scaling and strengthened-control studies use one locked seed and only 610 optimizer updates; update-matched scaling remains open. The strengthened hNET also shares recurrent mixers within each three-layer group to remain parameter-matched, whereas rank 8 uses depth-specific mixers, so the intervention tests the complete higher-capacity design rather than core width alone. Memory and throughput numbers are implementation-specific and unfused. AVHBench uses frozen PE-AV features, S13 is synthetic, and neither establishes end-to-end raw multimodal pretraining or competitive large-scale language modeling. Finally, *low bandwidth* is a behavioral description of the tested interface, not a mutual-information estimate.

The experiments study foundational adaptation mechanisms and release no new personal data or high-risk model. More reliable modality conditioning may improve accessibility and efficient multimodal systems, but overstating a lossy condition channel could produce ungrounded outputs in applications that require sensory fidelity. The practical mitigation is the paper's evaluation policy: direct signal probes, wrong-condition controls, and explicit reporting of which modality evidence actually affects behavior. Dataset and model licenses are documented in the artifact package; the affect audit uses previously generated model outputs and is not used to make claims about people.

# 13 Conclusion

Input-dependent weights are not characterized by controller size alone. Their function is set jointly by the information in the condition, the distinctions rewarded by the objective, the reachable transformations, the optimization path, and the state maintained by the host. This view turns an apparently inconsistent record into a causal map. Gains robustly set length and persistent policy. Structured signals and direct losses recover temporal and cross-modal distinctions. On AVHBench, a non-degenerate dense-core initialization improves hNET by 8.66 points at identical size, and explicit audio–visual contrasts lift official macro balanced accuracy from $78.11\%$ to $82.85\%$ across five paired seeds.

At model level, generated transforms do something more specific than activation modulation: they implement input-programmed operators. They reach $83.84\%$ on unseen ordered programs versus $12.68\%$ for matched FiLM, then beat matched GDN, FiLM, and static low-rank recurrence across three WikiText-103 seeds at two million tokens. But the longer run and the strengthened-control intervention locate the boundary. Multiplying the generated core from 64 to 576 coefficients makes controller shuffling more than ten times as damaging while leaving language NLL unchanged; GDN's associative state still leads by 0.041 NLL at 20 million tokens. Control strength and content memory are distinct resources.

That distinction is the constructive result. An hNET can be the shared control plane that interprets modality, task, and program signals into persistent computation; an associative recurrent state can be the data plane that writes and retrieves token content; periodic full attention can refresh both. The evidence does not support replacing GDN with hNET. It supports a more capable decomposition: *generated control over learned memory*. This is the architecture the mechanism predicts, and the next experiment the paper makes unavoidable.

# References


Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha S. Srinivasa. Git Re-Basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023. URL `https://arxiv.org/abs/2209.04836`.

Bert De Brabandere, Xu Jia, Tinne Tuytelaars, and Luc Van Gool. Dynamic filter networks. In *Advances in Neural Information Processing Systems*, 2016. URL `https://papers.nips.cc/paper/2016/hash/8bf1211fd4b7b94528899de0a43b9fb3-Abstract.html`.

David Ha, Andrew Dai, and Quoc V. Le. Hypernetworks. In *International Conference on Learning Representations*, 2017. URL `https://arxiv.org/abs/1609.09106`.

Xiaolong Han, Zehong Wang, Bo Zhao, Binchi Zhang, Jundong Li, Damian Borth, Rose Yu, Haggai Maron, Yanfang Ye, Lu Yin, and Ferrante Neri. A survey of weight space learning: Understanding, representation, and generation. *arXiv preprint arXiv:2603.10090*, 2026. URL https://arxiv.org/abs/2603.10090.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. LoRAHub: Efficient cross-task generalization via dynamic LoRA composition. *arXiv preprint arXiv:2307.13269*, 2023. URL https://arxiv.org/abs/2307.13269.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2212.04089.

Sung-Bin Kim, Hyun-Bin Oh, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. AVHBench: A cross-modal hallucination benchmark for audio-visual large language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=jTEKTdI3K9.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Conference on Empirical Methods in Natural Language Processing*, 2021. URL https://aclanthology.org/2021.emnlp-main.243/.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Annual Meeting of the Association for Computational Linguistics*, 2021. URL https://aclanthology.org/2021.acl-long.353/.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2205.05638.

Kai Lv, Yuqing Yang, Tengxiao Liu, Qipeng Gao, Qibin Guo, and Xipeng Qiu. HyperLoRA: Efficient cross-task generalization via constrained low-rank adapters generation. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024. URL https://aclanthology.org/2024.findings-emnlp.956/.

Rabeeh Karimi Mahabadi, Sebastian Ruder, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. In *Annual Meeting of the Association for Computational Linguistics*, 2021. URL https://aclanthology.org/2021.acl-long.47/.

Otniel-Bogdan Mercea, Lukas Riesch, A. Sophia Koepke, and Zeynep Akata. Audio-visual generalised zero-shot learning with cross-modal attention and language. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2203.03598.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1609.07843.

Aviv Navon, Aviv Shamsian, Idan Achituve, Ethan Fetaya, Gal Chechik, and Haggai Maron. Equivariant architectures for learning in deep weight spaces. In *International Conference on Machine Learning*, 2023. URL https://proceedings.mlr.press/v202/navon23a.html.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018. URL https://arxiv.org/abs/1709.07871.

Apoorv Vyas, Heng-Jui Chang, Cheng-Fu Yang, Po-Yao Huang, Luya Gao, Julius Richter, Sanyuan Chen, Matthew Le, Piotr Dollär, Christoph Feichtenhofer, Ann Lee, and Wei-Ning Hsu. Pushing the frontier of audiovisual perception with large-scale multimodal correspondence learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2512.19687.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.01708.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. *arXiv preprint arXiv:2412.06464*, 2024. URL https://arxiv.org/abs/2412.06464.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. HyperPELT: Unified parameter-efficient language model tuning for both language and vision-and-language tasks. In *Findings of the Association for Computational Linguistics: ACL*, 2023. URL https://aclanthology.org/2023.findings-acl.725/.

# A Proof of Proposition 1

Let $M_i = b_i a_i^\top$. Equation 1 gives $\Delta W(g) = \sum_{i=1}^{r} g_i M_i$, hence every reachable update lies in $\mathrm{span}\{M_1, \ldots, M_r\}$, whose dimension is at most $r$ (and may be smaller if the atoms are linearly dependent). This conclusion is independent of the function class that maps $c$ to $g$.

For a dense core $C$, expand

$$BCA = \sum_{i=1}^{r} \sum_{j=1}^{r} C_{ij} b_i a_j^\top . \tag{5}$$

The reachable family lies in the span of at most $r^2$ fixed outer-product atoms. For a Cayley core $Q(S) = (I - S)(I + S)^{-1}$ with skew-symmetric $S$, the parameter space has $r(r-1)/2$ independent coordinates wherever $I + S$ is invertible. Its image is a nonlinear subset of the dense-core family with local dimension no greater than the parameter-space dimension. These are upper bounds; degeneracy in $A$, $B$, or the controller image can reduce them.

# B Detailed S7 implementation

**Conditions and split.** All QA rows associated with a video remain in one SHA-256 split. Policy labels are assigned after sorting video identifiers by a stable hash, giving exact or near-exact balance within every split. Positive and negative policy conditions are a fixed random unit vector and its negation. Identity rows use the corresponding audio-only or visual-only PE-AV vector; relation rows use the audiovisual vector. Cyclic shuffles preserve demand and modality while breaking video alignment. Wrong-modality controls exchange audio and visual, and replace audiovisual with audio; they are undefined for policy.

**Actuators.** Weight-space arms adapt $q_{\mathrm{proj}}$, $k_{\mathrm{proj}}$, $v_{\mathrm{proj}}$, and $o_{\mathrm{proj}}$ at rank 2 with $\alpha = 48$ and dropout 0.05. Diagonal gains initialize to the identity. Dense cores are generated as identity plus an unconstrained $2 \times 2$ residual. Both use a shared normalized 1,024→256 stem and a two-layer width-128 SiLU controller. The ARAM arm maps the same stem through a width-1,000 controller to bounded coefficients over eight deterministic relative-position bases in every layer. Its first smoke test stopped before any optimizer step because fused SDPA did not expose gradients through the additive mask with frozen Q/K/V. The preregistration was amended before any ARAM validation or test result to use mathematically equivalent eager attention with an explicit causal mask. The shared prefix maps the 256-D stem through a 552-D GELU bottleneck to one 2,560-D token. The comparator uses three disjoint 1,024→169→2,560 projectors.

**Training and metrics.** Each epoch contains every official training row and one policy row per unique training video. Demands are deterministically interleaved. Binary cross-entropy is implemented as two-candidate next-token cross-entropy. The secondary counterfactual diagonal arm adds $0.25 \max(0, 0.5 - m_{\mathrm{matched}} + m_{\mathrm{shuffled}})$. BAcc uses a zero threshold on the Yes-minus-No score. Linear probes standardize features on train, fit a regularized logistic head for 200 AdamW steps, and evaluate on test. Cluster intervals resample test videos so all rows from a video move together.

Table 7: Locked S7 primary interactions in BAcc percentage points.

| Comparator | Mean | Paired seeds | Video-cluster 95% interval |
|---|---|---|---|
| Dense core | +1.55 | +4.33, +0.27, +0.04 | $[-1.06, +4.65]$ |
| Attention bias | $-5.21$ | +1.14, $-17.52$, +0.75 | $[-16.83, +2.27]$ |
| Shared prefix | +1.76 | +3.85, +1.49, $-0.08$ | $[-0.66, +4.40]$ |
| Specialist prefixes | +0.90 | +2.16, +1.22, $-0.67$ | $[-1.31, +3.16]$ |

# C Rescue logic and additional controls

**AVHBench consolidation.** The shared rank-1 hNET uses 1,068,176 trainable parameters, compared with 737,280 for a condition-free static LoRA and 1,076,768 for the parameter-matched shared prefix. The video-disjoint split contains 3,693/770/839 train/validation/test questions and all methods receive four epochs. Across three seeds, the hNET obtains $72.46 \pm 2.55$ macro BAcc and condition

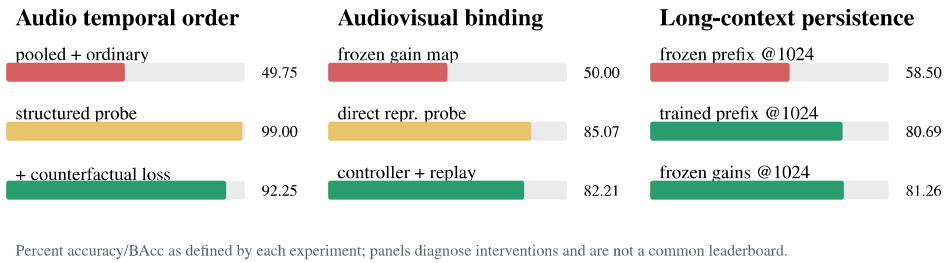


Figure 3: Large improvements follow targeted repairs to signal, objective, or training distribution. Increasing controller size alone does not produce these rescues.

shuffling reduces overall BAcc by $12.96 \pm 0.59$ points. The shared prefix obtains $75.17 \pm 0.26$. Three prefix experts total three times the adapter storage but do not significantly exceed the shared prefix.

**Affect latent-signal test.** This experiment deliberately asks whether the conditioning path can discover and exploit an implicit tone/register variable. Fixed warm and cold endpoints are behaviorally distinct, showing detectable endpoint structure, while a dynamic interpolation through a nominal neutral target is not resolved by the blind measure. The corrected audit uses 110 individual 1–7 warmth ratings and a prompt-based positive control. Earlier lexical divergence scores are excluded because they fall below their own noise floor and anticorrelate with the blind measure. The $+0.08$ dynamic effect, versus a 3.45-point prompt span, therefore measures the sensitivity boundary of the tested channel rather than a general inability to control affect.

**Systems boundary.** The cache-resident experiment runs on an RTX 5060 Ti with 16 GB memory, a measured 32 MiB L2 cache, and a 20 MiB persisting region. The 0.98M-parameter hypernetwork occupies 15.0 MiB including parameters, gradients, and Adam states. Its microbenchmark and end-to-end results separate a true hardware effect from a negligible workload contribution. This experiment motivates reporting full-backbone traffic and per-sample batching, rather than adapter bytes alone.

# D Per-test evidence ledger

# E Reproducibility and artifact map

The supplement contains the frozen S7 and S8 protocols, runners, launchers, aggregation code, all per-seed JSON outputs and raw predictions, checkpoints, cross-domain corrected reports, and a machine-readable evidence registry. The central paths are:

- `PROTOCOL_S7.md`; `audio_probe/run_s7_actuator_matrix.py`;
- `artifacts/evaluation/s7_actuator_matrix_summary.json`;
- `PROTOCOL_S8_EXPERT_ALIGNED.md` and `PROTOCOL_S9_FREEZE_ISOLATION.md`; S8–S9 summaries under `artifacts/evaluation/`;
- `PROTOCOL_S10_CONFIRMATION.md`, `RESULT_S10.md`, and `artifacts/evaluation/s10_contrast5_summary.json`;
- `artifacts/evaluation/mechanism_evidence.json`;
- `PROTOCOL_S13_HYBRID_ARCHITECTURE.md` and `artifacts/evaluation/s13_hybrid_summary.json`;
- `PROTOCOL_S23_REAL_INTERPRETER_TRANSFER.md` through `PROTOCOL_S26_HYPERMEMORY_V2.md`, with matching result reports and `artifacts/evaluation/s23_*`–`artifacts/evaluation/s26_*` JSON outputs;
- `TEST_LEDGER.md` and `EVIDENCE_LEDGER.md`; and
- corrected domain reports under `artifacts/external_audit/`.