TRACING PROMPT-CONDITIONED CONTROL FROM ALLOCATION TO REALIZED ACTUATION IN FROZEN VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

We analyze inference-time control authority through the measurable sequence $(\mathbf{a} \rightarrow \mathbf{m} \rightarrow \mathbf{w} \rightarrow \mathbf{u} \rightarrow \mathbf{r} \rightarrow \boldsymbol{\ell} \rightarrow \mathbf{p} \rightarrow \mathbf{y})$. In one paired checkpoint, Dynamic-to-Uniform composition yields equal measured accuracies. Across five training seeds, the replacement changes the Prompt-derived state while retaining a zero-median allocation response and unchanged dominant allocation indices. Direct allocation and Prompt-input replacements produce downstream continuous responses at different controller coordinates, revealing a recurrent dissociation between conditioning variation, continuous allocation, realized actuation, and decision-boundary crossing.

1 INTRODUCTION

Adapting frozen Vision-Language Models (VLMs) requires controlling representation changes while retaining pretrained knowledge Houlsby et al. (2019); Hu et al. (2022; 2023); Zhang et al. (2024a); Xu et al. (2023); Zhang et al. (2023). Multimodal models also exhibit cross-modal representation gaps Liang et al. (2022), while adaptation can introduce forgetting Kirkpatrick et al. (2017); He et al. (2022). We use a compact frozen VLM, SmolVLM Marafioti et al. (2025), to support repeated paired interventions and intermediate-state capture. However, learned adaptation and inference-time control remain distinct evaluation targets, leaving a separate question unresolved: whether a learned conditioning variable retains measurable inference-time authority over downstream controller states.

While adaptation evaluates whether training modifies a target criterion, control assesses whether internal variables maintain measurable authority over model states. To examine this distinction, we analyze a controller that decouples prompt composition, conditioning, continuous allocation, and representation displacement. Although learned prompt conditioning alters states across examples, dynamic-to-centroid replacement yields a zero-median allocation response without altering dominant routes across seeds. Direct allocation and zero-input replacements produce continuous downstream responses at different control coordinates. Together, the measurements separate conditioning variation, continuous allocation, realized actuation, and decision-boundary crossing in the audited controller.

To determine how input-conditioned control signals transmit from prompt composition to the frozen model's readout, we employ VEFT Kawamae (2025c) as an audited instance and learned representation controller Elhage et al. (2022); Kawamae (2025a;c); Anonymous (2026b). Specifically, we address three key questions:

1. Does changing the learned prompt-composition policy alter the prompt-derived conditioning state?
2. How much of that change reaches continuous allocation and realized actuation?
3. Is the response reached through the prompt policy distinguishable from responses obtained by direct allocation and prompt-input interventions?

Our contributions are:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

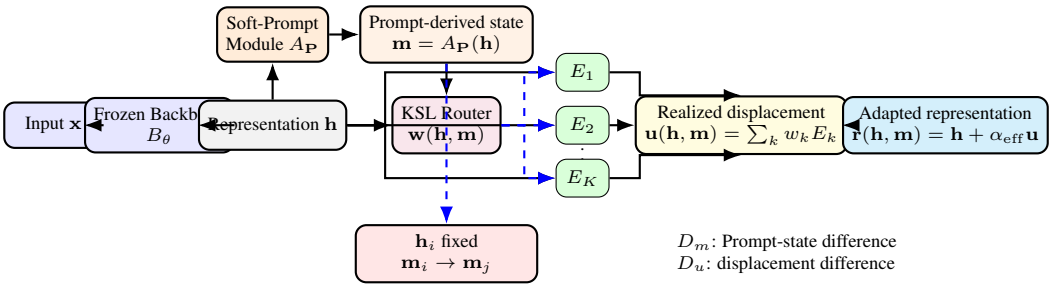

Figure 1: Audited controller path. The Prompt-derived state conditions continuous allocation over state-dependent transformations. The reported interventions manipulate Prompt composition, the Prompt input to the allocator, or the allocation coordinate and record responses at subsequent stages.

- We identify a recurrent dissociation between prompt-derived conditioning variation, continuous allocation, realized actuation, and decision-boundary crossing in a trained controller.
- We organize these measurements as a stage-wise audit that pairs each controller coordinate with a specific intervention and observable.
- We find that the zero-median, unchanged-index pattern recurs across five training seeds, while continuous response magnitudes vary across checkpoints.
- Direct allocation and prompt-input replacements distinguish the prompt-policy response from responses initiated at other controller coordinates.
- These findings motivate a stage-wise evaluation decomposition: conditioning variation, continuous allocation, realized actuation, distributional response, and boundary crossing can be reported separately when these coordinates are available.

Task performance identifies an endpoint but does not localize where a learned control signal is attenuated, redirected, or preserved within the controller.

2 Previous Work

Continuous prompt methods adapt frozen models through learned representations Lester et al. (2021); Liu et al. (2021), with input-dependent extensions Qin & Eisner (2021); Asai et al. (2022); Choi et al. (2023). Complementary steering and representation-level interventions modify internal states rather than relying solely on weights Wu et al. (2024b); Turner et al. (2023); Li et al. (2023a); Lee et al. (2025); Sun et al. (2025), while causal analyses test functional roles Geiger et al. (2021); Meng et al. (2022). Although frameworks like VEFT and KLAFT prioritize external behavior and efficiency Kawamae (2025c); Anonymous (2026b), their internal dynamics remain under-explored.

Activation patching and interchange interventions replace internal states to measure downstream responses (Geiger et al., 2021; Zhang & Nanda, 2024), with interpretations depending on the replacement policy and response metric. Rather than introducing a new replacement operator, we apply stage-specific interventions to a prompt-conditioned representation controller. Causal abstraction studies faithful relations between low-level neural mechanisms and interpretable high-level models (Geiger et al., 2025). VEFT is an audited instantiation of the controller path rather than the scientific object of the paper. PAAF Anonymous (2026a) uses the same substrate to study optimization-time Schedule and Action selection, including KnG-based Actions Kawamae (2025b). This work holds the trained controller fixed and audits evaluation-time transmission across its conditioning, allocation, actuation, readout, and decision coordinates.

3 Problem Definition

We analyze the shallow feedforward controller applied to each token between the frozen backbone and the language-model head. Its stage variables occupy

$$\mathbf{a} \in \Delta^{M-1}, \quad \mathbf{m} \in \mathbb{R}^{d_m}, \quad \mathbf{w} \in \Delta^{K-1}, \quad \mathbf{u}, \mathbf{r} \in \mathbb{R}^{d_h}, \quad \boldsymbol{\ell} \in \mathbb{R}^{V}, \quad \mathbf{p} \in \Delta^{V-1}, \quad \mathbf{y} \in \mathcal{Y}.$$

These variables represent the measurable coordinates: composition coordinate ($\mathbf{a}$), conditioning state ($\mathbf{m}$), continuous allocation ($\mathbf{w}$), realized actuator displacement ($\mathbf{u}$), adapted representation ($\mathbf{r}$), logits ($\boldsymbol{\ell}$), output distribution ($\mathbf{p}$), and behavioral evaluation ($\mathbf{y}$). It follows $\mathbf{a} \to \mathbf{m} \to \mathbf{w} \to \mathbf{u} \to \mathbf{r} \to \boldsymbol{\ell} \to \mathbf{p} \to \mathbf{y}$.

To analyze internal dynamics without assuming full identifiability, we define finite mechanism interventions $\mathcal{I} = (t, \widetilde{f}_t)$ that replace mechanism f_t at stage t while holding non-intervened parameters fixed. Stage-specific responses are quantified operationally via controlled within-model contrasts. Without additional injectivity or margin assumptions, upstream variations do not strictly guarantee downstream functional changes (e.g., $\mathcal{A}_m > 0 \not\Rightarrow \mathcal{A}_w > 0$). Thus, fixing intermediate variables defines intervention-specific controlled responses rather than unique global causal pathways. Our primary evaluation investigates an *Evaluation-Time Composition-Policy Sensitivity Test* to determine whether removing input-dependent Prompt weighting produces detectable contrasts under test conditions. This evaluation is interpreted strictly as an external sensitivity test rather than a general causal-identification or equivalence test.

4 METHOD

4.1 INTERVENTION TAXONOMY AND ESTIMANDS

The implementation instantiates the adaptation pathway through specific architectural components. Here, $\mathbf{a}$ is the Prompt-composition coordinate and $\mathbf{m}$ is the resulting conditioning state, treated as the implemented composition rule.

Specifically, a learned Prompt bank $\mathbf{P} = [\mathbf{p}_1^\top; \ldots; \mathbf{p}_M^\top] \in \mathbb{R}^{M \times d_m}$ is accessed through representation-conditioned weighting motivated by the attention mechanism (Vaswani et al., 2017) and mixture of prompts (Asai et al., 2022; Anonymous, 2026b):

$$\mathbf{q}(\mathbf{h}) = W_Q \mathbf{h}, \qquad \mathbf{K}_P = \mathbf{P} W_K^\top, \tag{1}$$

$$\mathbf{a}(\mathbf{h}) = \mathrm{softmax}\left(\frac{\mathbf{K}_P \mathbf{q}(\mathbf{h})}{\sqrt{d_r}}\right) \in \Delta^{M-1}, \qquad \mathbf{m}(\mathbf{h}) = \mathbf{P}^\top \mathbf{a}(\mathbf{h}) = \sum_{j=1}^{M} a_j(\mathbf{h}) \mathbf{p}_j, \tag{2}$$

where $W_Q \in \mathbb{R}^{d_r \times d_h}$ and $W_K \in \mathbb{R}^{d_r \times d_m}$ denote projection matrices, d_r is the projection dimension, and $\mathbf{h} \in \mathbb{R}^{d_h}$ is the backbone hidden state. The resulting conditioning state satisfies $\mathbf{m}(\mathbf{h}) \in \mathrm{conv}\{\mathbf{p}_1, \ldots, \mathbf{p}_M\}$. We treat activation replacement, edge blocking, and state substitution as mechanism interventions, following intervention-based analyses of neural representations and causal abstraction (Geiger et al., 2021; 2025). Because intervention conclusions depend on the replacement policy and response metric, we pair each intervention with an explicit stage-specific observable (Zhang & Nanda, 2024). Joint supervision learns the Prompt bank, composition map, allocator, and transformation bank. We treat knowledge relationally through the conditioning of state-dependent transformations applied to the current frozen representation.

To evaluate mechanism behavior, we establish an intervention sequence governed by core research questions. Dynamic-to-Uniform composition compares $\mathbf{a}_{\mathrm{dyn}}(\mathbf{h})$ against a uniform baseline $\mathbf{a}_{\mathrm{uni}} = (1/M, \ldots, 1/M)^\top$. Direct allocation replacement substitutes the effective allocation coordinate, fixed-allocation replay checks deterministic implementation consistency, and Prompt-input zero replacement provides supplementary sensitivity analysis. Each intervention is paired with the first downstream coordinate needed to answer its corresponding question.

For paired intervention conditions, we quantify stage-specific responses using the following operational measures:

$$A_a = \|\mathbf{a}_1 - \mathbf{a}_0\|_1, \quad A_m = \|\mathbf{m}_1 - \mathbf{m}_0\|_2, \quad A_w = \|\mathbf{w}_1 - \mathbf{w}_0\|_1, \quad A_u = \|\mathbf{u}_1 - \mathbf{u}_0\|_2, \tag{3}$$
$$A_r = \|\mathbf{r}_1 - \mathbf{r}_0\|_2, \quad A_\ell = \|\boldsymbol{\ell}_1 - \boldsymbol{\ell}_0\|_2, \quad A_p = \|\mathbf{p}_1 - \mathbf{p}_0\|_1, \quad A_y = \mathbf{1}[\arg\max \mathbf{p}_1 \neq \arg\max \mathbf{p}_0].$$

Exact zeros correspond to entries in the stored audit statistics; because numerical prevalence depends on tolerance, we additionally report support above an L_1 threshold of 10^{-5} where relevant.

Let ψ denote the parameters of the transformation bank $\{E_k\}_{k=1}^K$, and define $R_\phi : \mathbb{R}^{d_h} \times \mathbb{R}^{d_m} \to \Delta^{K-1}$ as the allocator. In the audited VEFT controller, the KSL module defines

the allocation coordinate $\mathbf{w}(\mathbf{h}, \mathbf{m}) = R_\phi(\mathbf{h}, \mathbf{m})$, the realized actuator displacement $\mathbf{u}(\mathbf{h}, \mathbf{m}) = \sum_{k=1}^{K} w_k(\mathbf{h}, \mathbf{m}) E_k(\mathbf{h})$, and the adapted representation $\mathbf{r}(\mathbf{h}, \mathbf{m}) = \mathbf{h} + \alpha_{\text{eff}} \mathbf{u}(\mathbf{h}, \mathbf{m})$. While normalized routing is also used in conditional-computation systems (Zhou et al., 2022), KSL uses dense probabilistic weighting over subnetwork outputs, whereas sparse MoE selects a limited subset. This decomposition distinguishes the computational allocation coordinate $\mathbf{w}$ from the realized displacement $\mathbf{u}$. Routing displacement and realized actuation are distinct. For fixed $\mathbf{h}$ and simplex-valued $\mathbf{w}, \mathbf{w}'$,

$$\|\mathbf{u}(\mathbf{w}') - \mathbf{u}(\mathbf{w})\|_2 \leq \frac{D_E(\mathbf{h})}{2} \|\mathbf{w}' - \mathbf{w}\|_1, \quad D_E(\mathbf{h}) = \max_{k,l} \|E_k(\mathbf{h}) - E_l(\mathbf{h})\|_2. \tag{4}$$

The bound shows that allocation displacement and realized actuation are different quantities. The interpretation of A_w therefore depends on the geometry of the transformation outputs represented by $D_E(\mathbf{h})$. Accordingly, $\mathbf{u}$ is the realized actuator displacement used in our mechanistic formulation, whereas $\mathbf{w}$ is the corresponding allocation coordinate.

The present Dynamic-to-Uniform experiment evaluates downstream behavior rather than directly establishing a mechanistic effect from $\mathbf{m}$ to $\mathbf{u}$ unless the corresponding internal quantities are explicitly measured. We record same-model paired contrasts in $\mathbf{a}$, $\mathbf{m}$, $\mathbf{w}$, $\mathbf{u}$, and $\mathbf{r}$ under both Dynamic and Uniform prompt compositions. While this finite-intervention audit localizes attenuation along the realized pathway, complementary analyses are required to fully characterize the response to arbitrary prompt-derived states. We evaluate the framework via a VEFT/KSL instantiation, leveraging its hierarchical auditable variables and edge-level interventions. To assess sensitivity, we substitute the input-dependent composition policy with a uniform one ($a_k^{\text{uni}} = 1/M$), yielding the prompt centroid $\mathbf{m}_{\text{uni}} = \frac{1}{M} \sum_{k=1}^{M} \mathbf{p}_k$ as a deterministic removal of input-dependent weighting while preserving the learned prompt bank. This intervention preserves the prompt bank, KSL parameters, subnetwork transformations, and training/evaluation setup, isolating the composition policy:

$$\mathbf{m}_{\text{dyn}}(\mathbf{h}) \quad \longrightarrow \quad \mathbf{m}_{\text{uni}}, \tag{5}$$

followed by standard routing and actuation. Because actuation occurs directly beneath the LM head, the actuator-to-loss Jacobian omits downstream Transformer block products: $\partial \mathcal{L} / \partial \theta_{\text{act}} = (\partial \mathcal{L} / \partial \boldsymbol{\ell}) W_{\text{LM}} \alpha_{\text{eff}} (\partial \mathbf{u} / \partial \theta_{\text{act}})$. This placement removes downstream Transformer blocks from the observed actuation-to-logit path. We use it to capture representation and logit responses directly, rather than inferring improved optimization stability from placement alone. We therefore interpret the Dynamic-to-Uniform comparison as a finite intervention contrast:

$$\Delta_{\text{ext}}^{\mathcal{M}} (G_{\text{uni}}, G_{\text{dyn}}) = \mathcal{M} (G_{\text{uni}}) - \mathcal{M} (G_{\text{dyn}}), \tag{6}$$

where $\mathcal{M}$ denotes the selected external measurement.

4.2 Intervention-to-Estimand Map

Table 8 pairs each intervention with its measured response and supported inference. The same-model audit measures finite paired contrasts in $\mathbf{a}$, $\mathbf{m}$, $\mathbf{w}$, $\mathbf{u}$, and $\mathbf{r}$ under the Dynamic-to-Uniform replacement. Appendix characterizes the scope of unmeasured directions. The reported estimands are finite paired responses under specified coordinate replacements.

The transformation module is inserted immediately before the language model head, such that

$$\boldsymbol{\ell} = W_{\text{LM}} (\mathbf{h} + \alpha_{\text{eff}} \mathbf{u}) + \mathbf{b}. \tag{7}$$

This placement provides a short and explicit observation path from the realized transformation to logits while preserving the frozen foundation-model backbone. Although the explicit credit-assignment path is shorter than for modules inserted deep in the backbone, we do not infer improved optimization stability or immunity to vanishing or exploding gradients from placement alone.

Let $\widehat{\eta} = \left(\widehat{\mathbf{P}}, \widehat{Q}, \widehat{K}, \widehat{\phi}, \widehat{\psi}, \widehat{\alpha}_{\text{eff}} \right)$ denote the adaptation parameters optimized using token-level cross-entropy on the training corpus while the foundation-model backbone remains frozen:

$$\widehat{\eta} = \arg\min_{\eta} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}_{\text{train}}} \left[\mathcal{L}_{\text{CE}} \left(\mathbf{y}, p_\eta(\cdot \mid \mathbf{x}) \right) \right]. \tag{8}$$

The reported intervention audits are performed after training with $\widehat{\eta}$ fixed, characterizing the response structure learned by training rather than changes induced by additional optimization. Task,

role, instruction, and multimodal context can influence the learned conditioning and allocation maps through the supervised objective. Knowledge is treated relationally through the learned conditioning and allocation maps. The analysis assigns no exclusive semantic label to individual Prompt vectors or subnetworks.

Local differential interpretation. Holding $\mathbf{h}$ fixed, let

$$\mathbf{E}(\mathbf{h}) = [E_1(\mathbf{h}), \ldots, E_K(\mathbf{h})] \in \mathbb{R}^{d_h \times K}, \quad \text{and} \quad J_{X \leftarrow Y} = \frac{\partial \mathbf{X}}{\partial \mathbf{Y}}. \tag{9}$$

A local change in prompt composition yields the first-order approximations:

$$\delta \boldsymbol{\ell} \approx W_{\mathrm{LM}} \alpha_{\mathrm{eff}} \mathbf{E}(\mathbf{h}) J_{w \leftarrow m} J_{m \leftarrow a} \delta \mathbf{a}, \qquad \delta \mathbf{p} \approx J_{p \leftarrow \ell}(\boldsymbol{\ell}) \delta \boldsymbol{\ell}, \tag{10}$$

where $J_{w \leftarrow m} \in \mathbb{R}^{K \times d_m}$ and $J_{m \leftarrow a} = \mathbf{P}^\top \in \mathbb{R}^{d_m \times M}$ denote the local stage-transition Jacobians. This factorization identifies candidate stages shaping the local response. Because the reported results reflect finite paired contrasts rather than fitted individual Jacobian factors, a weak downstream response does not prove the absence of prompt conditioning. The evidential scope of this study is strictly restricted to establishing prompt-policy interventions as external response contrasts, rather than identifying underlying causal actuator mechanisms.

5 Experiments

5.1 Experimental Setting

For controlled cross-study comparison, we use the same frozen-VLM substrate and matched data construction as related optimization-time studies Anonymous (2026b). We adopt SmolVLM-Instruct Marafioti et al. (2025)[1] as the base Vision-Language Model, supporting the standard Transformers `GenerationMixin` interface, `lm_head`, and `output_hidden_states=True` for shallow routing module insertion. We focus on frozen-backbone adaptation, omitting input-level prompt tuning and full-parameter fine-tuning. The experiments follow the intervention map defined in Section 4. Dynamic-to-Uniform composition provides the primary Prompt-policy analysis across five training seeds, while direct allocation replacement serves as a downstream response control. Fixed-allocation replay checks implementation consistency, and prompt-input zero replacement provides a secondary conditioning-direction analysis on a single checkpoint. Representation and decision statistics characterize the resulting downstream responses.

For the dataset, we curate a reasoning-centric subset from HuggingFaceM4/M3IT Li et al. (2023b) using a streaming loader to handle corpus scale. Specifically, we extract 1,000 samples (900/100 train/eval split) from VCR and Visual-MRC containing answer choices and rationale sentences.

Evaluation spans three benchmark families: **(i)** Q-Bench and Q-Bench+ via Q-Bench2-HF[2] provide multiple-choice low-level visual perception endpoints Wu et al. (2024a); Zhang et al. (2024b), **(ii)** VisIT-Bench for open-ended instruction following Bitton et al. (2023), and **(iii)** drift metrics for temporal stability (Section 3).

5.2 Analysis Plan and Intervention Roles

We evaluate four intervention families following the stage variables defined in Section 4: Dynamic-to-Uniform Prompt composition, direct effective-allocation replacement, fixed-effective-allocation blocking, and Prompt-input zero replacement. For each paired contrast, we record stage-specific distances in $\mathbf{a}$, $\mathbf{m}$, $\mathbf{w}$, $\mathbf{u}$, $\mathbf{r}$, $\boldsymbol{\ell}$, and $\mathbf{p}$, alongside dominant-allocation-index and top-1 decision changes. Dynamic-to-Uniform composition assesses the tested Prompt policy, direct allocation replacement measures downstream responses from the allocation coordinate, fixed-allocation replay verifies implementation consistency, and Prompt-input zero replacement evaluates secondary conditioning on a single checkpoint. Representation and decision statistics characterize the resulting downstream response. These observables correspond directly to the coordinates defined in Section 3 and the response metrics defined in Section 4.

[1] `https://huggingface.co/HuggingFaceTB/SmolVLM-Instruct`

[2] `https://huggingface.co/datasets/q-future/Q-Bench2-HF`

Table 1: Evaluation-time intervention on SP4 Prompt composition. Both conditions use the same VEFT-only, tokenwise, concat, hybrid, softmax-router configuration and training seed. Dynamic attention is input-dependent, whereas Uniform fixes all four Prompt weights to $1/4$. Differences are descriptive results from one paired seed.

Metric	Dynamic	Uniform	Δ (Uni.-Dyn.)
Q-Bench Acc. (%)	42.00	42.00	0.00
Q-Bench2 Acc. (%)	54.00	54.00	0.00
BLEU (%)	2.12	2.12	0.00
ROUGE-L (%)	11.63	11.65	+0.02
METEOR (%)	16.99	17.02	+0.03
BERTScore F1 (%)	82.63	82.63	0.00
SBERT cosine	0.4138	0.4147	+0.0009
Early–late SBERT	0.7025	0.7046	+0.0021
Early–reference	0.3562	0.3562	0.0000
Late–reference	0.2954	0.2953	-0.0001

5.3 Prompt-Policy Response Across Five Seeds

This subsection estimates the finite response associated with the Prompt-policy substitution. We evaluate Dynamic-to-Uniform Prompt composition across five training seeds using a fixed regime: a four-vector Prompt bank, five transformation subnetworks, a softmax router, concat injection, hybrid decode routing, and $\lambda_{\mathrm{GSA}} = \lambda_{\mathrm{stab}} = 0$. Training uses dynamic composition in both conditions, while evaluation replaces input-dependent Prompt weights with the uniform vector $(1/4, 1/4, 1/4, 1/4)$, making the composition policy the sole manipulated variable.

The external sensitivity contrast is defined as:

$$\Delta^{\mathcal{M}}_{\mathrm{ext}} = \mathcal{M}(G_{\mathrm{uni}}) - \mathcal{M}(G_{\mathrm{dyn}}). \tag{11}$$

Table 1 reports one paired external evaluation checkpoint, showing equal accuracies and minor generation metric differences, while a five-seed internal audit evaluates controller responses.

Across all seeds, the intervention modifies the Prompt-derived state with a zero-median allocation response whose mean and support vary across checkpoints: the median A_w is zero, while the mean is positive with nonzero support on 14 to 31 of 100 examples. Seed-level means are $A_m = 0.0527 \pm 0.0273$ and $A_w = (1.50 \pm 1.04) \times 10^{-4}$ (sample standard deviation). The dominant allocation index remains unchanged in every seed.

An internal audit on 100 held-out examples reveals that Dynamic-to-Uniform composition alters Prompt weights ($D_a^{L1} = 0.1837$, 95% CI $[0.1783, 0.1886]$) and states ($D_m^{\mathrm{rel}} = 0.2250$, 95% CI $[0.2193, 0.2303]$; mean cosine similarity 0.9740). Routing and displacement contrasts are zero at stored precision for 69 examples and nonzero for 31. Within the stored-nonzero subset, allocation and actuation responses strongly co-vary ($\rho_{\mathrm{Spearman}} = 0.9998$), and residual-state contrast follows the applied gain $\alpha_{\mathrm{eff}} = 0.5933$. Downstream routing contrast averages $D_w^{L1} = 7.10 \times 10^{-5}$ (95% CI $[2.46 \times 10^{-5}, 1.30 \times 10^{-4}]$) with no routing-argmax changes, characterizing the evaluated replacement by its zero median, continuous magnitude, and thresholded support.

5.4 Direct-Allocation Control

This subsection evaluates the downstream response initiated directly at the allocation coordinate. The replacement produces nonzero actuation, representation, and logit responses over the tested replacement family. Direct effective-allocation replacement produces consistent nonzero changes in weighted actuation (mean responses ranging from 35.53 to 38.14 versus 0.00141 to 0.00811 under Prompt contrast), showing variation in the downstream map over the tested direct replacements, where direct replacement is a downstream response control, not a magnitude-matched estimate of transmission from the Prompt policy. Fixed-allocation replays yield zero error across all audited examples, verifying path consistency rather than serving as an independent mediation test. The control distinguishes Prompt-policy response from downstream response initiated directly at the allocation coordinate.

Table 2: Five-seed intervention audit on 100 paired examples per seed. Panel A reports the natural Dynamic-to-Uniform Prompt-control contrast. Panel B reports direct effective-allocation replacement and fixed-effective-allocation blocking. Numerical magnitudes are compared within, but not divided across, intervention families because their locations and perturbation scales differ.

Panel A: Natural Prompt-Control Response

Seed	A_m	A_w	NZ	$S_{10^{-5}}$	A_u	A_r	A_ℓ	Route sw.
20250917	0.098847	7.10×10^{-5}	31	15	0.001991	0.001181	0.20994	0
20250918	0.041677	9.97×10^{-5}	20	14	0.002666	0.001582	0.20470	0
20250919	0.054593	2.85×10^{-4}	29	27	0.008109	0.004811	0.40505	0
42	0.037811	5.69×10^{-5}	14	8	0.001410	0.000837	0.13386	0
7	0.030392	2.37×10^{-4}	29	29	0.006911	0.004100	0.34646	0
Mean	0.052664	1.50×10^{-4}	24.6	18.6	0.004217	0.002502	0.26000	0
Seed SD	0.027267	1.04×10^{-4}	7.0	8.9	0.003068	0.001820	0.11179	0

Panel B: Direct allocation response and fixed-effective-allocation blocking. Direct responses are means over 100 paired examples. Blocking columns report exact nonzero counts.

	Direct Replacement (Mean)				Blocking (Nonzero Count)			
Seed	A_w	A_u	A_r	A_ℓ	A_w	A_u	A_r	A_ℓ
20250917	1.5836	37.0628	21.9912	277.65	0	0	0	0
20250918	1.5855	35.8686	21.2826	278.77	0	0	0	0
20250919	1.5186	37.8794	22.4750	368.56	0	0	0	0
42	1.5863	35.5282	21.0804	315.72	0	0	0	0
7	1.5036	38.1379	22.6288	414.65	0	0	0	0

Table 3: Same-checkpoint Prompt-input ON/OFF audit on 100 Q-Bench examples with probability-based diagnostics. "Nonzero" reports exact nonzero values in the stored float32 audit statistics; effect magnitude and prevalence are reported separately.

Stage	Mean	Median	p05	p95	Nonzero
Allocation A_w	0.003032	0.003036	0.001439	0.005091	100/100
Actuation A_u	0.080040	0.080314	0.038981	0.133560	100/100
Representation A_r	0.047490	0.047653	0.023128	0.079247	100/100
Logits A_ℓ	2.219840	2.196731	1.739846	2.982045	100/100
MCQ distribution A_p	0.000275	1.29×10^{-6}	4.53×10^{-8}	0.001588	100/100
Routing argmax change	0	0	0	0	0/100
MCQ prediction change	0	0	0	0	0/100

5.5 Prompt-Input Zero-Replacement Sensitivity

This subsection evaluates a second replacement direction at the Prompt-derived allocator input. To evaluate the sensitivity of the allocator input, we conduct a complementary zero-replacement intervention on a trained checkpoint. While the ON condition retains the Prompt-derived conditioning state $\mathbf{m}(\mathbf{h})$, the zeroing condition substitutes this input with a zero tensor, holding all remaining parameters, inputs, and evaluation examples fixed.

For each intervention stage, we quantify the operational response using the stage-specific distances:

$$A_w = \|\mathbf{w}_{\text{on}} - \mathbf{w}_{\text{off}}\|_1, \quad A_u = \|\mathbf{u}_{\text{on}} - \mathbf{u}_{\text{off}}\|_2, \quad A_r = \|\mathbf{r}_{\text{on}} - \mathbf{r}_{\text{off}}\|_2, \quad A_\ell = \|\boldsymbol{\ell}_{\text{on}} - \boldsymbol{\ell}_{\text{off}}\|_2, \tag{12}$$

$$A_p = \|\mathbf{p}_{\text{on}} - \mathbf{p}_{\text{off}}\|_1, \quad A_y = \mathbf{1}\left[\arg\max \mathbf{p}_{\text{on}} \neq \arg\max \mathbf{p}_{\text{off}}\right].$$

5.6 Continuous Response and Boundary Crossing

This subsection characterizes how the continuous response relates to representation geometry and the observed decision margins. The normal condition was evaluated twice with distinct stochastic seeds to estimate repeat variation. ON-repeat allocation JSD was zero, whereas Prompt-input ON/OFF allocation JSD was 2.038×10^{-5}. Representation CKA Kornblith et al. (2019) was ≈ 1

Table 4: Repeat control and representation-level response in the same-model Edge audit.

Metric	Normal repeat	Prompt-input ON/OFF	Difference
Routing JSD	0.00000000	0.00002038	0.00002038
Representation CKA	≈ 1	0.99999988	$< 3\times10^{-7}$
Common-shift energy ratio	–	0.86152512	–

Table 5: Decision-margin audit for the 100 Prompt-input ON/OFF examples.

Quantity	Mean	Median	p05	p95
γ^{on}	0.955959	0.985513	0.861252	0.998747
$\gamma^{\mathrm{off}\|\mathrm{on}}$	0.956045	0.985744	0.861255	0.998749
$\Delta\gamma$	8.69×10^{-5}	2.38×10^{-7}	-6.12×10^{-4}	$+9.11\times10^{-4}$
A_p	2.75×10^{-4}	$+1.29\times10^{-6}$	$+4.53\times10^{-8}$	0.001588

numerical precision for the repeat comparison and 0.99999988 for the Edge ONOFF comparison. Both repeat and zeroing comparisons yielded CKA numerically indistinguishable from one at the reported precision.

The common-shift statistic is defined as

$$\rho_{\mathrm{common}} = \frac{\left\| \frac{1}{N} \sum_i \Delta r_i \right\|_2^2}{\frac{1}{N} \sum_i \|\Delta r_i\|_2^2}. \tag{13}$$

Linear CKA remains near one, indicating preserved centered cross-example geometry. The common-shift ratio of 0.8615 separately quantifies the fraction of displacement energy explained by the shared mean displacement.

The MCQ probability response was strongly right-skewed: its mean L_1 distance was 2.75×10^{-4}, while the median was only 1.29×10^{-6}. To distinguish distributional response from top-1 switching, we define the normal-condition decision margin, and evaluate the same choice identity under Edge OFF:

$$\gamma_i^{\mathrm{on}} = p_i^{\mathrm{on}}(y_i^{\mathrm{on}}) - \max_{j \neq y_i^{\mathrm{on}}} p_i^{\mathrm{on}}(j), \quad \gamma_i^{\mathrm{off}|\mathrm{on}} = p_i^{\mathrm{off}}(y_i^{\mathrm{on}}) - \max_{j \neq y_i^{\mathrm{on}}} p_i^{\mathrm{off}}(j). \tag{14}$$

As shown in Table 5, no evaluated example crossed the decision boundary, and top-1 predictions remained unchanged. The median relative margin change $|\Delta\gamma|/\gamma^{\mathrm{on}}$ of 6.81×10^{-7} (95th percentile: 1.63×10^{-3}) shows that the measured probability changes remain below the observed decision margins for most examples. All normal-condition predictions select option A, with high confidence relative to accuracy. The decision analysis therefore characterizes boundary crossing within this choice-concentrated checkpoint.

Evaluated on the same checkpoint, Q-Bench and Q-Bench2 accuracies reached 42% and 54% with average confidences of 0.978 and 0.986 (ECE: 0.558 and 0.446), placing the boundary-crossing result in a high-confidence, poorly calibrated operating regime. The readout module placed before the language-model head defines the explicit formulation:

$$\boldsymbol{\ell} = W_{\mathrm{LM}}\mathbf{r} + \mathbf{b}, \tag{15}$$

where $\mathbf{h}$ represents the normalized final hidden state and $\mathbf{r}$ the adapted state passed to the LM head without post-residual normalization, preserving the frozen-backbone interface. While edge zeroing altered continuous allocation across all 100 examples, the median L_1 distance was minimal (0.003036). The zero-input replacement produces continuous responses through choice probabilities, with zero observed dominant-index and MCQ boundary crossings in this checkpoint.

The Prompt-policy replacement changes the conditioning state across five seeds while retaining a zero-median allocation response and unchanged dominant indices. Direct allocation replacement produces nonzero downstream responses over a separate intervention family. Prompt-input zero replacement produces continuous downstream responses in one checkpoint without an observed dominant-index or MCQ boundary crossing.

Table 6: Matched-seed nominal-K probe. The comparison distinguishes response magnitude, exact nonzero support, thresholded support, and dominant-route identity in the instantiated system; it is not a general capacity-scaling experiment.

Metric	$K = 5$	$K = 10$	Interpretation
Mean Δ_m	0.098847	0.038442	Different upstream displacement
Mean Δ_w	7.10×10^{-5}	1.49×10^{-4}	Larger mean at $K = 10$
Median Δ_w	0	0	No broad differentiation
Nonzero Δ_w	31/100	27/100	Frequency did not increase
Mean A_u	0.001991	0.003987	Larger mean, single matched setting
Mean A_r	0.001181	0.002365	Larger mean, single matched setting
Mean Δ_ℓ	0.209938	0.204512	Nearly unchanged
Argmax changes	0/100	0/100	Dominant route unchanged

Table 7: Dataset structural audit for the 1,000-sample reasoning corpus.

Property	Value	Implication
Total rows	1,000	Scale of the reported case-study corpus
VCR subset	813	Majority component of the sampled corpus
VisualMRC subset	187	Factual / scientific subgroup
Choices present	1,000	Uniform MCQ structure
Reason flag	1,000	Reason field expected
Unique image paths	1,000	No existing same-image paired views
Existing paired views	0 confirmed	D_{view} must be constructed

6 Discussion and Limitations

In this audited controller, Dynamic-to-Uniform replacement alters the conditioning state across seeds while preserving a zero-median allocation and dominant indices. The core observation is that conditioning availability does not by itself establish realized control authority. Allocation support, realized actuation, and decision-boundary crossing must be measured at their corresponding controller coordinates. Across seeds, the allocation response exhibits a zero median, positive mean, and checkpoint-dependent support. In the reported internal audit, stored-nonzero allocation and actuation responses strongly co-vary ($\rho_{\text{Spearman}} = 0.9998$); for each fixed $\mathbf{h}$, $\Delta\mathbf{u} = \mathbf{E}(\mathbf{h})\Delta\mathbf{w}$. Direct allocation and prompt-input replacements shift coordinate-level responses, confirming the separation among conditioning variation, allocation, actuation, and decision boundaries. This hierarchical dissociation yields specific diagnostic targets: conditioning-to-allocation gaps require inspecting allocator coupling, saturation, hidden-state dominance, and replacement direction; allocation-to-actuation gaps necessitate examining transformation-output geometry and cancellation; and distribution-to-decision gaps evaluate operating margins. Although behavioral invariance does not preclude internal conditioning, finite interventions cannot isolate softmax saturation, weak coupling, or local insensitivity. The five-seed audit reveals recurrent response patterns with variable magnitudes, indicating that empirical evaluation must report both structural invariance and seed-level variability across independent operational stages rather than relying solely on task performance.

This analysis is instantiated on a single compact VLM and dataset regime. The reported recurrence reflects training seeds within this specific configuration, leaving architectural and scale transfer as open evaluation targets. The intervention set comprises centroid-policy, direct allocation, and zero-input replacements, while norm-matched random directions provide an additional specificity control. Fixed compositions, support region discrepancies, and high-confidence MCQ endpoints further bound the current operating regime.

7 Conclusion

We measured the inference-time response of a Prompt-conditioned degree of freedom in a VLM. Dynamic-to-Uniform replacement alters conditioning across seeds while maintaining zero-median allocation and dominant indices, while direct allocation and input replacements induce downstream responses at distinct coordinates. These findings reveal a recurrent dissociation between conditioning variation, allocation, actuation, and decision-boundary crossing.

REFERENCES

Anonymous. Auditing optimization and probabilistic reliability in frozen vision-language representation intervention. In *Under review*, - 2026a.

Anonymous. Decoding-aware view routing for stable frozen vlm adaptation. In *Under review*, 2026b.

Akari Asai, Mohammadreza Salehi, Matthew Peters, and Hannaneh Hajishirzi. ATTEMPT: Parameter-efficient multi-task tuning via attentional mixtures of soft prompts. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *EMNLP*, pp. 6655–6672, 2022.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. Visit-bench: A dynamic benchmark for evaluating instruction-following vision-and-language models. In *NeurIPS*, 2023.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Joon-Young Choi, Junho Kim, Jun-Hyung Park, Wing-Lam Mok, and SangKeun Lee. SMoP: Towards efficient and effective prompt tuning with sparse mixture-of-prompts. In *EMNLP*, pp. 14306–14316, 2023.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Prafulla Dhariwal, Beth Chen, Chris Olah, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2021/framework/index.html.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *NeurIPS*, pp. 9574–9586, 2021.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *JMLR*, 26(83):1–64, 2025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *ICLR*, 2022.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *ICML*, pp. 2790–2799, 2019.

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 *ICLR*. OpenReview.net, 2022.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Ka-Wei Lee. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *EMNLP*, pp. 5254–5276, 2023.

Noriaki Kawamae. Knowledge-aligned domain shift tuning for efficient adaptation in large language models. In *KDD*, pp. 1128–1138, 2025a.

Noriaki Kawamae. Multi level vision language adapter. In *ICDM*, pp. 1300–1309, 2025b.

Noriaki Kawamae. One-shot multi-view visual conversational recommendation. In *KDD*, pp. 1118–1127, 2025c.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, 114(13):3521–3526, 2017.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of neural network representations revisited. In *ICML*, pp. 3519–3529, 2019.

Alon Lavie and Abhaya Agarwal. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation, WMT@ACL 2007*, pp. 228–231, 2007.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre L. Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *ICLR*, 2025.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, pp. 3045–3059, 2021.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *NeurIPS*, 2023a.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. M^3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *CoRR*, abs/2306.04387, 2023b.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2022.

Chin-Yew Lin. Rouge: a package for automatic evaluation of summaries. In *ACL-workshop*, pp. 25–26, 2004.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. GPT understands, too. *CoRR*, abs/2103.10385, 2021.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`, 2022.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models. *CoRR*, abs/2504.05299, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *NeurIPS*, 2022.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pp. 311–318, 2002.

Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference. In *Proceedings of Machine Learning and Systems (MLSys)*, 2023.

Guanghui Qin and Jason Eisner. Learning how to ask: Querying LMs with mixtures of soft prompts. In *NAACL-HLT*, pp. 5203–5212, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, 2019.

Jiuding Sun, Sidharth Baskaran, Zhengxuan Wu, Michael Sklar, Christopher Potts, and Atticus Geiger. Hypersteer: Activation steering at scale with hypernetworks. *CoRR*, abs/2506.03292, 2025.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *CoRR*, abs/2308.10248, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In *NIPS*, pp. 5998–6008, 2017.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *ICLR*, 2024a.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. Reft: Representation finetuning for language models. In *NeurIPS*, pp. 63908–63962, 2024b.

Yan Xu, Mahdi Namazifar, Devamanyu Hazarika, Aishwarya Padmakumar, Yang Liu, and Dilek Hakkani-Tür. KILM: knowledge injection into encoder-decoder language models. In *ACL*, pp. 5013–5035, 2023.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *ICLR*, 2024.

Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *ICLR*, 2024a.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *ICLR*, 2020.

Zhengyan Zhang, Zhiyuan Zeng, Yankai Lin, Huadong Wang, Deming Ye, Chaojun Xiao, Xu Han, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Plug-and-play knowledge injection for pre-trained language models. In *ACL*, pp. 10641–10658, 2023.

Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin. Q-bench+: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):10404–10418, 2024b.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Y. Zhao, Andrew M. Dai, Zhifeng Chen, Quoc V. Le, and James Laudon. Mixture-of-experts with expert choice routing. In *NeurIPS*, pp. 7103–7114, 2022.

8 AI USE STATEMENT

We used these tools to edit the manuscript for grammar and readability, assist with literature searches and identify relevant literature, format references, and edit and format tables. All reported results were produced by the authors' experimental pipeline. We manually verified the cited literature, numerical results, code changes, and final claims.

A APPENDIX

A.1 PROSPECTIVE FIXED-STATE ANALYSIS

The finite-response formulation motivates a fixed-state conditional intervention. For a target representation $\mathbf{h}_i$, let

$$\mathbf{m}_i = A_{\mathbf{P}}(\mathbf{h}_i) \tag{16}$$

denote its observed prompt-derived state. A conditional intervention holds $\mathbf{h}_i$ fixed while replacing $\mathbf{m}_i$ with an alternative state $\widetilde{\mathbf{m}}$:

$$F_{\mathbf{h}_i}(\mathbf{m}_i) \quad \longrightarrow \quad F_{\mathbf{h}_i}(\widetilde{\mathbf{m}}). \tag{17}$$

Such an intervention characterizes conditional actuator response at a held backbone state.

Across five within-regime replications (detailed in Table 2 and Table 6), prompt composition consistently alters upstream conditioning while inducing sparse downstream responses. Direct allocation replacement yields larger downstream changes, confirming an active allocation-to-actuation path. In

Table 8: Each finite intervention supports a restricted response claim and leaves broader properties unidentified.

Control question	Intervention	Observable	Supported inference
Upstream state	Dynamic to Uniform	A_a, A_m	Tested composition contrast only
Allocation response	Dynamic to Uniform	A_w	Not the full reachable set
Direct response	Allocation replacement	A_u, A_r, A_ℓ	Not full actuator capacity
Replay consistency	Fixed allocation	A_u, A_r, A_ℓ	Deterministic downstream equality conditional on shared $\mathbf{h}$ and effective allocation
Prompt-input sensitivity	$\mathbf{m} \mapsto \mathbf{0}$ at allocator input	$A_w, \ldots, A_p$	Finite response to the zero-input replacement

the matched probe, $K = 10$ exhibits larger continuous responses despite smaller upstream displacement. Zeroing the router input generates nonzero continuous responses across all levels without altering dominant routing or top-1 decisions. These findings decouple response magnitude from route-identity switching, establishing intervention-relative boundaries rather than global reachability.

A.1.1 Conditional Response at a Fixed Backbone State

The evaluation set provides an observed collection of Prompt-derived states

$$\mathcal{M}_{\text{obs}} = \{A_{\mathbf{P}}(\mathbf{h}_j) : j \in \mathcal{D}_{\text{eval}}\}. \tag{18}$$

For a fixed target representation $\mathbf{h}_i$, the corresponding observed-donor displacement set is

$$\mathcal{U}_{\text{donor}}(\mathbf{h}_i) = \{\mathbf{u}(\mathbf{h}_i, \mathbf{m}_j) : \mathbf{m}_j \in \mathcal{M}_{\text{obs}},\ j \neq i\}. \tag{19}$$

When the router produces a simplex-valued weight vector, for example through softmax normalization, the resulting displacement satisfies

$$\mathbf{u}(\mathbf{h}_i, \mathbf{m}_j) \in \text{conv}\{E_1(\mathbf{h}_i), \ldots, E_K(\mathbf{h}_i)\} \tag{20}$$

This inclusion characterizes the convex-hull constraint induced by a simplex-valued router for a fixed representation. It does not imply that every point in the convex hull is reachable through the Prompt branch.

Because an observed donor state need not naturally co-occur with the target representation, donor substitutions are characterized by their Prompt-state norm, target–donor representation similarity, and distance within the observed Prompt-state set. These quantities describe the support of the evaluated intervention and are not treated as evidence that the substitutions constitute natural counterfactuals.

Prompt injection. Role-aware prompts are prepended to the input sequence using special tokens that are embedded and trained jointly with the adapter parameters. We define prompt tokens for visual context (e.g., `<vis_ctx>`), user queries (e.g., `<question>`), and system responses (e.g., `<answer>`), with each role-specific token initialized from the base vocabulary embedding and updated during fine-tuning.

Loss computation. Losses are computed at the token level with attention masks applied to exclude padding or prompt-only tokens. When enabled, the consistency loss and role-aware weights are incorporated into the final objective with minimal overhead. During ablation studies, these components can be toggled via configuration without modifying the core training loop.

A.2 IMPLEMENTATION

We implemented the framework in PyTorch 2.5.1+cu124 (CUDA 12.4). All adaptation methods are standard `nn.Modules` to integrate with the Trainer and generation interfaces with the Trainer API and the `generate()` interface. All experiments were executed on a single NVIDIA A100-SXM4-40GB GPU with `bfloat16`. We will release the codebase and configuration files to ensure reproducibility.

Automatic evaluation uses Hugging Face's PEFT Mangrulkar et al. (2022)[3] and Evaluate [4] libraries, reporting individual metrics as well as a standardized z-score composite (semantic/stability/cost) for secondary diagnostics. Q-Bench and Q-Bench+ measure multiple-choice accuracy, with predicted choices (A–D) extracted via a regex-based parser to handle minor formatting inconsistencies, augmenting accuracy with probabilistic scoring (ECE (Guo et al., 2017), Brier score (Brier, 1950), and NLL). VisIT-Bench reports lexical metrics (BLEU Papineni et al. (2002), ROUGE-L Lin (2004), METEOR Lavie & Agarwal (2007)), semantic metrics (BERTScore Zhang et al. (2020), SBERT cosine similarity Reimers & Gurevych (2019)), and exact match. Temporal stability is quantified using drift metrics. We treat adaptation as a design space across accuracy, semantic alignment, stability, and computational cost. Across discrete accuracy, semantic alignment, and temporal stability, generation latency is measured at `generated_tokens = 64` (excluding prefill) Pope et al. (2023). Training runtime is reported for PEFT baselines where gen64 entries are unavailable, whereas gen64 throughput is reported for KLAFT.

Implementation details are:

- **Prompt Formatting**: All prompts were formatted using the tokenizer.apply_chat_template method.
- **Output Normalization for Evaluation**: For the Q-Bench task, we implemented a custom, regex-based logic to handle minor variations in model output, correctly extracting the choice letter (A-D) even if the model's response was not perfectly formatted. This ensures that the evaluation accurately reflects the model's true performance.

[3] `https://github.com/huggingface/peft`

[4] `https://github.com/huggingface/evaluate`