APERTURE: FEW-SHOT REFOCUSING OF VISUAL EVIDENCE IN VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

In a new visual domain, a vision–language model may need to make many predictions using only a few labeled images. This situation arises, for example, when identifying tumor tissue at a new hospital or assessing building damage after a new disaster. We introduce APERTURE, a small, reusable module that adjusts image-token key/value states while keeping the backbone frozen. Gradients of the correct-answer loss on labeled support images identify sensitive directions in these states. The module learns a bounded transformation within those directions and applies it to each new image's activations, without demonstrations or further optimization. Across four BRIGHT events and three support sets per event, APERTURE increases Qwen2.5-VL's mean macro-F1 from 22.55% to 31.33% using only 1,024 learned coefficients, compared with 31.46% for LoRA-1E. On CAMELYON17–H2, it achieves lower mean raw candidate negative log-likelihood than LoRA-1E and LoRA-4E across the three evaluated Qwen model variants. In the measured Qwen2.5-VL configuration, its saved adaptation state, including the fixed bases, occupies approximately 1/291 of the storage required by rank-16 LoRA. These results show that APERTURE can achieve performance close to LoRA-1E on BRIGHT with substantially less adaptation storage, while its benefits vary across tasks and models.

1 INTRODUCTION

Pretrained vision–language models (VLMs) can use images and textual instructions to answer questions about visual content. Applying them in a new environment often requires adapting to conditions that differ from their training data. A pathology system entering a new hospital must identify tumor tissue despite changes in staining and image acquisition (Koh et al., 2021). After a disaster, a system may need to assess building damage using images collected before and after the event, with labels available for only a few buildings (Chen et al., 2025a). A robot moved to a new workspace may need to judge whether an action succeeded under unfamiliar viewpoints or backgrounds. In each case, a small set of labeled examples, called the support set, must help the model make predictions on many unseen images from the same target domain. Figure 1 illustrates this problem,rather than a specific method or experimental result.

One way to use these labels is to fine-tune the model. Parameter-efficient methods reduce the number of parameters trained during this process. LoRA, for example, keeps the pretrained weights fixed and learns small additional matrices that modify selected model computations (Hu et al., 2022; Zanella & Ben Ayed, 2024). These matrices can still contain millions of trainable variables, and few-shot results can depend on which support examples are used and how long adaptation is run (Mrabah et al., 2025; Peng et al., 2025). This motivates studying whether useful adaptation can be concentrated in a much smaller learned component.

To identify such a component, we focus on how the model uses its internal visual representations. Research on visual grounding shows that adjusting the influence of image evidence can improve predictions (Leng et al., 2024; Huang et al., 2024), while methods such as ReFT demonstrate the value of learning interventions on intermediate states (Wu et al., 2024). We focus on image-token key/value (K/V) states: keys influence how attention is allocated to visual evidence, and values determine the information aggregated from it. This provides a concrete location for visual correction,

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.

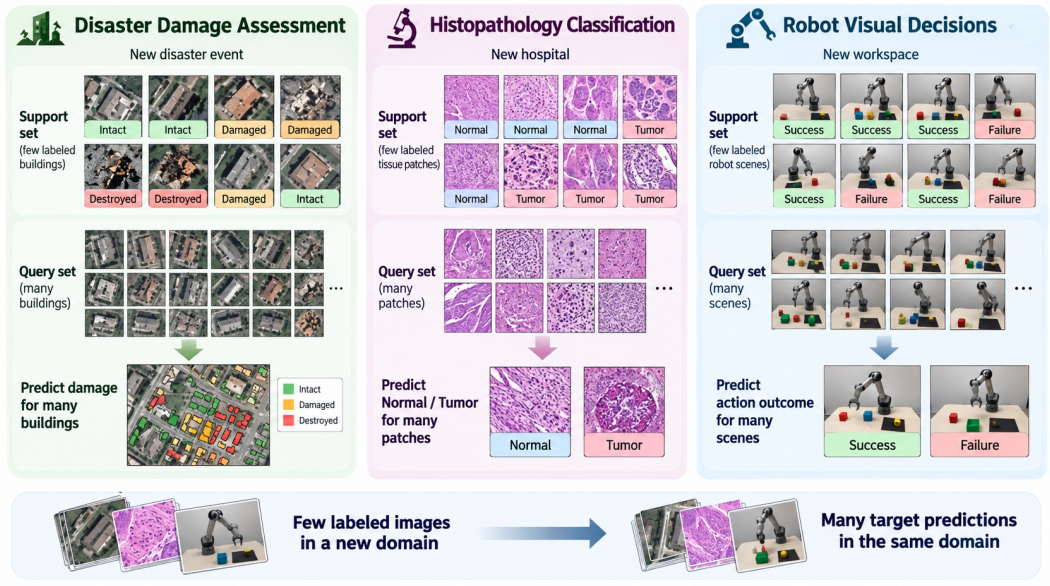

Figure 1: **Few labels must support many visual decisions.** Three examples of target-domain adaptation: building-damage classification after a disaster, normal/tumor tissue classification at a new hospital, and visual success/failure verification in a new robot workspace. Support and query images belong to the same target domain; each domain is adapted independently. Thumbnails, labels, and the map are illustrative, not experimental predictions.

but leaves a design question: which directions in these states should be modified, and how can a few labeled examples guide that choice?

We introduce APERTURE, a task-specific *internal visual lens* that combines support-guided direction selection with a bounded correction applied to visual K/V states. It is learned in two stages. First, gradients of the correct-answer loss on the support images identify directions to which the loss is sensitive. Second, we keep these directions fixed and train a small module to strengthen, attenuate, or mix activation components within them. Support labels thus serve two roles: they determine the space in which adaptation takes place and supervise the correction learned within that space. This combination specifies both what the intervention can modify and how its modification directions are chosen. The module uses a bounded residual transformation, is initialized to make no change to the model, and adds corrections directly only to image-token states. The backbone weights remain fixed throughout.

Once fitted, the module is reused across new images from the target domain. Each image supplies its own activations, so the same transformation can produce different corrections for different inputs. Predictions require neither appended demonstrations nor query-time optimization. In our two-layer configuration, four 16×16 controller matrices contain only 1,024 learned coefficients. The fixed direction vectors are stored alongside these coefficients and are included in our resource measurements.

Across four BRIGHT events and three support sets per event, APERTURE increases Qwen2.5-VL's mean macro-F1 from 22.55% to 31.33%, compared with 31.46% for LoRA-1E. On CAMELYON17–H2, it achieves lower mean raw candidate negative log-likelihood than both evaluated LoRA schedules on the three evaluated Qwen model variants. In the measured Qwen2.5-VL configuration, its saved adaptation state, including the fixed direction vectors, occupies approximately 1/291 of rank-16 LoRA's storage. These gains vary across tasks and models. We organize the evaluation around classification performance (**RQ1**), candidate-answer probabilities (**RQ2**), adaptation costs (**RQ3**), and the effects of direction choice, intervention location, visual input, and residual strength (**RQ4**).

Our contributions are **(1)** a reusable visual K/V module that uses support gradients to select sensitive directions and learns a bounded transformation within them; **(2)** an evaluation of prediction quality,

storage, fitting time, and inference cost under explicit few-shot budgets; and **(3)** controlled analyses that clarify the choices affecting the module's performance and the limits of its benefits across tasks and models.

2 Related work

Weight adaptation and representation intervention. LoRA learns low-rank weight updates while keeping the pretrained weights frozen (Hu et al., 2022). Few-shot VLM studies examine the effectiveness and generalization of weight adaptation under limited supervision (Zanella & Ben Ayed, 2024; Mrabah et al., 2025; Peng et al., 2025). ReFT and LoFiT learn localized interventions on intermediate representations (Wu et al., 2024; Yin et al., 2024). APERTURE follows this representation-intervention approach with a specific design: it directly modifies only image-token K/V states, selects fixed directions from labeled support examples, and learns a bounded residual transformation within those directions. This combination defines the location, direction space, and form of the learned adaptation.

Inference-time control. Test-time training and test-time prompt tuning use self-supervised or label-free objectives to adapt models under distribution shift (Sun et al., 2020; Shu et al., 2022). V-Reason optimizes value-cache corrections using an entropy objective for multimodal reasoning (Sridhar et al., 2026). APERTURE uses labeled examples from the target domain to fit a controller before query evaluation. The controller is fitted once and reused across subsequent images, with no optimization on individual queries. This places our method in supervised few-shot adaptation, where the support set provides the supervision for later predictions.

Visual grounding and hallucination mitigation. VCD contrasts predictions from original and distorted images, while OPERA intervenes during decoding to mitigate over-trust (Leng et al., 2024; Huang et al., 2024). ClearSight, ICT, prefill-time intervention, and CausalLens also adjust visual contributions or evidence-sensitive representations (Yin et al., 2025; Chen et al., 2025b; Zhang et al., 2026; Ji et al., 2026). These studies motivate changing how visual information influences a model's answers. We examine this idea in supervised target-domain adaptation: a few labeled examples are used to learn a transformation that can be reused on new images. Our evaluation focuses on classification performance and finite-candidate answer probabilities, rather than open-ended hallucination rates.

3 Reusable visual residual control

3.1 Task and intervention interface

A target domain e provides labeled support $S_e = \{(x_i, y_i)\}_{i=1}^{N}$ and disjoint queries Q_e. Each input contains an image or image pair and a question. The pretrained VLM f_θ is fixed. For selected decoder projection $j = (\ell, t)$, with $t \in \{K, V\}$, let $Z_{i,j} \in \mathbb{R}^{T_i \times d_j}$ be its output before head reshaping and architecture-specific normalization or positional transforms. A broadcast row mask M_i identifies visual tokens: the post-event image in BRIGHT and the image rows in single-image tasks. Text rows receive no direct residual write, although their downstream states respond to attention over edited visual evidence.

3.2 Measure sensitivity, then learn the correction

The first stage in Figure 2 uses support labels to identify visual K/V directions that affect the correct-answer loss. For answer-token span $\mathcal{T}_i$,

$$L_i = -\frac{1}{|\mathcal{T}_i|} \sum_{u \in \mathcal{T}_i} \log p_\theta(y_{i,u} \mid x_i, y_{i,<u}), \qquad G_{i,j} = \left[\frac{\partial L_i}{\partial Z_{i,j}} \right]_{M_i}. \tag{1}$$

Each selected K or V module receives a separate orthonormal basis:

$$C_j = \frac{1}{N} \sum_i G_{i,j}^\top G_{i,j}, \qquad B_j = \mathrm{TopEig}_r(C_j), \qquad B_j^\top B_j = I_r. \tag{2}$$

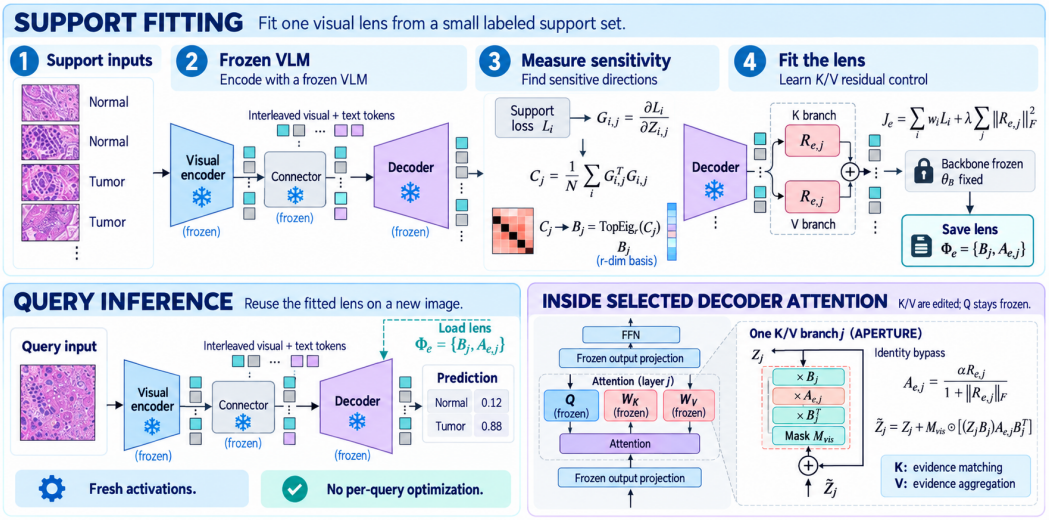

Figure 2: **Fit one lens on support; reuse it on fresh query activations.** Top: the same support images first identify sensitive directions and then fit controller matrices in the fixed bases. Bottom: the stored lens is reused without query-time optimization. The attention detail places separate residuals after K/V projection, retaining the identity path and leaving Q unedited. Backbone weights remain fixed throughout. Image patches are illustrative.

The uncentered second moment retains sensitivity even if examples request opposing changes along a shared direction. Its leading eigenspace retains the greatest observed support-gradient energy at rank r (Appendix A). These directions define where the controller can act; the second stage learns how to adjust the corresponding activations.

We keep B_j fixed and learn a dense controller $R_{e,j} \in \mathbb{R}^{r \times r}$:

$$\widetilde{Z}_{i,j} = Z_{i,j} + M_i \odot [(Z_{i,j} B_j) A_{e,j} B_j^\top], \qquad A_{e,j} = \frac{\alpha R_{e,j}}{1 + \|R_{e,j}\|_F}. \tag{3}$$

Only $R_{e,j}$ is optimized. Starting at zero exactly recovers the unchanged model. Full mixing can strengthen, attenuate, or exchange components within the selected directions; the identity branch retains the original state, including components outside this space. The bound $\|A_{e,j}\|_2 < \alpha$ limits the local transformation.

3.3 Support Fitting and Query Reuse

The second stage minimizes

$$\mathcal{J}_e(R) = \sum_{i=1}^{N} w_i L_i(f_{\theta,B,R}) + \lambda \sum_j \|R_{e,j}\|_F^2 \tag{4}$$

using Adam and gradient clipping. Each update accumulates the complete support. BRIGHT uses $w_i = 1/N$; the recorded single-image recipe uses $w_i = 1$. After fitting, both B_j and $A_{e,j}$ are fixed. Candidate answer c is scored with its recorded autoregressive score $s_c(x)$, giving $p(c \mid x) = \mathrm{softmax}(s(x))_c$. Algorithm 1 makes the sequential fitting and subsequent reuse explicit.

The fitted lens acts on each query's own activations, producing an input-dependent correction through $Z_{i,j} B_j$. For one module, let $D_M = \mathrm{diag}(M)$ and $H = \partial L / \partial \widetilde{Z}$. Holding the incoming state fixed, the coefficient derivative is

$$\frac{\partial L}{\partial A} = (D_M Z B)^\top H B. \tag{5}$$

The controller therefore combines support-image activations with answer-loss gradients to learn a reusable visual correction. The residual $\Delta Z = D_M Z B A B^\top$ satisfies $\Delta Z (I - B B^\top) = 0$ and $\|\Delta Z\|_F \le \alpha \|D_M Z B\|_F$. Appendix A gives the derivations and their local scope.

Algorithm 1 Aperture: domain-level fitting and query prediction

Input: Frozen f_θ; support S_e; sites $\mathcal{I}$; masks M_i; rank r;
bound α, regularizer λ, weights w_i, steps T, rate η, clip c.
Output: Reusable lens $\Phi_e = \{B_j, A_{e,j}\}_{j\in\mathcal{I}}$.

$C_j \leftarrow 0$ for every $j \in \mathcal{I}$
for $(x_i, y_i) \in S_e$ **do**
  $L_i \leftarrow \mathrm{AnswerLoss}(f_\theta, x_i, y_i)$
  for $j \in \mathcal{I}$ **do**
    $G_{i,j} \leftarrow [\nabla_{Z_{i,j}} L_i]_{M_i}$; $C_j \leftarrow C_j + G_{i,j}^\top G_{i,j}$
  end for
end for
for $j \in \mathcal{I}$ **do**
  $B_j \leftarrow \mathrm{stopgrad}(\mathrm{TopEig}_r(C_j/N))$; $R_{e,j} \leftarrow 0_{r\times r}$
end for
Initialize Adam state for $R = \{R_{e,j}\}_{j\in\mathcal{I}}$
for $t = 1, \ldots, T$ **do**
  $A_{e,j}(R) \leftarrow \alpha R_{e,j}/(1 + \|R_{e,j}\|_F)$ for all j
  $\mathcal{J}_e(R) \leftarrow \sum_i w_i L_i(f_{\theta,B,R}) + \lambda \sum_j \|R_{e,j}\|_F^2$
  $g \leftarrow \mathrm{clip}_c(\nabla_R \mathcal{J}_e(R))$
  $R \leftarrow \mathrm{Adam}(R, g; \eta)$ ▷ Only controller matrices change
end for
$\Phi_e \leftarrow \{B_j, \mathrm{stopgrad}(\alpha R_{e,j}/(1 + \|R_{e,j}\|_F))\}_{j\in\mathcal{I}}$
return Φ_e

procedure PREDICT$(x, \mathcal{C}(x), \Phi_e)$
  for $c \in \mathcal{C}(x)$ **do**
    $s_c \leftarrow \mathrm{AnswerScore}(f_{\theta,\Phi_e}, x, c)$ ▷ Use Eq. 3
  end for
  return $\arg\max_c s_c,\ \mathrm{softmax}(s)$
end procedure

For m sites, Full control learns mr^2 coefficients and stores $r\sum_j d_j$ basis values. Two K/V layers have four sites and hence 1,024 learned coefficients at rank 16. Query overhead is $O(T_i d_j r + T_i r^2)$ per site with no new input tokens. This accounting motivates measuring basis storage and fitting cost separately from the optimized-variable count.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Tasks and evaluation protocols. We evaluate whether a visual correction learned from a small labeled support set improves predictions on unseen images from the same target domain. Our main classification experiment uses four BRIGHT events: Hawaii wildfire, Libya flood, Noto earthquake, and Turkey earthquake (Chen et al., 2025a). For each event, Qwen2.5-VL-7B is adapted using 24 labeled buildings and evaluated on 300 balanced queries, with tile-disjoint support and query sets and three support seeds. Additional Qwen3-VL-8B and InternVL3-8B results use separately recorded support24/query300 splits (Bai et al., 2025b;a; Zhu et al., 2025). These experiments evaluate building-level classification. For pathology, we distinguish two CAMELYON17–H2 protocols: a Qwen2.5-VL study with three support seeds and 300 queries, and a cross-backbone study with two support seeds and 200 queries (Koh et al., 2021). Both allocate 12 labels to model adaptation and four to calibration. The cross-backbone study evaluates Qwen2.5-VL-7B, Qwen3-VL-4B/8B, and Gemma 3 4B IT (Gemma Team, 2025), and also includes PathMNIST–224 with 54 fitting labels, 18 calibration labels, and 180 queries (Yang et al., 2023). H2 splits are patient-disjoint; PathMNIST splits are image-content-disjoint because patient identifiers are unavailable. Supplementary robotics experiments cover execution verification and manipulation question answering (Pacaud et al., 2026; Zhao et al., 2025). Each domain is adapted independently, and Table 4 in Appendix B summarizes the separate evaluation protocols.

Table 1: **Few-shot building-damage classification on BRIGHT (RQ1). Panel A:** matched Qwen2.5-VL-7B comparisons using 24 support examples and 300 queries per event. Entries are macro-F1 (%) mean ± sample SD over three support seeds. The final column averages the four event means; the random-basis aggregate is approximate because it is calculated from rounded event means. **Panel B:** separately recorded four-event means on two additional backbones, using their archived LoRA comparators. Support-seed SD was not recorded for these comparisons. Higher is better.

Panel A: Qwen2.5-VL-7B, matched protocol

Method	Hawaii	Libya	Noto	Turkey	4-event mean
Frozen	$19.4_{\pm 0.0}$	$27.3_{\pm 0.0}$	$20.6_{\pm 0.0}$	$22.9_{\pm 0.0}$	22.55
LoRA-1E	$35.0_{\pm 3.8}$	$\mathbf{35.9}_{\pm 2.9}$	$31.3_{\pm 2.9}$	$23.6_{\pm 0.9}$	**31.46**
LoRA-4E	$28.4_{\pm 3.8}$	$24.4_{\pm 3.8}$	$26.8_{\pm 3.9}$	$\mathbf{29.9}_{\pm 3.0}$	27.39
Random basis	$\mathbf{36.7}_{\pm 2.8}$	$30.5_{\pm 4.4}$	$28.0_{\pm 7.8}$	$25.5_{\pm 2.0}$	≈ 30.2
APERTURE	$33.3_{\pm 4.4}$	$34.0_{\pm 3.2}$	$\mathbf{32.4}_{\pm 2.7}$	$25.6_{\pm 3.5}$	31.33

Panel B: additional backbones, four-event mean macro-F1 (%)

Backbone	Frozen	LoRA	APERTURE
Qwen3-VL-8B	16.67	20.34	**34.92**
InternVL3-8B	18.48	16.67	**28.60**

Adaptation and training budgets. The main APERTURE configuration uses rank 16, K/V projections at decoder layers 14 and 27, full mixing, $\alpha = 3$, four full-support updates, a learning rate of 0.05, and $\lambda = 0.001$. LoRA uses rank 16, scaling 32, and decoder Q/K/V/O projections. LoRA-1E and LoRA-4E denote one and four complete passes over the fitting support set. Methods share the total labeled-support budget and query set within each comparison, while their optimizer updates and computational costs differ. In particular, APERTURE requires an additional support pass for direction estimation. We report both fixed LoRA schedules in the main tables and treat the number of support epochs as an adaptation hyperparameter. Each backbone retains its native processor, with image preprocessing and complete training budgets specified in Appendix B.

Controls and metrics. Frozen provides the unadapted reference. The random-basis control retains the same trainable visual controller as APERTURE but replaces the support-derived directions with random directions. Output Bias fits candidate-score offsets without modifying internal representations, while positive temperature scaling adjusts candidate probabilities without changing a fixed model's predicted class. These controls distinguish direction selection, internal adaptation, and changes to output scores. Their fitting and calibration allocations remain within the stated total label budgets. We report macro-F1 for classification and raw candidate negative log-likelihood (NLL) and Brier score for probability prediction. Here raw denotes probabilities before temperature scaling. Means and sample standard deviations summarize variation across support seeds with fixed query sets. We report raw and temperature-scaled results separately because an improvement in candidate probabilities need not coincide with an improvement in classification.

4.2 RQ1: DOES THE LEARNED CORRECTION IMPROVE TARGET-DOMAIN DECISIONS?

Under the matched three-support-seed BRIGHT protocol, APERTURE improves mean macro-F1 over Frozen on all four events, increasing the four-event average from 22.55% to 31.33% (+8.78 percentage points; Table 1). Its mean is close to LoRA-1E at 31.46% and exceeds LoRA-4E at 27.39%, using 1,024 learned coefficients. The gains therefore show that a correction fitted to a small support set can improve predictions on unseen images from the same event. The relative performance of the two LoRA schedules also demonstrates why the training budget must be specified: four support epochs do not consistently improve on one. The random-basis controller is also competitive, indicating that learning a compact visual correction accounts for a substantial part of the benefit. We examine the additional contribution of support-derived directions in Section 4.5.

The additional backbone comparisons provide complementary evidence: APERTURE reaches 34.92% on Qwen3-VL-8B and 28.60% on InternVL3-8B, exceeding both Frozen and the corresponding LoRA comparator in each case. These comparisons use separate splits and are not pooled with the matched

Table 2: **Classification and candidate probability prediction on H2 (RQ2). Panel A:** Qwen2.5-VL-7B with three support seeds and 300 queries; entries are mean ± sample SD. **Panel B:** mean raw candidate NLL under the separate two-support-seed, 200-query protocol. Both panels allocate 12 labels to APERTURE and LoRA fitting and four to calibration. Output Bias fits candidate-score offsets within the same total label budget. NLL and Brier scores shown here are computed before temperature scaling. Full cross-backbone SDs and temperature-scaled results appear in Appendix C.3.

Panel A: Qwen2.5-VL-7B, three-support-seed H2

Method	F1 (%) ↑	Raw NLL ↓	Brier ↓
Frozen	$54.21_{\pm 0.00}$	$0.6669_{\pm 0.0000}$	$0.4746_{\pm 0.0000}$
Output Bias	$59.42_{\pm 2.75}$	$0.6526_{\pm 0.0031}$	$0.4608_{\pm 0.0030}$
LoRA-1E	$41.41_{\pm 11.11}$	$0.7371_{\pm 0.0767}$	$0.5365_{\pm 0.0620}$
LoRA-4E	$\mathbf{64.29}_{\pm 7.37}$	$1.3259_{\pm 0.5812}$	$0.5782_{\pm 0.1274}$
APERTURE	$61.66_{\pm 9.29}$	$\mathbf{0.6076}_{\pm 0.0198}$	$\mathbf{0.4227}_{\pm 0.0183}$

Panel B: raw candidate NLL across backbones on H2

Method	Qwen2.5-VL 7B	Qwen3-VL 4B	Qwen3-VL 8B	Gemma 3 4B
Frozen	0.812	1.952	1.669	2.060
Output Bias	0.797	1.860	0.959	**0.735**
LoRA-1E	0.846	1.494	1.022	1.667
LoRA-4E	3.062	2.685	1.489	3.542
APERTURE	**0.774**	**1.467**	**0.865**	2.267

Qwen2.5-VL experiment. The broader evaluation also identifies limits to the classification gains. On PathMNIST, APERTURE improves over Frozen on all four evaluated backbones, but LoRA-4E remains stronger. On H2, the fixed controller does not improve classification on every backbone, and LoRA also retains substantial advantages on the manipulation tasks. Thus, the BRIGHT results establish useful adaptation in this setting, while the supplementary tasks show that the same correction recipe is not uniformly effective. Complete comparisons are provided in Appendices C, C.3, and G.

4.3 RQ2: How does adaptation affect candidate-answer probabilities?

We next examine whether adaptation changes the probability assigned to the correct candidate, alongside its effect on class selection. Table 2, Panel A reports the three-support-seed Qwen2.5-VL H2 study. APERTURE increases macro-F1 from 54.21% to 61.66%, reduces raw candidate NLL from 0.6669 to 0.6076, and reduces Brier score from 0.4746 to 0.4227. It also improves all three metrics over Output Bias, showing a benefit beyond that score-offset control in this protocol. LoRA-4E achieves higher F1 at 64.29%, but has higher raw NLL and Brier score. This comparison separates two outcomes of adaptation: selecting the correct class more often and assigning more useful probabilities across the candidate answers. APERTURE has the lowest mean raw NLL and Brier score among the displayed methods, while LoRA-4E provides the strongest classification result.

Under the separate cross-backbone H2 protocol, APERTURE achieves the lowest mean raw candidate NLL on all three Qwen backbones, while Output Bias performs best on Gemma (Panel B). These raw-NLL gains do not establish a general advantage over calibrated weight adaptation. Temperature scaling changes some rankings: on Qwen2.5-VL, LoRA-4E reaches a calibrated NLL of 0.632, below APERTURE's raw NLL of 0.774; on Qwen3-VL-8B, calibrated LoRA-1E reaches 0.847, compared with 0.865 for APERTURE. The complete raw and calibrated results are reported in Appendix C.3. Together, the two H2 studies show that the learned visual correction can improve candidate probability prediction, but the extent of that benefit depends on the backbone, protocol, and whether output calibration is applied.

4.4 RQ3: What is the cost of fitting and reusing a lens?

Table 3 accounts for both the learned controller and its fixed direction vectors in the measured Qwen2.5-VL configuration. APERTURE optimizes 1,024 coefficients, compared with 10,092,544 for rank-16 LoRA, and stores an additional 32,768 fixed-basis values. Its complete serialized adaptation

Table 3: **Adaptation storage and measured computational cost (RQ3).** Qwen2.5-VL-7B on Hawaii support seed 0, measured on an RTX 4090. State size includes all serialized task-specific parameters and fixed bases. It excludes backbone storage and does not measure peak GPU memory. Fit time includes direction extraction for APERTURE. Query latency is the measured mean time per query.

Method	Learned coefficients	State (KiB)	Fit (s)	Query (ms)
Frozen	0	0	—	485.1
LoRA-1E	10,092,544	39,462.46	7.92	512.3
LoRA-4E	10,092,544	39,462.46	25.06	511.0
APERTURE	1,024	135.43	22.14	506.8

state occupies 135.43 KiB, approximately 1/291 of LoRA's storage. Query latency is 506.8 ms, about 4.5% above Frozen and close to the measured LoRA latencies. Fitting takes 22.14 s including direction extraction, compared with 7.92 s for LoRA-1E and 25.06 s for LoRA-4E. The principal resource advantage in this configuration is therefore the small amount of state stored per domain; fewer optimized coefficients do not guarantee faster fitting. Once fitted, the lens is reused across subsequent queries without demonstration tokens or per-query backward passes.

4.5 RQ4: WHICH ASPECTS OF THE VISUAL CORRECTION AFFECT ITS USEFULNESS?

The benefit of support-derived directions varies across events. We first test the direction-selection stage while keeping the trainable controller and residual construction fixed. In the matched rank-16 BRIGHT comparison, support-derived directions outperform random directions on Libya (34.0% versus 30.5%) and Noto (32.4% versus 28.0%), give similar performance on Turkey (25.6% versus 25.5%), and underperform on Hawaii (33.3% versus 36.7%). Both controllers improve mean F1 over Frozen on every event. These results support the value of learning a compact correction within visual K/V states, while showing that the additional benefit of gradient-based direction selection is task-dependent. A separate rank-5 comparison finds a larger advantage for uncentered gradients over centered gradients and random directions (Table 15); it provides evidence under a different configuration and is reported separately. The gradient basis should therefore be understood as a support-informed choice of correction space whose usefulness is evaluated empirically.

Intervention location has task-dependent effects. We next examine whether changing the intervention layers improves the same form of visual correction. On Turkey, using layer 14 alone increases three-support-seed mean F1 from 25.57% to 32.12%, while reducing the number of learned coefficients from 1,024 to 512. Applying this same one-layer rule to all four BRIGHT events raises their mean from 31.33% to 32.07%, with gains on Noto and Turkey and decreases on Hawaii and Libya (Table 16). These results show that increasing the number of intervention sites does not necessarily improve performance, and that a useful placement can differ across target domains. The follow-ups are exploratory and do not replace the fixed configuration used in the main comparisons.

Adaptation gains depend on the visual input. We hold fitted controllers fixed and compare their gains over Frozen on real images and controlled inputs. These separate seed-0 experiments use Qwen2.5-VL on Hawaii and Libya and InternVL3 on Camelyon. For a control input c, we define

$$\begin{aligned} D_c = &[F_1(\text{APERTURE}, \text{real}) - F_1(\text{Frozen}, \text{real})] \\ &- [F_1(\text{APERTURE}, c) - F_1(\text{Frozen}, c)]. \end{aligned} \tag{6}$$

Here F_1 is expressed on the unit interval, and a positive D_c indicates a larger adaptation gain on real images than on the control. Neutral-image contrasts correspond to 16.86, 9.65, and 21.23 F1 percentage points on Hawaii, Libya, and Camelyon, respectively. Shuffle contrasts are smaller. Output Bias also produces positive neutral-image contrasts, since its adjusted scores retain the frozen model's dependence on the image (Appendix G). The results therefore establish that the measured adaptation gain changes with visual input, while leaving open which particular visual features or decision biases account for that change.

The fitted residual contributes to candidate probability prediction. Finally, we scale an already fitted residual as $Z + \gamma\Delta Z$ without retraining, directly testing the contribution of the learned

correction. On fixed seed-0 subsets of 60 queries from each of eight domains, full strength ($\gamma = 1$) yields lower NLL than disabling the residual ($\gamma = 0$) in all eight cases, and lower NLL than half strength in six (Table 19). This links the fitted residual to the observed probability improvements on these subsets. Its preferred strength nevertheless depends on the domain and metric. On the H2 subset, full strength lowers NLL but reduces F1 from 56.94% to 48.41%, whereas half strength attains 66.06% F1. The experiment thus reinforces the distinction between candidate probability prediction and class selection: the learned correction contributes to prediction, but its optimal magnitude cannot be inferred from direction sensitivity alone.

5 Discussion and conclusion

We introduced APERTURE, a reusable controller that learns task-specific corrections to visual K/V states from a small labeled support set. Support-answer gradients identify sensitive directions, and a bounded residual transformation learns how to modify activation components within those directions while keeping the backbone frozen. On the three-support-seed BRIGHT evaluation, APERTURE improves the four-event mean macro-F1 from 22.55% to 31.33%, close to LoRA-1E at 31.46%, using 1,024 learned coefficients. Its measured adaptation state, including the fixed bases, occupies approximately 1/291 of the storage required by the corresponding rank-16 LoRA adapter. These results demonstrate that limited supervision can produce a compact correction that remains useful when applied to new images from the same target domain.

The benefits depend on the task, backbone, and evaluation metric. On H2, APERTURE achieves lower mean raw candidate NLL than both LoRA schedules on the three evaluated Qwen backbones, but this advantage does not consistently extend to classification or comparisons after temperature scaling. Random-basis controllers also remain competitive, showing that support-derived directions are not uniformly preferable. The placement and residual-strength experiments further indicate that the usefulness of a correction depends on where it is applied and how strongly it modifies the visual states. These findings support supervised visual-state correction as a practical adaptation approach while motivating further study of how limited support labels can guide the choice of directions, locations, and correction strength.

AI USE STATEMENT

Generative AI assisted research discussion, paper polishing and manuscript revision. We do not use it for the paper design and method design.

REPRODUCIBILITY STATEMENT

REPRODUCIBILITY STATEMENT

Parameter counts, serialized adaptation-state sizes, fitting time, and inference overhead are reported separately in Appendix E, since a small number of optimized coefficients does not by itself imply lower fitting cost.

ETHICS AND REPRODUCIBILITY STATEMENTS

These offline evaluations do not establish clinical safety, disaster-response readiness, or robot-control reliability. The appendix specifies data roles, fitting budgets, scoring, and diagnostic scope. The source package retains the recorded summaries, visualizations, and verification scripts; reproducing model evaluations requires the corresponding model and image assets.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Hongruixuan Chen, Jian Song, Olivier Dietrich, Clifford Broni-Bediako, Weihao Xuan, Junjue Wang, Xinlei Shao, Yimin Wei, Junshi Xia, Cuiling Lan, Konrad Schindler, and Naoto Yokoya. BRIGHT: A globally distributed multimodal building damage assessment dataset with very-high-resolution for all-weather disaster response. *arXiv preprint arXiv:2501.06019*, 2025a.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. ICT: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4209–4221, 2025b.

Gemma Team. Gemma 3. *Technical report*, 2025.

Guangya Hao, Yitong Shang, Yunbo Long, Zhuokai Zhao, and Hanxue Liang. Self-policy distillation via capability-selective subspace projection. *arXiv preprint arXiv:2605.22675*, 2026.

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.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

Junyang Ji, Qifan Liu, Wenming Yang, and Zhihai He. CausalLens: Sensitivity-guided multi-head causal intervention for hallucination mitigation in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4199–4209, 2026.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2021.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Nairouz Mrabah, Nicolas Richet, Ismail Ben Ayed, and Eric Granger. Sparsity outperforms low-rank projections in few-shot adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3143–3152, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Mrabah_Sparsity_Outperforms_Low-Rank_Projections_in_Few-Shot_Adaptation_ICCV_2025_paper.html.

Paul Pacaud, Ricardo Garcia, Shizhe Chen, and Cordelia Schmid. Guardian: Detecting robotic planning and execution errors with vision-language models. *arXiv preprint arXiv:2512.01946*, 2026. Version 4, revised August 2026; the repository retains the Guardian/FailCoT naming.

Yuqi Peng, Pengfei Wang, Jianzhuang Liu, and Shifeng Chen. GLAD: Generalizable tuning for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 4369–4379, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025W/MMFM/html/Peng_GLAD_Generalizable_Tuning_for_Vision-Language_Models_ICCVW_2025_paper.html.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, 2022.

Deepak Sridhar, Kartikeya Bhardwaj, Jeya Pradha Jeyaraj, Nuno Vasconcelos, Ankita Nayak, and Harris Teague. Video reasoning without training. *arXiv preprint arXiv:2510.17045*, 2026. Version 2, revised June 2026.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, 2020.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In *Advances in Neural Information Processing Systems*, 2024.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Fangcong Yin, Xi Ye, and Greg Durrett. LoFiT: Localized fine-tuning on LLM representations. In *Advances in Neural Information Processing Systems*, 2024.

Hao Yin, Guangzong Si, and Zilei Wang. ClearSight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14625–14634, 2025.

Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1593–1603, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024W/PV/html/Zanella_Low-Rank_Few-Shot_Adaptation_of_Vision-Language_Models_CVPRW_2024_paper.html.

Chengsheng Zhang, Chenghao Sun, Xinyan Jiang, Wei Li, and Xinmei Tian. Prefill-time intervention for mitigating hallucination in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25293–25303, 2026.

Enyu Zhao, Vedant Raval, Hejia Zhang, Jiageng Mao, Zeyu Shangguan, Stefanos Nikolaidis, Yue Wang, and Daniel Seita. ManipBench: Benchmarking vision-language models for low-level robot manipulation. In *Conference on Robot Learning*, volume 305, pp. 3413–3462, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A MATHEMATICAL DETAILS

A.1 NOTATION

For each projection, $B_j \in \mathbb{R}^{d_j \times r}$ is orthonormal and $R_j \in \mathbb{R}^{r \times r}$ is trainable. The actual projection width d_j need not equal the language hidden width, including in grouped-query attention. We suppress domain/module indices below and write the broadcast mask as $D_M = \mathrm{diag}(M)$.

The intervention is $\widetilde{Z} = Z + D_M Z B A B^\top$. Support-derived B is detached before fitting R; both the basis and controller belong to the saved adaptation state.

A.2 WHY THE SECOND MOMENT SELECTS THESE DIRECTIONS

Proposition 1 (Maximum retained support energy)**.** *Let $C = N^{-1} \sum_i G_i^\top G_i \succeq 0$ have descending eigenvalues $\lambda_1, \ldots, \lambda_d$. Among all $d \times r$ matrices with orthonormal columns,*

$$\max_{B^\top B = I_r} \mathrm{tr}(B^\top C B) = \sum_{k=1}^{r} \lambda_k, \qquad \mathrm{tr}(B^\top C B) = \frac{1}{N} \sum_i \|G_i B\|_F^2.$$

Proof. Write $C = U \Lambda U^\top$ and $W = U^\top B$. Since $W^\top W = I_r$, the row energies $q_k = \|W_{k,:}\|_2^2$ satisfy $0 \le q_k \le 1$ and $\sum_k q_k = r$. The trace equals $\sum_k \lambda_k q_k$, which is maximized by assigning unit mass to the r largest eigenvalues. Taking the corresponding eigenvectors as columns of B attains this value. Cyclicity of the trace yields the Frobenius-norm identity. □

This classical trace result optimizes support energy, not classification. Tied eigenvalues permit different bases; common positive scaling preserves eigenspaces. Token rows are accumulated without per-image token-count normalization, which would change image weighting.

Technical context. Correctness-derived K/V directions have also been used to control self-generated training data (Hao et al., 2026). Here direction estimation is followed by fitting a bounded, visual-row-selective residual that is reused at deployment; neither self-generated distillation data nor subsequent backbone fitting is part of the procedure. The eigenspace primitive is not claimed as a new contribution.

A.3 RESIDUAL PROPERTIES

Proposition 2 (Identity, support, and local amplitude)**.** *For $A = \alpha R/(1 + \|R\|_F)$ with finite R, orthonormal B, and $\alpha > 0$, the following hold: (i) $R = 0$ implies $\widetilde{Z} = Z$; (ii) $(I - D_M)(\widetilde{Z} - Z) = 0$; (iii) $(\widetilde{Z} - Z)(I - BB^\top) = 0$; and (iv)*

$$\|A\|_2 \le \|A\|_F < \alpha, \qquad \|\widetilde{Z} - Z\|_F \le \alpha \|D_M Z B\|_F.$$

Proof. The first claim follows from $A = 0$. The second follows from $(I - D_M) D_M = 0$. The third follows from $B^\top (I - BB^\top) = 0$. Finally, $\|A\|_F = \alpha \|R\|_F/(1 + \|R\|_F) < \alpha$, and submultiplicativity with $\|B^\top\|_2 = 1$ gives the residual bound. □

This is a local bound; global logit changes additionally depend on the downstream Jacobian. Alpha 3 permits amplification, and retaining the identity path does not imply that $I + BAB^\top$ is invertible.

Orientation within the same space. For orthogonal Q, replacing (B, A) with $(BQ, Q^\top A Q)$ preserves $BAB^\top$ and its Frobenius bound. Full control is thus coordinate-invariant within the selected space; Diagonal control generally is not, because basis rotation can make a diagonal action dense.

A.4 THE COEFFICIENT GRADIENT DIFFERS FROM AN ACTIVATION GRADIENT

Let $H = \partial L/\partial \widetilde{Z}$ and $U = D_M ZB$. Holding the incoming Z fixed, Equation 5 gives $\partial L/\partial A = (D_M ZB)^\top HB$. This follows from $d\widetilde{Z} = U(dA)B^\top$ and $dL = \mathrm{tr}(H^\top d\widetilde{Z})$. The fitted correction couples image coordinates ZB with answer sensitivity HB, so a common controller can produce different residuals on different images. For $n = \|R\|_F > 0$, write $G_A = \partial L/\partial A$. The chain rule gives

$$\nabla_R L = \frac{\alpha}{1+n} G_A - \frac{\alpha \langle G_A, R\rangle_F}{n(1+n)^2} R. \tag{7}$$

At zero, $A(R) = \alpha R + O(\|R\|_F^2)$, so $\nabla_R L|_{R=0} = \alpha G_A$. With multiple intervened layers, automatic differentiation additionally propagates downstream edits through earlier modules; Equation 5 is the corresponding local chain-rule term.

High activation-gradient overlap is insufficient. For $B = (1,0)^\top$, $z = (0,1)$, and $h = (1,0)$, all gradient energy is in the selected direction, yet $zB = 0$ makes the residual and coefficient gradient zero. Controller updates depend on both activations and sensitivities, not overlap alone.

A.5 A LOCAL DESCENT STATEMENT AND ITS SCOPE

For a smooth query loss $L_Q(\phi)$ with β-Lipschitz gradient, an idealized support-gradient step $\phi' = \phi - \eta g_S$ satisfies

$$L_Q(\phi') \le L_Q(\phi) - \eta \langle g_Q, g_S\rangle + \frac{\beta \eta^2}{2} \|g_S\|^2,$$

where $g_Q = \nabla_\phi L_Q$ and $g_S = \nabla_\phi L_S$ are gradients in the same controller coordinates. This descent-lemma interpretation is not an accuracy guarantee: activation-space cosine is not the controller-gradient inner product, and the executed optimizer additionally uses Adam, clipping, and regularization.

B EVALUATION PROTOCOLS, FITTING BUDGETS, AND SCORING

B.1 DATA AND EVALUATION BLOCKS

BRIGHT uses building-level classification on paired 448-pixel crops: support24, balanced query300, and separated support/query tiles. The matched reproduction regenerates deterministic folds from checksum-verified official assets because historical private manifests were unavailable. Extra-backbone results use their archived folds; neither block is the full dense damage-mapping estimator.

The five-arm Qwen2.5-VL suite uses H2 support16/query300 and ManipBench Q1 support32/query400, with query seed 1729 and three support seeds. H2 is supervised adaptation, not the WILDS domain-generalization estimator. ManipBench labels index each question's options. Guardian verification joins before/after observations in one image. New group-disjoint RoboFail has 113 queries; the archived split has 137.

Table 4: **Recorded evaluation blocks used by the research questions.** Fit/calibration counts are per task or event. A dash means that no separate calibration split is used in that block, not a missing measurement.

Block	Fit / cal.	Query	Supports	Main use
Matched BRIGHT	24 / —	300/event	3	RQ1
Archived fixed-support BRIGHT	24 / —	300/event	1 locked	RQ1 history
Archived BRIGHT robustness	24 / —	300/event	3 (KV only)	RQ1 history
Archived extra-backbone BRIGHT	24 / —	300/event	archived	RQ1
Four-backbone H2	12 / 4	200	2	RQ1–2
Four-backbone PathMNIST	54 / 18	180	2	RQ1–2
Final H2 calibration	12 / 4	300	3	RQ1–2
Final BRIGHT calibration	18 / 6	300/event	3	RQ1, RQ3
Final manipulation calibration	24 / 8	400/domain	3	RQ2 scope
Fixed-strength subsets	existing lens	60/domain	seed0	RQ4

B.2 Training budgets and model interfaces

An epoch visits the complete support once; Table 5 separates visits from optimizer updates. Both LoRA variants have rank16. Prompts and image/data budgets match within each comparison. Corrected historical InternVL3 uses 224 pixels/64 visual tokens; the Qwen3 task block uses 448 pixels/256 tokens.

Table 5: **What one and four LoRA epochs mean.** Rank is fixed at 16. A visit is one training-loss evaluation on one support example; direction extraction is an additional pass for Aperture.

Method	Rank	LR	Epochs	Updates	Example visits
LoRA-1E (BRIGHT)	16	10^{-4}	1	8	24; accumulation 3
LoRA-4E (BRIGHT)	16	10^{-4}	4	32	96; accumulation 3
Aperture (BRIGHT)	16	0.05	4	4	96 + 24 extraction
Random (BRIGHT)	16	0.05	4	4	96; random basis
LoRA-4E (single image)	16	2×10^{-4}	4	$4N$	$4N$; accumulation 1
Aperture (Camelyon)	16	0.05	4	4	64 + 16 extraction
Aperture (RoboFail)	16	0.01	8	8	128 + 16 extraction

K/V-output masks persist through backward and checkpoint replay. Captured outputs retain their upstream graph; fixed bases and modules are excluded from the optimizer. The backend evaluates candidate sequences, not a validated open-ended generation-cache interface.

B.3 Loss spans and probability metrics

Mean answer-token loss is reduced across support by a mean in BRIGHT and a sum in the recorded single-image recipe; switching reduction changes relative regularization. A shared answer prefix cancels in deterministic summed candidate normalization but can contribute adaptation gradients.

For P0 candidate a_c, $s_c(x) = \sum_{u \in \mathcal{T}_c} \log p(a_{c,u} \mid x, a_{c,<u})$ and $p_c = \mathrm{softmax}(s)_c$. The field named `mean_log_scores` in that backend contains these sums. Historical scoring/formatting contracts are retained with their own blocks. NLL is $-n^{-1} \sum_i \log p_{i,y_i}$ over normalized candidate probabilities; Brier is the mean sum of squared class-probability errors; ECE uses the recorded confidence bins. Mean/SD describes support-seed variation, not repeated independent query datasets. Query-cluster intervals condition on the fitted support and do not replace support-level replication.

C RQ1–RQ2: Complete classification and cross-backbone results

C.1 The complete historical five-arm suite

Table 6 reports three-seed F1/NLL for the earlier full-support Qwen2.5-VL states. Aperture has lower NLL, while LoRA has higher manipulation F1. The later calibration-held-out controls are separate (Appendix D).

The archived Camelyon Aperture F1/NLL are 0.6781 ± 0.0161 and 0.5895 ± 0.0156; LoRA gives 0.4054 ± 0.0831 and 5.7935 ± 3.4386. Mean Brier/ECE for Aperture is .403/.061 on Camelyon, .661/.056 on Bridge, .750/.113 on DROID-place, and .777/.116 on DROID-arti. Corresponding LoRA values are .917/.460, .430/.210, .528/.254, and .466/.230. These show why the NLL finding is not a blanket statement about all probability metrics. Complete five-arm Brier/ECE values are retained in the packaged JSON.

Table 6: **Archived five-arm results: macro-F1 / NLL.** Three-support-seed means. Exact per-seed values and SD remain in the supplied source summaries.

Method	Camelyon	Bridge	DROID-place	DROID-arti
Frozen	.554 / .666	.233 / 1.654	.155 / 1.890	.148 / 1.961
LoRA-4E	.405 / 5.793	.774 / 2.839	.714 / 2.342	.754 / 2.430
Random visual	.536 / .674	.226 / 1.567	.197 / 1.722	.157 / 1.836
Hidden residual	.470 / .690	.345 / 1.332	.250 / 1.573	.279 / 1.576
APERTURE	.678 / .589	.426 / 1.214	.288 / 1.394	.250 / 1.458

C.2 MATCHED BRIGHT, HISTORICAL PROTOCOL LINEAGE, AND ADDITIONAL-BACKBONE METRICS

Table 7: **Original matched BRIGHT macro-F1.** Mean and SD over three support seeds per event; all events and both fixed LoRA budgets are retained.

Method	Hawaii	Libya	Noto	Turkey
Frozen	$.194_{\pm.000}$	$.273_{\pm.000}$	$.206_{\pm.000}$	$.229_{\pm.000}$
LoRA-1E	$.350_{\pm.038}$	$.359_{\pm.029}$	$.313_{\pm.029}$	$.236_{\pm.009}$
LoRA-4E	$.284_{\pm.038}$	$.244_{\pm.038}$	$.268_{\pm.039}$	$.299_{\pm.030}$
Random visual	$.367_{\pm.028}$	$.305_{\pm.044}$	$.280_{\pm.078}$	$.255_{\pm.020}$
APERTURE	$.333_{\pm.044}$	$.340_{\pm.032}$	$.324_{\pm.027}$	$.256_{\pm.035}$

Archived fixed-support raw-backbone campaign. The historical reports called the method *Full Gradient-Cov KV* or *Full KV-TTT*. Its Full row below uses the same core operator as APERTURE: a correctness-gradient second-moment basis, visual K/V-only intervention, layers 14 and 27, rank 16, bounded Full mixing, four full-support updates, and 1,024 learned coefficients. The current paper renames and generalizes this operator. Table 8 is retained because it established the approximately 32% BRIGHT result before the current three-seed fold regeneration.

Table 8: **Archived fixed-support BRIGHT comparison from the raw backbone.** Macro-F1 is in percent; NLL is the four-event mean. Each event used one locked support24 and query300, so no support-seed SD exists for this table. "Learned" excludes fixed basis values.

Method	Hawaii	Libya	Noto	Turkey	Mean F1 ↑	NLL ↓	Learned
Raw VLM	19.92	25.85	20.87	22.73	22.34	1.3233	0
LoRA (fixed 8 updates)	**35.26**	27.76	**34.95**	21.47	29.86	**1.1276**	10,092,544
Diagonal precursor	29.18	25.60	28.20	28.79	27.94	1.1715	**64**
APERTURE (archived Full)	29.57	**37.39**	31.12	**30.25**	**32.08**	1.1367	1,024

Archived support-seed robustness. A separate historical suite repeated the Full and Diagonal controllers over three independently sampled, balanced, query-tile-disjoint support24 sets while fixing each event's query IDs. Table 9 reports mean±sample SD over those support seeds. This suite supports robustness relative to the fixed Frozen model; it does not provide a matched three-seed LoRA comparison. Its manifests and implementation snapshot differ from the current Table 1, so the two means are descriptive protocol replications rather than pooled estimates.

Table 9: **Archived BRIGHT support robustness.** Macro-F1 (%), mean ± sample SD over three support24 seeds per event. The final column is the equal-weight mean of the four event means.

Method	Hawaii	Libya	Noto	Turkey	4-event mean
Full controller	$27.95_{\pm3.17}$	$30.44_{\pm2.23}$	$26.11_{\pm5.66}$	$23.17_{\pm4.98}$	26.92
Diagonal controller	$28.83_{\pm3.44}$	$31.11_{\pm5.30}$	$19.82_{\pm1.16}$	$25.57_{\pm2.93}$	26.33

Earlier source-initialized precursor. Before the raw-backbone protocol was adopted, a support12 leave-one-event-out experiment began from a LoRA trained on the remaining disasters and used a bounded diagonal visual-state controller. Across the first five completed folds, the source model obtained $35.30 \pm 7.05\%$ macro-F1 and the visual-state precursor obtained $37.68 \pm 7.27\%$, with a gain of 2.38 ± 1.07 points and 5/5 F1 and NLL wins. These SDs are across events, not support seeds. Adding Haiti produced a six-fold mean of $33.67 \pm 7.47\%$ for the source model and $33.20 \pm 12.76\%$ for the precursor, leaving 5/6 event wins but exposing an event-specific failure. This earlier setting is methodological lineage, not a directly comparable estimate for the current raw-backbone controller.

In Frozen/LoRA/Random/Aperture order, the archived Qwen3-VL four-event balanced accuracies are .3333/.3317/.3342/.3558, with NLL 9.4245/1.3891/8.1916/2.1515. InternVL3 gives BA .3258/.3333/.3367/.3583 and NLL 1.4763/1.6927/1.3085/1.3684. These accompany Table 1; model-specific protocols and score scales are not pooled.

C.3 Complete fixed-configuration pathology matrix

Table 10 retains the full 112-entry matrix under the protocol in Section 4. Eight model–task audits, four support-repeat gates, and prediction-based metric checks precede export.

Table 10: **Complete fixed-configuration pathology comparison.** Completed fixed-configuration study; mean and sample SD over two support seeds. F1 is in percent. NLL+T uses held-out-support temperature scaling. Bold marks the best mean within each backbone, task, and metric.

		CAMELYON17–H2			PathMNIST–224		
Backbone	Method	F1 ↑	NLL ↓	NLL+T ↓	F1 ↑	NLL ↓	NLL+T ↓
Qwen2.5-VL 7B / Qwen	Frozen	$51.75_{\pm 0.00}$	$0.812_{\pm 0.000}$	$1.623_{\pm 0.573}$	$35.07_{\pm 0.00}$	$2.277_{\pm 0.000}$	$1.876_{\pm 0.005}$
	Output bias	$51.42_{\pm 0.10}$	$0.797_{\pm 0.001}$	*n/a*	$34.82_{\pm 2.14}$	$1.640_{\pm 0.008}$	*n/a*
	LoRA-1E	$33.33_{\pm 0.00}$	$0.846_{\pm 0.049}$	$0.752_{\pm 0.042}$	$39.51_{\pm 4.22}$	$\mathbf{1.337}_{\pm 0.041}$	$1.341_{\pm 0.041}$
	LoRA-4E	$\mathbf{61.32}_{\pm 18.68}$	$3.062_{\pm 3.002}$	$\mathbf{0.632}_{\pm 0.175}$	$\mathbf{64.35}_{\pm 2.20}$	$1.590_{\pm 0.405}$	$\mathbf{0.885}_{\pm 0.006}$
	Aperture	$55.34_{\pm 0.92}$	$\mathbf{0.774}_{\pm 0.009}$	$1.015_{\pm 0.028}$	$39.87_{\pm 1.80}$	$1.504_{\pm 0.011}$	$1.500_{\pm 0.017}$
Qwen3-VL 4B / Qwen	Frozen	$66.52_{\pm 0.00}$	$1.952_{\pm 0.000}$	$4.390_{\pm 3.300}$	$44.62_{\pm 0.00}$	$4.792_{\pm 0.000}$	$1.459_{\pm 0.049}$
	Output bias	$59.79_{\pm 6.63}$	$1.860_{\pm 0.106}$	*n/a*	$58.85_{\pm 3.13}$	$2.363_{\pm 0.134}$	*n/a*
	LoRA-1E	$52.93_{\pm 17.33}$	$1.494_{\pm 0.695}$	$\mathbf{1.915}_{\pm 1.667}$	$64.63_{\pm 5.87}$	$\mathbf{1.077}_{\pm 0.387}$	$0.937_{\pm 0.219}$
	LoRA-4E	$\mathbf{78.73}_{\pm 7.91}$	$2.685_{\pm 2.580}$	$2.665_{\pm 2.608}$	$\mathbf{70.56}_{\pm 0.31}$	$1.208_{\pm 0.060}$	$\mathbf{0.875}_{\pm 0.019}$
	Aperture	$57.13_{\pm 5.38}$	$\mathbf{1.467}_{\pm 0.341}$	$4.014_{\pm 0.284}$	$59.07_{\pm 4.00}$	$1.866_{\pm 0.158}$	$1.107_{\pm 0.020}$
Qwen3-VL 8B / Qwen	Frozen	$59.19_{\pm 0.00}$	$1.669_{\pm 0.000}$	$4.405_{\pm 5.040}$	$38.46_{\pm 0.00}$	$4.646_{\pm 0.000}$	$1.732_{\pm 0.036}$
	Output bias	$61.49_{\pm 1.41}$	$0.959_{\pm 0.003}$	*n/a*	$60.26_{\pm 1.22}$	$1.421_{\pm 0.016}$	*n/a*
	LoRA-1E	$58.12_{\pm 13.16}$	$1.022_{\pm 0.038}$	$\mathbf{0.847}_{\pm 0.064}$	$50.09_{\pm 0.79}$	$1.260_{\pm 0.042}$	$1.138_{\pm 0.018}$
	LoRA-4E	$\mathbf{75.50}_{\pm 7.45}$	$1.489_{\pm 0.869}$	$1.778_{\pm 1.860}$	$\mathbf{81.40}_{\pm 4.96}$	$\mathbf{0.632}_{\pm 0.104}$	$\mathbf{0.639}_{\pm 0.049}$
	Aperture	$57.61_{\pm 10.45}$	$\mathbf{0.865}_{\pm 0.266}$	$2.052_{\pm 0.290}$	$61.01_{\pm 2.34}$	$1.651_{\pm 0.001}$	$1.245_{\pm 0.008}$
Gemma 3 4B / Gemma	Frozen	$55.97_{\pm 0.00}$	$2.060_{\pm 0.000}$	$1.865_{\pm 0.995}$	$17.09_{\pm 0.00}$	$13.291_{\pm 0.000}$	$2.114_{\pm 0.014}$
	Output bias	$\mathbf{73.19}_{\pm 0.36}$	$\mathbf{0.735}_{\pm 0.006}$	*n/a*	$25.72_{\pm 0.20}$	$4.469_{\pm 0.044}$	*n/a*
	LoRA-1E	$33.33_{\pm 0.00}$	$1.667_{\pm 0.274}$	$\mathbf{0.715}_{\pm 0.008}$	$24.32_{\pm 1.67}$	$\mathbf{1.984}_{\pm 0.062}$	$1.979_{\pm 0.059}$
	LoRA-4E	$56.05_{\pm 26.79}$	$3.542_{\pm 0.458}$	$2.244_{\pm 2.219}$	$\mathbf{34.43}_{\pm 0.21}$	$2.959_{\pm 0.342}$	$\mathbf{1.750}_{\pm 0.019}$
	Aperture	$44.16_{\pm 11.44}$	$2.267_{\pm 0.652}$	$2.158_{\pm 1.837}$	$23.19_{\pm 4.56}$	$3.920_{\pm 1.032}$	$2.028_{\pm 0.037}$

Output bias uses all support labels; a second temperature is not fitted (n/a). NLL is rounded to three decimals for readability; full-precision means/SD are retained in the source export. 1E/4E denote one/four complete support epochs; all measured arms are shown.

D RQ2: Complete final-round calibration comparisons

This 96-state study uses three supports on eight Qwen2.5-VL domains: BRIGHT fit/calibration18/6, H2 12/4, and ManipBench24/8. Temperature and schedule selection precede query evaluation. NLL uses log-sum-exp rather than clipping. These states differ from both the full-support suite and Table 10.

D.1 Stronger LoRA controls and dataset-specific gains

Table 11: **Completed calibration-round NLL.** Means over three support seeds. Each cell is raw NLL / temperature-calibrated NLL. Bias has no second temperature. H2 uses 300 queries; each manipulation domain uses 400. Complete per-seed SD is in the exported result tables.

Method	H2	Bridge	DROID-place	DROID-arti
Frozen	.667 / .804	1.674 / 1.342	1.916 / 1.394	1.983 / 1.451
Output bias	.653 / n/a	1.236 / n/a	1.485 / n/a	1.574 / n/a
LoRA-1E	.737 / .720	.858 / 1.021	1.110 / .953	1.146 / 1.075
LoRA-4E	1.326 / 2.746	2.045 / .924	2.557 / .895	2.862 / .821
Aperture	.608 / .692	1.138 / 1.143	1.349 / 1.345	1.506 / 1.577
LoRA (support-selected)	1.326 / 2.746	1.377 / 1.035	1.110 / .895	1.146 / 1.008

The H2 lens has the lowest mean NLL, including against calibrated alternatives. On all three manipulation tasks, LoRA-1E has lower NLL and temperature scaling strengthens LoRA-4E. Raw and calibrated LoRA (support-selected) results can choose different schedules; any resulting F1 difference is due to schedule selection, not temperature changing a fixed model's argmax.

Table 12: **Classification in the same final-round states.** Macro-F1 mean and sample SD over three supports. This three-seed, 300/400-query protocol is distinct from the two-seed cross-family matrix.

Method	H2	Bridge	DROID-place	DROID-arti
Frozen	$.5421_{\pm.0000}$	$.2300_{\pm.0000}$	$.1552_{\pm.0000}$	$.1344_{\pm.0000}$
Output bias	$.5942_{\pm.0275}$	$.4483_{\pm.0064}$	$.3217_{\pm.0101}$	$.2736_{\pm.0054}$
LoRA-1E	$.4141_{\pm.1111}$	$.7697_{\pm.0381}$	$.6549_{\pm.0267}$	$.6504_{\pm.0593}$
LoRA-4E	$.6429_{\pm.0737}$	$.7754_{\pm.0319}$	$.6791_{\pm.0146}$	$.7457_{\pm.0329}$
Aperture	$.6166_{\pm.0929}$	$.5015_{\pm.0437}$	$.3031_{\pm.0770}$	$.2231_{\pm.0555}$

E RQ3: Resource accounting and reproducibility

E.1 State size and source records

For m sites, Full control learns mr^2 coefficients and stores $r\sum_j d_j$ basis values; Diagonal learns mr. At two K/V layers and rank16, these counts are 1,024 and 64; one-layer Full uses 512. Extraction requires support backpropagation and module-wise eigendecomposition. Query control adds $O(Td_jr + Tr^2)$ per module without new tokens. Parameter count, serialized state size, and wall-clock cost are different quantities.

The source package retains the 83-job P0 campaign, 29 follow-ups, archived multisupport suites, 96-state calibration round, completed 64-state pathology matrix, and both exploratory support searches. Their data splits and result versions are not pooled. This manuscript update adds no GPU runs.

E.2 Resources and repeated fits

The Hawaii RTX 4090 timing records report 7.92 seconds for LoRA-1E fitting, 25.06 seconds for LoRA-4E, and 22.14 seconds for Aperture including 4.88 seconds of extraction. Thus compact storage does not imply universally faster fitting. The 135.43-KiB controller includes 32,768 fixed-basis values and 1,024 learned coefficients. The approximately 291-fold serialized-file reduction is distinct from the 9,856-fold reduction in optimized variables.

Five archived reruns keep query IDs and labels identical. Frozen scores match exactly, but trainable arms show run-to-run variation. Hawaii seed 1 Aperture changes 187/300 predictions, F1 .3134 to .3342, and NLL 1.4207 to 1.1254; Noto seed 2 changes 116/300 predictions and F1 .3449 to .3709. These results preclude a bitwise-stability claim. A separate DROID-arti rerun moves only descriptive operator-norm calculation to CPU after a cuSOLVER allocation failure and remains labeled as a diagnostic source change. None of these reruns replaces the sealed primary values.

F RQ4: SENSITIVITY, CORRECTION DIRECTION, AND INTERVENTION SITES

Table 13: **Basis quality and intervention location are distinct.** Macro-F1 on archived InternVL3 seed-0 splits. The query-basis column uses query labels only to estimate directions; coefficients are fitted on support in every column. Other columns use support-derived directions.

Task	KV	Query-basis KV	Q only	O only	QKVO
Camelyon	0.5994	0.6632	0.5125	0.4172	0.6664
RoboFail	0.4389	0.4119	0.4811	0.4506	0.4538
DROID-place	0.2193	0.2482	0.1548	0.1631	0.2625

F.1 ENERGY, SIGN, AND LOCATION

Energy ratios aggregate projected second moments; signed agreement aggregates projected mean gradients. K/V-only overlaps below differ from all-QKVO values in Table 14. Recipes are Camelyon4 updates/rate0.05, RoboFail8/0.01, and ManipBench8/0.05. Query labels are used only for retrospective diagnostics.

For K/V, Q-only, O-only, and QKVO respectively, Camelyon NLL is .6629/.7057/.7590/.6333 and energy coverage .9362/.5942/.4884/.7786. RoboFail has NLL .7219/.8908/.9488/.7446 and coverage .7441/.5039/.2876/.6077. DROID-place has NLL 1.4898/1.8415/1.8497/1.4610 and coverage .7242/.4873/.3306/.6122. These complement the F1 comparisons in Table 13.

Table 14: Unsigned energy coverage and signed agreement by projection. All denotes the combined Q/K/V/O diagnostic modules, not K/V alone.

Task	Statistic	Q	K	V	O	All
Camelyon	ρ	0.600	0.964	0.751	0.497	0.784
	κ	+0.982	+0.960	+0.991	+0.994	+0.974
RoboFail (old)	ρ	0.507	0.863	0.394	0.290	0.610
	κ	+0.034	-0.101	-0.298	-0.361	-0.207
DROID-place	ρ	0.493	0.820	0.480	0.338	0.617
	κ	+0.731	+0.441	+0.756	+0.815	+0.609

Table 13 reports query-basis and site F1, with complementary energy and NLL values above. Broader QKVO control increases both location coverage and coefficient count; these probes are not upper bounds over all activation controllers.

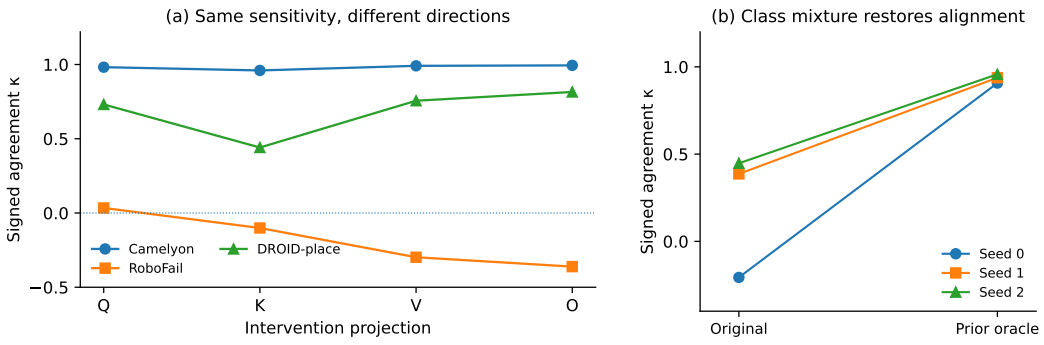

Figure 3: **Correction directions depend on task and class mixture.** Left: projection-wise support–query agreement on the archived seed-0 diagnostics. Right: query-prior reweighting restores directional agreement across three original RoboFail supports. The oracle changes a diagnostic gradient mixture, not a deployed classifier.

F.2 Class mixtures and replicated directional transfer

Camelyon's all-module three-support-seed statistics are $\rho = 0.7846 \pm 0.0049$ and $\kappa = 0.9773 \pm 0.0079$. Its normal/tumor class-conditional mean agreements are 0.9876/0.9802. RoboFail has $\rho = 0.620 \pm 0.008$ and $\kappa = 0.209 \pm 0.361$, but class-conditional means 0.962/0.970. DROID-place retains positive aggregate agreement across all three supports (0.741 ± 0.142), although it does not close the LoRA classification gap.

The original RoboFail support has eight examples per class while query has 116 success/21 failure examples. Keeping the basis fixed and substituting oracle query proportions in the support class-gradient mixture changes seed-0/1/2 agreement from -.2065/.3866/.4475 to .9070/.9375/.9574 (Figure 3). Module-wise post-reweighting means are Q=0.9521, K=0.9433, V=0.9143, and O=0.9267. The recovery spans modules, supporting the mixture explanation.

Archived support12 BRIGHT diagnostics give Hawaii $\rho = 0.4328$, $\kappa = 0.9754$ and Libya $\rho = 0.3689$, $\kappa = 0.6769$, with strong within-class alignment. They provide qualitative evidence on a distinct support protocol; they are not paired with support24 F1 in a predictive correlation. Likewise, the original 137-query RoboFail directional analysis is separate from the later group-disjoint 113-query evaluation.

G RQ4: Visual controls, placement, and transfer scope

G.1 Direction source and controller structure

Table 15: **Direction choice changes the fitted visual correction.** RQ4: archived four-event BRIGHT ablation; NLL reduction is relative to Frozen. The three rank-5 arms provide the rank-matched comparison; mean gradient is a rank-1 diagnostic. Random varies basis seeds, not support sets.

Basis	Rank	F1	ΔBA	NLL reduction
Uncentered gradients	5	.2964	-.0008	.1435
Mean gradient	1	.2506	-.0100	.0984
Centered gradients	5	.2721	-.0133	.1440
Random directions	5	.2414	-.0028	.0468

The direction-source ablation appears in Table 15. On the newer Hawaii supports, the learned basis reduces support loss more than Random but gives lower query F1. Appendix A explains the distinct invariances of Full and Diagonal control.

G.2 Position and residual budget: two task-specific probes

Table 16: **Fixed placement rule across every BRIGHT event.** Three-support-seed mean F1. The one-layer rule uses 512 rather than 1,024 coefficients; both positive and negative changes are retained.

Event	Layers 14+27	Layer 14	Change (pp)
Hawaii	.3329	.2883	−4.46
Libya	.3404	.3354	−0.50
Noto	.3242	.3378	+1.36
Turkey	.2557	.3212	+6.55

The source JSON retains all 29 follow-ups; Table 16 applies the same layer-14 rule across events. On Hawaii, the original alpha-3 mean is .3329 (seed 0: .3630). Alpha 1 yields seed values .4019/.2362/.3352 and mean .3244: a favorable seed-0 change does not persist in the mean. On Turkey, layers 14+27 give .2845/.2659/.2166 across seeds; layer 14 alone gives .3574/.2867/.3195. Four-pass LoRA gives .3272/.3028/.2676. The means are .2557/.3212/.2992, respectively.

In the same seed-0 Turkey probe, eight updates give F1 0.3378; two updates 0.2553; rate 0.01 gives 0.2517; rank 8 gives 0.2445. These one-factor probes do not replace the fixed-rule means in Table 16.

G.3 OBJECTIVE CONTROLS

On Camelyon seed 0, mean/full and sum/full answer losses give F1/NLL 0.5458/0.6861 and 0.5457/0.7252. Their label-only counterparts both have F1 0.3333 (NLL 1.0251 and 1.1276). The tested full-answer objective therefore supports useful adaptation in this setting. On the new RoboFail split, the same four variants remain at F1 0.3005–0.3080. Answer-span and support reduction are recorded design choices, not interchangeable formatting details.

G.4 OUTPUT BIAS AND FIXED-STATE VISUAL DEPENDENCE

Output Bias is retained here as a mechanism control rather than a primary adaptation baseline: it changes only support-fitted candidate-score offsets and does not modify internal VLM states. Output

Table 17: **Internal visual control and output calibration.** Real-input seed-0 F1 / NLL, approximately reconstructed from reported Frozen scores and exact gains. The image-control experiment uses four-epoch LoRA. Camelyon uses InternVL3; the BRIGHT events use Qwen2.5-VL.

Domain	Frozen	Output bias	LoRA	APERTURE
Hawaii	0.194 / 1.245	0.339 / 1.109	0.306 / 1.127	0.363 / 1.096
Libya	0.273 / 1.323	0.408 / 1.105	0.228 / 1.171	0.369 / 1.109
Camelyon	0.333 / 0.966	0.629 / 0.642	0.545 / 3.099	0.546 / 0.725
Learned scalars	0	$K-1$	LoRA-r16	1,024

bias fits $K-1$ zero-sum offsets by support candidate cross-entropy with ℓ_2 regularization. It retains frozen image-dependent scores. Table 17 is approximately reconstructed from Frozen scores and exact gains on separate seed-0 states, not the five-arm suite.

Table 18: Fixed-state visual dependence: gain contrast D_c from Eq. 6. Higher positive values mean a larger adaptation gain on real images than on that control.

Domain	Method	Shuffle	Neutral
Hawaii	Output bias	+0.0387	+0.1448
Hawaii	LoRA	+0.0408	+0.1119
Hawaii	APERTURE	+0.0745	+0.1686
Libya	Output bias	+0.0138	+0.1352
Libya	LoRA	-0.0449	-0.0451
Libya	APERTURE	+0.0026	+0.0965
Camelyon	Output bias	+0.1269	+0.2960
Camelyon	LoRA	+0.0941	+0.2120
Camelyon	APERTURE	+0.0534	+0.2123

Post-only shuffle/neutral contrasts are 0.0684/0.1610 on Hawaii and -0.0164/0.1369 on Libya. Models, text, and token budgets stay fixed; neutral images retain their size and shuffled donors are deterministic and nonidentity. D_c is a gain contrast, not a raw performance drop. Competitive bias and weaker shuffle effects preclude a universal denoising explanation; source records retain cluster intervals.

G.5 CLASSIFICATION BOUNDARIES ACROSS RECORDED SUITES

Corrected three-seed InternVL3 Aperture/LoRA F1 is 0.5107/0.6325 on Camelyon, 0.2993/0.5551 on Bridge, 0.3307/0.6759 on DROID-arti, and 0.2839/0.4304 on DROID-place. These LoRA advantages concern different inputs and splits from the cross-family matrix and different states from the five-arm NLL suite.

On the archived execution-verification folds, Frozen/LoRA/Aperture with eight updates gives mean F1 0.4509/0.4432/0.4905 on UR5-Fail, 0.4479/0.5010/0.3952 on RoboFail, and 0.4043/0.5159/0.5204 on RoboVQA. The corresponding four-update Aperture values are 0.4376/0.4229/0.4923. This comparison uses the original folds and both recorded fitting schedules.

The newer group-disjoint RoboFail split has 93 success and 20 failure queries. Frozen F1 is 0.4115 and the canonical two-layer lens 0.3080. Support covers only five of ten query task families. On covered/uncovered families, Frozen gives 0.4653/0.3611; a layer-27 probe gives 0.4844/0.3007, and the canonical controller reaches 0.1711 on uncovered families. These observations localize negative transfer to support coverage and the induced decision mixture. The old query-prior analysis does not use this new split.

G.6 Fixed residual strength and input perturbations

Table 19: **Turning on a fitted lens reduces NLL on every measured subset.** Fixed seed-0 subsets of 60 queries; no refitting. All eight domains are shown. Lower is better.

γ	Hawaii	Libya	Noto	Turkey	H2	Bridge	D.-place	D.-arti
0	1.2975	1.1976	1.2373	1.3443	.6570	1.5979	2.0274	2.1999
0.5	1.0879	1.1411	1.1029	1.1766	.6205	1.3066	1.5036	1.8376
1	1.0828	1.1604	1.0705	1.1178	.6247	1.2377	1.3077	1.5999

Table 19 scales fixed controllers on 60-query seed-0 subsets without refitting. H2 F1 is .5694/.6606/.4841 at strengths0/.5/1, despite lower NLL at both nonzero strengths. Gaussian-noise and JPEG probes likewise preserve model, text, and token budgets; their domain-dependent, sometimes nonmonotonic effects do not isolate semantic denoising. Complete dose and corruption CSVs retain these trade-offs.