PROVENANCE IS NOT UNIVERSAL: DISCOVERING AND DISTILLING FORENSIC SUBSPACES IN FROZEN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated image detectors often achieve strong in-distribution performance but degrade on unseen generators and datasets, raising the question of whether pretrained vision-language models (VLMs) encode more transferable provenance signals. Although frozen features from foundation models have shown promising performance for synthetic image detection, it remains unclear how such provenance information is organized within VLM representations. We address this gap by studying prompt-conditioned hidden states from seven frozen VLMs. We introduce Behavioral Provenance Signatures (BPS) to characterize separability, generator-conditioned forensic subspaces, cross-domain transferability, and distillability. We observe that frozen representations consistently support strong in-distribution separation but exhibit substantial variation across models and source generators. Generator-conditioned probes have partially aligned yet distinct discriminant directions, showing that strong separability does not imply a universal provenance axis. Distillation is similarly heterogeneous across teachers, target domains, and student architectures. Nevertheless, the resulting image-only models eliminate VLM execution and prompt processing at test time, achieving more than a $700\times$ measured inference speed-up. These results distinguish the availability of provenance information from its robustness and deployability.

1 INTRODUCTION

Recent image-generation models can produce increasingly realistic content at low cost and large scale. Reliable AI-generated image (AIGI) detection is therefore important for digital forensics, content moderation, and media verification. Many AIGI detectors exploit spatial, frequency, noise, or reconstruction artifacts associated with a particular generator (Wang et al., 2020; 2023a). These methods can perform well when the training and test images come from similar sources. However, their performance often degrades substantially when the images come from different generators, datasets, or processing pipelines. Large-scale evaluations, including GenImage, Chameleon, and AIGIBench, highlight a gap where high in-distribution performance does not reliably translate to out-of-distribution (OOD) detection (Zhu et al., 2023; Yan et al., 2025; Li et al., 2026).

Pretrained foundation vision-language models (VLMs) provide generalization features for detecting unseen generators. Simple classifiers trained on frozen CLIP features have shown stronger generalization to unseen generators than many task-specific detectors (Ojha et al., 2023; Cozzolino et al., 2024). Subsequent work has explored intermediate encoder blocks, patch tokens, and newer visual foundation models for AIGI detection (Koutlis & Papadopoulos, 2024; Abdullah et al., 2026; Zhou et al., 2026). These findings suggest that pretrained representations already expose information about image provenance.

However, detection performance alone provides limited insight into *how* provenance information is organized or *why* it transfers. Three questions remain open: **(i)** whether provenance-relevant information is consistently represented across frozen VLMs; **(ii)** whether detectors trained on different source generators rely on shared or source-specific directions; and **(iii)** whether the decision-relevant signal is sufficiently compact and learnable to transfer into lightweight image models. These dis-

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.

tinctions are important because separability, cross-domain transferability, and student distillability need not coincide.

Shortcut learning further complicates this interpretation, where detectors may exploit differences in image content, resolution, or compression that correlate with the real/synthetic label (Grommelt et al., 2024; Rajan et al., 2025). Source-attribution studies show that generators leave distinct fingerprints (Wang et al., 2023b), but attribution performance does not explain how these differences appear within frozen VLM representations or affect binary real/synthetic detection. Likewise, knowledge distillation has been used to build smaller AIGI detectors (Lim et al., 2024), but its benefits under distribution shift remain insufficiently understood.

In this work, we study these questions through *Behavioral Provenance Signatures* (BPS), prompt-conditioned representations derived from the hidden states of frozen VLMs. Using a fixed bank of visual-description prompts, we analyze both raw representations and PCA-compressed representations. We examine three properties of provenance representations: **(i) Separability** between real and synthetic images, **(ii) Cross-domain transferability** across generators and datasets, and **(iii) Distillability** into lightweight image-only detectors.

We first evaluate seven frozen VLMs to determine whether provenance-relevant separability, transfer and compressibility are consistent across model families. We then construct a generator-conditioned analysis using five source generators: SD1.5, ADM, GLIDE, Wukong, and VQDM to compare their OOD transfer, low-dimensional structure, and discriminant directions. Finally, we distill these VLM-derived provenance signals into lightweight MobileNetV3 and ResNet-18 students and evaluate how well their representations and detection behavior are preserved without VLM execution or prompt processing at inference.

Our results reveal a distinction between separability, transferability, and distillability. Frozen VLM representations provide strong in-distribution separation. However, their OOD transfer and compressibility vary across models and source generators. Generator-conditioned probes exhibit partially aligned but distinct discriminant directions, suggesting shared provenance structure alongside source dependence. Distillation is similarly heterogeneous. Its OOD effects vary across teachers, target domains, and student architectures, while representation fidelity alone is only modestly associated with OOD retention. Together, these findings show that the presence of a provenance signal does not by itself guarantee cross-domain robustness or efficient transfer to a smaller model.

Our main contributions are:

- We systematically study prompt-conditioned hidden states from seven frozen VLMs, evaluating the separability, compressibility, and OOD transfer of their prompt-conditioned provenance representations.
- We characterize forensic geometry across five training generators by jointly analyzing external transfer, PCA compressibility, discriminant-direction alignment, and OOD transfer profiles, showing that strong source separability coexists with source dependence.
- We investigate the distillability of VLM-derived provenance representations into lightweight image-only detectors across teachers, target domains, and student architectures, and measure the effects of representation matching, decision weighting, cosine alignment, logit supervision, and label supervision.

2 Problem Statement

Let $\mathcal{X}$ denote the image space and let $\mathcal{D}_g = \{(x_i, y_i)\}_{i=1}^{N_g} \subset \mathcal{X} \times \{0, 1\}$ be a dataset associated with source generator g. Here, $x_i \in \mathcal{X}$ is an image and we set $y_i = 0$ if x_i is a real image and $y_i = 1$ if x_i is a synthetic image.

Given a frozen VLM, we aim to determine whether its internal responses contain a provenance signal. We then study whether this signal transfers to unseen image sources and whether it can be learned by a lightweight student model. We refer to the resulting prompt-conditioned representation as a *BPS*. We address three research questions:

- **RQ1:** Can the hidden states of a frozen VLM separate real and synthetic images?
- **RQ2:** Does the learned provenance signal transfer to unseen generators and datasets?

- **RQ3:** Can this signal be transferred to a small image model without test-time VLM execution?

The VLM remains frozen in all experiments. The models, generators, datasets, and training settings are described in §4.

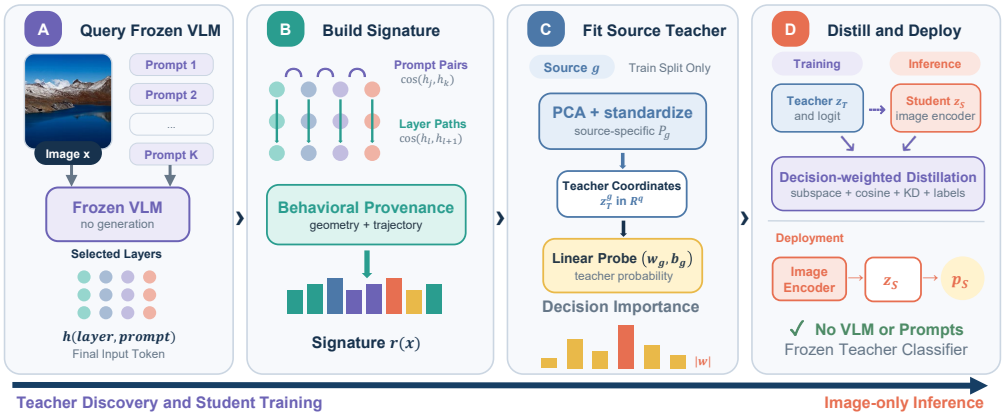


Figure 1: Overview of BPS. **A:** A frozen VLM processes an image with multiple prompts and extracts hidden states from selected layers. **B:** These activations and their within-layer prompt relationships define BPS representations. **C:** Source-specific PCA, standardization and a linear probe produce the generator-conditioned teacher representation and prediction. **D:** The signal is distilled into a lightweight image encoder. At inference, detection requires neither the VLM nor the prompts.

3 METHOD

3.1 OVERVIEW

Figure 1 summarizes the two stages of BPS. First, we use the source-specific PCA, standardization, and a linear provenance probe to construct a generator-conditioned forensic teacher from frozen VLM activations. Throughout this paper, *teacher* $\mathcal{T}_g$ refers to the complete generator-conditioned forensic model, composed of a frozen VLM feature extractor, a source-specific PCA and standardization transform, and a linear provenance probe $(\mathbf{w}_g, b_g)$. Second, we train a lightweight student to reproduce the teacher's provenance representation. Deployment requires only the student and the fixed linear head. The VLM is used only during teacher construction and student training. After distillation, the student performs provenance detection directly on the input image.

3.2 BEHAVIORAL PROVENANCE SIGNATURE

Let $\mathcal{P} = \{p_k\}_{k=1}^{K}$ be a fixed set of prompts and let $\mathcal{L} = \{\ell_l\}_{l=1}^{L}$ be the selected layers.

For image $x \in \mathcal{X}$, prompt p_k, layer $\ell \in \mathcal{L}$, we extract the hidden state of the last valid input token:

$$\mathbf{h}_{\ell,k}(x) = F_\ell(x, p_k) \in \mathbb{R}^d, \tag{1}$$

where F_ℓ denotes the frozen VLM up to layer ℓ. The model is used in evaluation mode without training or text generation. Collecting these states across the selected layers and prompts gives

$$\mathbf{H}(x) = [\mathbf{h}_{\ell,k}(x)]_{\ell\in\mathcal{L},\, 1\le k\le K} \in \mathbb{R}^{L\times K\times d}. \tag{2}$$

We describe within-layer prompt relationships using stabilized cosine similarities:

$$G_{ij}^{(\ell)}(x) = \frac{\mathbf{h}_{\ell,i}(x)^\top \mathbf{h}_{\ell,j}(x)}{\|\mathbf{h}_{\ell,i}(x)\|_2 \|\mathbf{h}_{\ell,j}(x)\|_2 + \epsilon}, \qquad 1 \le i < j \le K, \tag{3}$$

where constant $\epsilon > 0$ stabilizes the denominator.

The strictly upper-triangular entries from all selected layers form the prompt-geometry signature:

$$\mathbf{s}_{\text{geom}}(x) = \underset{\substack{\ell\in\mathcal{L}\\ 1\le i<j\le K}}{\text{Concat}} G_{ij}^{(\ell)}(x) \in \mathbb{R}^{L\binom{K}{2}}, \tag{4}$$

where the Concat operator concatenates the entries in a fixed order.

For teacher training, we use the concatenated prompt states from the final selected layer:

$$\mathbf{r}(x) = \mathrm{Concat}_{k=1}^{K} \mathbf{h}_{\ell_L,k}(x) \in \mathbb{R}^{Kd}. \tag{5}$$

The final-layer concatenation in equation 5 is the primary representation used to construct the generator-conditioned teachers. The geometry representation above, together with additional BPS variants, is evaluated in the representation analysis. Detailed signature definitions are given in § A.

3.3 Teacher Provenance Probe

For each source generator g, the teacher $\mathcal{T}_g$ combines the frozen feature map $\mathbf{r}(\cdot)$ with source-specific PCA, standardization, and a linear provenance probe. We fit the PCA transform, standardization statistics, and probe using only the training split of $\mathcal{D}_g$.

Let $\mathbf{P}_g \in \mathbb{R}^{q\times Kd}$ contain the retained PCA directions as rows and let $\boldsymbol{\mu}_g^r \in \mathbb{R}^{Kd}$ denote the training mean of the raw activations. The PCA coordinates are

$$\mathbf{u}_g(x) = \mathbf{P}_g \left(\mathbf{r}(x) - \boldsymbol{\mu}_g^r\right) \in \mathbb{R}^q. \tag{6}$$

Let $\boldsymbol{\mu}_g^u \in \mathbb{R}^q$ denote the training mean of the PCA coordinates, and let

$$\mathbf{S}_g = \mathrm{diag}(\sigma_{g,1}, \ldots, \sigma_{g,q}) \in \mathbb{R}^{q\times q}, \qquad \sigma_{g,j} > 0, \tag{7}$$

be the diagonal matrix of standardization scales. $\sigma_{g,j}$ denotes the training-set standard deviation of the jth PCA coordinate for source generator g. The standardized teacher representation is

$$\mathbf{z}_T^g(x) = \mathbf{S}_g^{-1} \left(\mathbf{u}_g(x) - \boldsymbol{\mu}_g^u\right) \in \mathbb{R}^q. \tag{8}$$

A logistic-regression probe with weights $\mathbf{w}_g \in \mathbb{R}^q$ and bias $b_g \in \mathbb{R}$ produces the teacher logit:

$$a_T^g(x) = \mathbf{w}_g^\top \mathbf{z}_T^g(x) + b_g. \tag{9}$$

The predicted probability that x is synthetic is

$$p_T^g(x) = \mathrm{sigmoid}\left(a_T^g(x)\right), \tag{10}$$

where $\mathrm{sigmoid}(t) = 1/(1+\exp(-t))$ is the standard sigmoid function. The VLM, PCA transform, standardization parameters, and linear probe remain fixed during student training and evaluation.

3.4 Student Distillation

For each teacher $\mathcal{T}_g$, we train a student with parameters θ to map an image directly to the teacher's standardized PCA space

$$\mathcal{S}_\theta : \mathcal{X} \to \mathbb{R}^q, \qquad \mathbf{z}_\mathcal{S}(x) = \mathcal{S}_\theta(x). \tag{11}$$

The student reuses the same frozen teacher probe

$$a_\mathcal{S}^g(x) = \mathbf{w}_g^\top \mathbf{z}_\mathcal{S}(x) + b_g. \tag{12}$$

For a fixed teacher, the complete student-training objective is

$$\mathcal{J} = \lambda_{\mathrm{sub}}\mathcal{L}_{\mathrm{sub}} + \lambda_{\cos}\mathcal{L}_{\cos} + \lambda_{\mathrm{KD}}\mathcal{L}_{\mathrm{KD}} + \lambda_y\mathcal{L}_y. \tag{13}$$

The four terms supervise the teacher coordinates, representation direction, teacher logit, and ground-truth label, respectively. In decision-weighted distillation, coordinate j receives a weight

$$\alpha_{g,j} = |w_{g,j}|, \qquad j = 1, \ldots, q. \tag{14}$$

The magnitude $|w_{g,j}|$ measures the teacher's logit sensitivity to a change in the standardized coordinate j. Therefore, larger weights emphasize coordinates where a given prediction error produces a larger change in the logit. Full loss definitions are given in § A.

3.5 Inference Time and Analysis

At inference time, the student predicts

$$p_{\mathcal{S}}^{g}(x) = \text{sigmoid}\left(\mathbf{w}_g^{\top}\mathbf{z}_S\left(x\right) + b_g\right). \tag{15}$$

It requires only the image encoder and frozen probe, with no VLM execution or prompt processing.

We measure teacher transfer using mean AUROC across external domains and assess student retention by comparing student and teacher AUROC on the same domains. Let $\mathcal{E}$ denote the set of external domains and $\mathcal{A}$ the set of student architectures. Teacher transferability is

$$T_g = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\text{AUROC}_T(g,e), \tag{16}$$

and student retention loss is

$$L_g = \frac{1}{|\mathcal{A}||\mathcal{E}|}\sum_{a\in\mathcal{A}}\sum_{e\in\mathcal{E}}\left[\text{AUROC}_T(g,e) - \text{AUROC}_S(g,a,e)\right]. \tag{17}$$

A smaller L_g indicates better retention. For the comparison in Figure 4b, L_g is evaluated using the matched label-only students.

We also compare source-specific provenance directions after mapping the probes to a shared raw activation space defined by the same VLM, layer, and ordered prompt bank. For source generator g, the corresponding raw-space discriminant direction is

$$\mathbf{v}_g = \mathbf{P}_g^{\top}\mathbf{S}_g^{-1}\mathbf{w}_g, \tag{18}$$

and the signed alignment between source generators g and h is

$$C_{g,h} = \frac{\mathbf{v}_g^{\top}\mathbf{v}_h}{\|\mathbf{v}_g\|_2\|\mathbf{v}_h\|_2}. \tag{19}$$

These directional comparisons are made only within a shared VLM activation space, whereas comparisons across VLMs focus on detection performance.

Student-teacher representation fidelity is measured using mean cosine similarity,

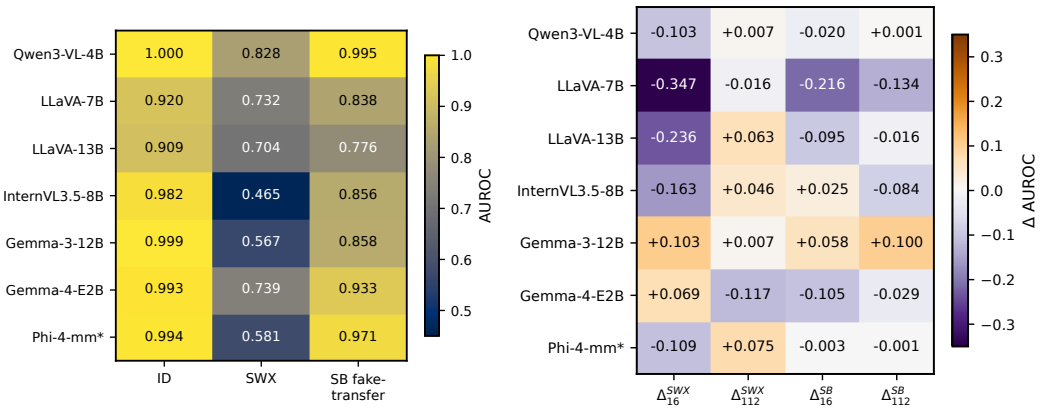
$$F_{\cos} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{z}_S(x_i)^{\top}\mathbf{z}_T^{g}(x_i)}{\|\mathbf{z}_S(x_i)\|_2\|\mathbf{z}_T^{g}(x_i)\|_2 + \epsilon}. \tag{20}$$

Additional analysis measures and diagnostic statistics are defined in § A.

4 Experiments

4.1 Evaluation protocol

We evaluate whether frozen VLM hidden states separate real and synthetic images **(RQ1)**, whether this provenance signal transfers to unseen generators and datasets **(RQ2)**, and whether it can be distilled into a lightweight image-only detector **(RQ3)** using AUROC as the primary metric. We report the pooled AUROC for SynthWildX, the macro-average AUROC across nine generators for Synthbuster, and the macro-average AUROC across 24 subsets for AIGI-Bench. Every generator or subset is given equal weight within its benchmark. The primary teacher is Qwen3-VL-4B, with its 20480-dimensional activation representation compressed to $q = 112$ dimensions before fitting a linear probe. We train separate probes for SD1.5, ADM, GLIDE, Wukong, and VQDM, and distill them into MobileNetV3-Large and ResNet-18 students. Full details on model, prompt, split, training, and uncertainty are provided in § B. For the student comparisons, distillation effects are measured as AUROC differences relative to label-only students that are matched on source-generator training data, data split, student architecture, and random seed. External benchmark images are excluded from PCA fitting, standardization, teacher-probe fitting, student training, and checkpoint selection. The deployment configuration reported in Table 1 is pre-specified and not selected based on the external benchmark results. Confidence intervals for the paired student comparisons are estimated using a hierarchical paired bootstrap. Full sampling and uncertainty details are in § B.



(a) Raw-activation AUROC.

(b) AUROC change after PCA-16 and -112 compression.

Figure 2: Cross-VLM replication. * indicates a provisional loader.

4.2 RQ1: Frozen VLM states expose a provenance signal

We first test whether provenance separability is specific to one VLM. Figure 2a shows that, under the same SD1.4 training protocol, raw frozen activations yield in-distribution AUROCs between 0.909 and 1 across seven VLMs. Therefore, the signal is not unique to Qwen3-VL-4B. External transfer shows stronger variation. SynthWildX AUROC ranges from 0.465 to 0.828, while Synthbuster fake-transfer AUROC ranges from 0.776 to 0.995. Thus, frozen VLM states consistently separate the source task, while external performance depends on the model.

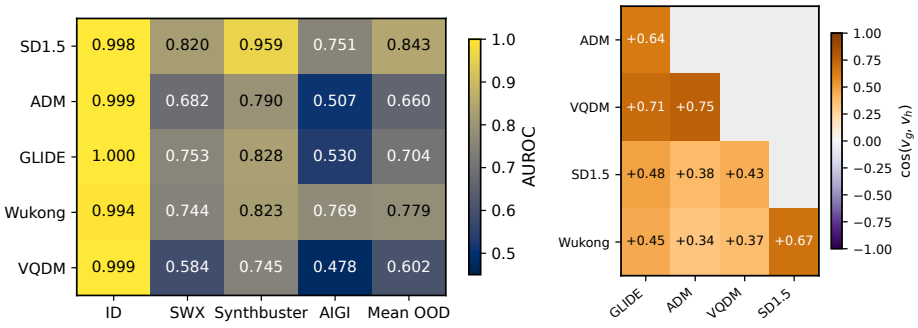
Compression also has model- and domain-dependent effects. Figure 2b shows that PCA-112 changes SynthWildX AUROC by -0.117 to +0.075 and Synthbuster AUROC by -0.134 to +0.100. The largest observed loss under more aggressive PCA-16 compression is -0.347 for LLaVA-7B on SynthWildX. Thus, a low-dimensional provenance signal exists broadly, but no single compression level is uniformly optimal across VLMs. Table S14 reports the complete values.

The representation study using Qwen3-VL-4B further distinguishes the available signatures. Raw concatenation achieves 1 in-distribution and 0.965 cross-generator AUROC, while PCA-112 retains 0.999 and 0.967, respectively. Geometry and cross-layer displacement yield cross-generator AUROCs of 0.829 and 0.912, respectively, which are below those of the final-layer representation. Figure S1 and Table S1 provide the full comparison.

4.3 RQ2: Transfer is substantial but source-dependent

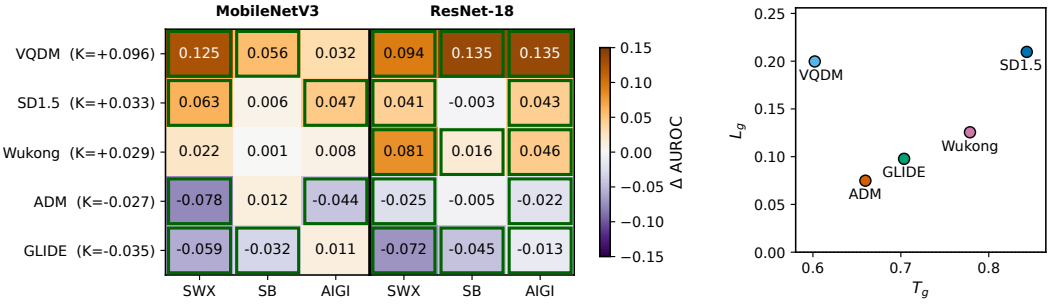
Figure 3a shows that the five generator-conditioned Qwen3-VL-4B teachers achieve nearly perfect AUROC of 0.994 to 1 on their own held-out validation data, yet their external transfer profiles differ. SD1.5 is the most transferable teacher, with a three-benchmark macro AUROC of 0.843, followed by Wukong at 0.779, GLIDE at 0.704, ADM at 0.660, and VQDM at 0.602. The ordering also changes across benchmarks. For example, Wukong is strongest on AIGI-Bench with an AUROC of 0.769, whereas SD1.5 is strongest on SynthWildX with an AUROC of 0.820 and on Synthbuster with an AUROC of 0.959. High in-distribution accuracy is therefore insufficient to predict external transfer. Compression is also source dependent. The per-teacher PCA sweep is non-monotonic, where different source teachers favor different retained dimensions, and the raw representation remains preferable for some sources. The complete sweep is reported in Figure S2a and Table S2.

Figure 3b shows that the raw-space discriminant directions are all positively aligned, with pairwise cosines ranging from 0.336 to 0.751. This partial alignment supports a shared component of the provenance signal, but the alignment is far from identity. Figure S8a further shows that mean alignment and transfer exhibit an exploratory negative rank association with descriptive Spearman $\rho = -0.60$ at $n = 5$. These observations are more consistent with partially shared, source-specific directions than with one universal provenance axis.



(a) In-distribution and external teacher AUROC. (b) Raw-space discriminant alignment.

Figure 3: Teacher transfer and discriminant structure.



(a) Weighted-KD gain. Marked cells have 95% CIs excluding zero.

(b) Teacher transferability versus student retention loss

Figure 4: Teacher-dependent distillation.

Benchmark averages also conceal variation. Figure S6a and Table S9 show that across the 24 AIGI-Bench subsets, individual teachers move from below-chance to near-perfect AUROC. Consequently, the benchmark macros summarize variation across diffusion, personalization, GAN, face-manipulation, and in-the-wild subsets. Transfer varies across training generators and target domains, with some teacher–domain pairs performing below chance.

4.4 RQ3: Lightweight students retain part of the VLM signal

We distill each teacher into an image-only student that predicts its 112-dimensional standardized PCA coordinates and reuses its frozen linear head. Figure 4a shows that the decision-weighted KD configuration does not provide a uniform gain relative to matched label-only training. Averaged across both backbones and the three external benchmarks, the AUROC changes are +0.096 for VQDM, +0.033 for SD1.5, and +0.029 for Wukong, compared with -0.027 for ADM and -0.035 for GLIDE. The paired hierarchical bootstrap identifies significant positive and negative cells, demonstrating that the distillation benefit is a property of the teacher-target pair rather than a universal effect. The component ablation in Table S16 further shows that, in the MobileNetV3 setting with SD1.5, teacher-logit supervision yields the clearest OOD improvement, whereas decision weighting and cosine alignment do not provide additional OOD gain in this ablation. We therefore treat these components as teacher- and setting-dependent rather than uniformly beneficial.

Figure 4b demonstrates that teacher transferability and student retention are distinct. T_g is the scalar transferability score for the teacher, and L_g is the label-only OOD retention loss for the student. The strongest external teacher, SD1.5, with $T_g = 0.843$, has the largest L_g of 0.210, while the weakest teacher, VQDM, with $T_g = 0.602$, reports the second-largest L_g of 0.200. Across teacher-architecture cells in Figure S8d, teacher identity explains 86% of the descriptive sum of squares, compared with 7% for architecture. Selecting a teacher for transfer and selecting one for distillability are, therefore, different decisions.

Table 1: External OOD detection and inference efficiency.

(a) External OOD Detection AUROC

Method	Params	SWX	SB	AIGI
PROBE-DINO-L	304M	**0.992**	**0.999**	**0.930**
PROBE-RN50	25.6M	<u>0.784</u>	<u>0.879</u>	0.744
HSIC Bottleneck	428M	0.525	0.581	<u>0.928</u>
UnivFD	428M	0.437	0.758	0.702
Effort	428M	0.489	0.762	0.826
Ours (MV3; SD1.5)	3.1M	0.633	0.786	0.631
Ours (RN18; SD1.5)	11.2M	0.593	0.736	0.618

(b) Inference Cost and Mean OOD AUROC

Model	MACs	Latency	VRAM
Qwen3-VL-4B	–	2032 ms	10.4 GB
MV3	0.22G	2.77 ms	0.03 GB
RN18	1.82G	2.71 ms	0.07 GB

Model	AUROC
Qwen3-VL-4B	0.843
MV3	0.683
RN18	0.649

Note: SWX=SynthWildX; SB=Synthbuster. Ours denotes a single pre-specified deployment detector: SD1.5 teacher distilled with weighted KD into the 112-d subspace, with one student per backbone and results averaged over five seeds. No teacher or recipe is selected. The five generator-conditioned teachers are separate detectors and are analyzed per teacher in Table S13 and Figure S12, rather than averaged here. The Qwen3-VL-4B rows in (b) report the SD1.5 teacher configuration corresponding to the student rows. Inference costs are measured on one RTX 6000 Ada at batch size 1. Bold and underline indicate the best and second-best AUROC, respectively. PROBE-DINO-L uses patch-aggregated scoring with a 304M-parameter fine-tuned DINOv2-L detector. Therefore, its AUROC is not directly comparable to single-forward detectors.

Higher coordinate fidelity alone does not guarantee better detection transfer. Figure 5 shows descriptive correlations of only $\rho = 0.52, 0.20, 0.37$ between in-distribution subspace cosine similarity and chance-normalized OOD retention for labels-only, PCA-16, and weighted-KD students, respectively. Figure S7a further shows that reconstruction quality reduces with PCA rank even though most OOD activation variance is concentrated in early components. Matching the teacher's full representation is thus unnecessary and insufficient for retaining its external decision rule.

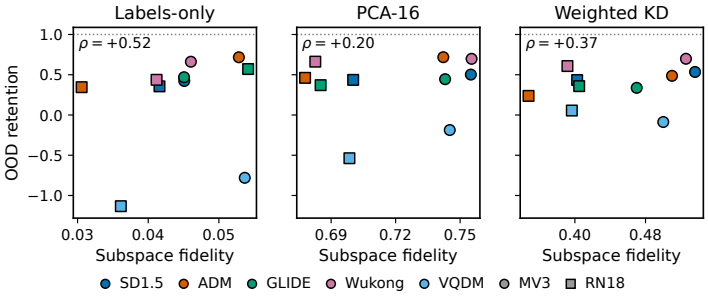


Figure 5: Representation fidelity and detector retention.

Accuracy-efficiency trade-off. Table 1 evaluates the pre-specified SD1.5-conditioned deployment configuration, rather than recipes selected on the external benchmarks. The teacher achieves an external macro-AUROC of 0.843, compared with 0.683 for MobileNetV3-Large and 0.649 for ResNet-18. These correspond to raw student-to-teacher AUROC ratios of 81.0% and 77.0%, respectively. Each student replaces eight prompt-conditioned forwards through a 4.44B-parameter VLM with one image-only forward. Latency decreases from 2032 ms to 2.77 and 2.71 ms, providing approximately 734- and 750-fold speedups. The deployed parameter counts decrease to 3.1M and 11.2M, reducing the peak VRAM usage from 10.4 GB to 0.03 GB and 0.07 GB for MobileNetV3-Large and ResNet-18, respectively.

5 Related Work

AIGI detection. Existing detectors identify synthetic images using generator artifacts, frequency statistics, local pixel relationships, reconstruction errors, or pretrained visual features (Wang et al., 2020; 2023a; Ojha et al., 2023; Ricker et al., 2024). Benchmarks such as GenImage, Synthbuster, and AIGI-Bench show that strong in-domain performance does not necessarily transfer to unseen generators or post-processing conditions (Zhu et al., 2023; Bammey, 2024; Li et al., 2026). Recent methods improve generalization through more diverse training generators or hard-example construction (Park & Owens, 2025; Cao et al., 2026). Rather than designing another dedicated detector, we investigate whether frozen VLM representations already contain a transferable provenance signal.

VLM representations. Pretrained vision-language models provide transferable features for synthetic-image detection. UnivFD, for example, trains a linear classifier on frozen CLIP embeddings (Ojha et al., 2023). Other work uses VLMs directly for authenticity reasoning, requiring the full model during inference (Xia et al., 2025). We instead extract prompt-conditioned hidden states without generating text or updating the VLM. Behavioral Provenance Signatures characterize both these states and their relationships across prompts and layers.

Knowledge distillation. Knowledge distillation transfers predictions or intermediate representations from a large teacher to a smaller student (Hinton et al., 2015; Romero et al., 2014). Direct feature matching, however, may emphasize coordinates that have little influence on the teacher's decision. Our student predicts a compressed VLM provenance representation and reuses the frozen teacher head. Decision-weighted matching gives greater importance to coordinates that contribute more strongly to that decision, enabling image-only inference without running the VLM.

6 Discussion and Limitations

Our results distinguish separability, transferability, and distillability. Frozen VLM activations contain differences that support real-versus-synthetic classification, but this should not be interpreted as an explicit or causal representation of image provenance. Cross-domain performance instead suggests that the activation space combines shared provenance structure with source- and dataset-specific cues. Consequently, strong in-domain performance alone does not guarantee robust transfer.

Distillability also depends on whether the teacher's decision rule can be learned by the student's architecture. High teacher accuracy does not necessarily imply strong student retention, and accurate global reconstruction may not preserve the coordinates most relevant to detection. Decision-weighted matching addresses this issue by emphasizing coordinates used by the teacher probe, although their importance may change under distribution shifts. Teacher selection should therefore consider transferability and student compatibility in addition to source-domain accuracy.

The practical contribution of BPS is its accuracy-efficiency trade-off. Distillation removes the VLM and prompts from the inference process, enabling deployment with a small image encoder. Although specialized detectors remain more accurate, BPS provides an efficient way to study and transfer provenance signals derived from VLMs.

Limitations. The main transfer and distillation experiments use five generator-conditioned teachers derived primarily from Qwen3-VL-4B, while the broader cross-VLM evaluation follows a narrower protocol. The teachers also require offline VLM feature extraction and may capture dataset, preprocessing, or compression cues alongside generator artifacts. A broader evaluation across VLMs, generators, transformations, and synthesis models would further establish their generality.

Correlations across teachers remain descriptive due to the limited number of teacher configurations. Our evaluation also focuses on binary ranking and a single hardware environment, rather than on attribution, localization, calibration, adversarial robustness, or platform-independent efficiency. Future work includes broader multi-source teachers and a more comprehensive robustness evaluation.

7 Conclusion

This work introduced BPS, which derives provenance representations from the prompt-conditioned hidden states of frozen VLMs and distills them into lightweight image encoders. Across seven VLMs, frozen activations support linear separation of real and synthetic images (RQ1), while transfer to unseen generators and datasets varies across models and source generators (RQ2). Generator-conditioned probes exhibit partially aligned but distinct discriminant directions, suggesting both shared and source-dependent information. Lightweight students retain part of the teacher signal without test-time VLM execution (RQ3), but distillation gains depend on the teacher, target domain, and student architecture. Higher representation fidelity does not consistently yield better external detection, underscoring the need to evaluate separability, transferability, and distillability separately. The deployed students achieve measured inference speedups exceeding $700\times$ on the evaluated hardware, at the cost of lower external AUROC than the teacher. Future work should test whether combining teachers from multiple generators improves transfer and whether these benefits persist across VLM families and student architectures.

AI USE STATEMENT

Generative AI tools were used to provide feedback on research methodology and experimental design, assist with the interpretation and verification of experimental results, and support the drafting and revision of the manuscript. They were also used for literature search and summarization, identifying related work and research gaps, improving readability and organization, editing and formatting, and assisting with code editing and debugging.

All AI-assisted methodological suggestions, code changes, numerical results, equations, citations, and manuscript claims were reviewed and verified by the authors against the implemented experiments, source code, experimental artifacts, and cited literature. Generative AI was not used to generate the experimental datasets or to replace author judgment in determining the final methodology, experimental conclusions, or scientific claims. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies the detection of AI-generated images using representations from frozen vision-language models. The intended applications include digital forensics, content moderation, and media verification. However, synthetic-image detection is inherently dual use: improved understanding of detector behavior may also inform efforts to evade detection. Our work does not develop image-generation attacks or evasion procedures, and we report limitations and distribution-dependent behavior to discourage treating the proposed detectors as universal authenticity classifiers.

The study uses existing research datasets and benchmark images and does not involve the recruitment of human participants or the collection of new personal data. Because detector performance varies across generators, datasets, and processing conditions, predictions from BPS-based detectors should not be used as the sole basis for high-stakes decisions about the authenticity, provenance, or ownership of individual images. In particular, false positives may incorrectly classify authentic content as synthetic, while false negatives may fail to identify generated content. We therefore view the proposed framework primarily as a tool for studying and transferring provenance-related representations rather than as a standalone mechanism for automated enforcement.

REPRODUCIBILITY STATEMENT

We provide detailed information needed to reproduce the experiments throughout the main paper and appendix. § 4 specifies the evaluation protocol, datasets, models, and primary metrics. § B documents the frozen-VLM extraction procedure, including the exact prompt bank, selected layers, token-extraction rule, model-specific preprocessing, numerical precision, source-domain construction, and train/validation splits. It also specifies the PCA and logistic-regression settings, student architectures, loss configurations, random seeds, optimization hyperparameters, data augmentation, checkpoint-selection procedure, and deployment measurement protocol. § A gives the complete mathematical definitions of the representations, distillation objectives, discriminant-direction comparisons, and transfer and retention metrics. The supplementary tables report the underlying per-generator, per-domain, cross-VLM, compression, robustness, and ablation results, including five-seed summaries and uncertainty estimates. External evaluation data are kept isolated from teacher fitting, student training, and checkpoint selection, and the deployment configuration reported in the main paper is fixed independently of the external benchmark results.

For prompt-paraphrase evaluation, § B.5 specifies training-only fitting of PCA (where used), scaling, and the $C = 0.1$ probe; validation images are used only for paired scoring. The robustness tables and figures use common saved clean/paraphrased scores (`robustness_results.json` and `robustness_scores.npz`), with source-cache hashes and image/label pairing checks. In particular, both axes of Figure S5 derive from that same frozen pipeline, and its AUROC drops are exact sign reversals of the corresponding signed changes in Figure S8c.

REFERENCES

Ahmed Abdullah, Nikolas Ebert, and Oliver Wasenmüller. Tap into the patch tokens: Leveraging vision foundation model features for ai-generated image detection. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, CVPRW, pp. 379–389, June 2026.

Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2024. doi: 10.1109/OJSP.2023.3337714.

Zijie Cao, Weijie Tu, Yao Xiao, Weijian Deng, Weiyan Chen, Liang Lin, and Pengxu Wei. Where detectors fail: Probing generative space for generalizable AI-generated image detection. In *Forty-third International Conference on Machine Learning*, ICML, 2026.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Niesner, and Luisa Verdoliva. Raising the Bar of AI-generated Image Detection with CLIP . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, CVPRW, pp. 4356–4366, Los Alamitos, CA, USA, Jun 2024. IEEE Computer Society. doi: 10.1109/CVPRW63382.2024.00439.

Patrick Grommelt, Louis Weiss, Franz-Josef Pfreundt, and Janis Keuper. Fake or jpeg? revealing common biases in generated image detection datasets. In *Computer Vision – ECCV 2024 Workshops: Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXII*, ECCV, pp. 80–95, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-92088-2. doi: 10.1007/978-3-031-92089-9_6. URL https://doi.org/10.1007/978-3-031-92089-9_6.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

Christos Koutlis and Symeon Papadopoulos. Leveraging Representations from Intermediate Encoder-Blocks for Synthetic Image Detection. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXII*, ECCV, pp. 394–411, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73219-5. doi: 10.1007/978-3-031-73220-1_23. URL https://doi.org/10.1007/978-3-031-73220-1_23.

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is Artificial Intelligence Generated Image Detection a Solved Problem? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, NeurIPS, 2026.

Yewon Lim, Changyeon Lee, Aerin Kim, and Oren Etzioni. DistilDIRE: A small, fast, cheap and lightweight diffusion synthesized deepfake detection. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024. URL https://openreview.net/forum?id=OiPHxdXei1.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards Universal Fake Image Detectors that Generalize Across Generative Models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pp. 24480–24489, 2023. doi: 10.1109/CVPR52729.2023.02345.

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR, pp. 8245–8257, June 2025.

Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images. In *The Thirteenth International Conference on Learning Representations*, ICLR, 2025.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pp. 9130–9140, June 2024.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *CoRR*, abs/1412.6550, 2014.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pp. 8692–8701, 2020. doi: 10.1109/CVPR42600.2020.00872.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for Diffusion-Generated Image Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, ICCV, pp. 22445–22455, October 2023a.

Zhenting Wang, Chen Chen, Yi Zeng, Lingjuan Lyu, and Shiqing Ma. Where did i come from? origin attribution of ai-generated images. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36 of *NeurIPS*, pp. 74478–74500. Curran Associates, Inc., 2023b. doi: 10.52202/075280-3256.

Cheng Xia, Manxi Lin, Jiexi Tan, Xiaoxiong Du, Yang Qiu, Junjun Zheng, Xiang-Heng Kong, Yuning Jiang, and Bo Zheng. Mirage: Towards ai-generated image detection in the wild. In *AAAI Conference on Artificial Intelligence*, AAAI, 2025.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A Sanity Check for AI-generated Image Detection. In *The Thirteenth International Conference on Learning Representations*, ICLR, 2025.

Yue Zhou, Xinan He, Kaiqing Lin, Bing Fan, Feng Ding, and Bin Li. Simplicity prevails: The emergence of generalizable aigi detection in visual foundation models, 2026. URL `https://arxiv.org/abs/2602.01738`.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. GenImage: a million-scale benchmark for detecting AI-generated image. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

A Additional Method Details

This appendix provides the detailed definitions, implementation settings, and supplementary analyses supporting Sections 3 and 4. The main paper presents the primary BPS construction and analysis quantities. Here we provide the complete loss definitions, secondary geometry and fidelity measures, data and training protocols, uncertainty procedures, and extended diagnostic results. Unless stated otherwise, all transformations that define a teacher are estimated only from the corresponding source-domain training split and are frozen before student training or external evaluation. A teacher includes PCA directions, centering statistics, standardization scales, and the linear provenance probe.

A.1 Notation and Signature Construction

Let K denote the number of visual-description prompts and let L denote the number of selected layers. We write the selected layers as $\mathcal{L} = \{\ell_l\}_{l=1}^{L}$, ordered from shallow to deep. In the experiments, $K = 8$ and $L = 4$ (§ B). For image x, prompt p_k, and selected layer ℓ_l, the frozen VLM activation is $\mathbf{h}_{\ell_l,k}(x) \in \mathbb{R}^d$.

Prompt geometry. For each selected layer, we form the stabilized cosine similarity between every unordered prompt pair,

$$G_{ij}^{(\ell_l)}(x) = \frac{\mathbf{h}_{\ell_l,i}(x)^\top \mathbf{h}_{\ell_l,j}(x)}{\|\mathbf{h}_{\ell_l,i}(x)\|_2 \|\mathbf{h}_{\ell_l,j}(x)\|_2 + \epsilon}, \qquad 1 \le i < j \le K. \tag{21}$$

Concatenating the strictly upper-triangular entries in a fixed layer–prompt order gives

$$\mathbf{s}_{\mathrm{geom}}(x) = \mathrm{Concat}_{\substack{1\le l\le L \\ 1\le i<j\le K}} G_{ij}^{(\ell_l)}(x), \tag{22}$$

with dimension

$$d_{\mathrm{geom}} = L\binom{K}{2} = \frac{LK(K-1)}{2}. \tag{23}$$

The fixed concatenation order is retained for fitting and evaluation, so coordinates have the same semantic meaning across samples within a VLM.

Layer trajectory. To measure how a prompt-conditioned state changes across adjacent selected layers, we use

$$t_{l,k}(x) = \frac{\mathbf{h}_{\ell_l,k}(x)^\top \mathbf{h}_{\ell_{l+1},k}(x)}{\|\mathbf{h}_{\ell_l,k}(x)\|_2 \|\mathbf{h}_{\ell_{l+1},k}(x)\|_2 + \epsilon}, \qquad 1 \le l < L, \quad 1 \le k \le K. \tag{24}$$

The complete trajectory vector is

$$\mathbf{s}_{\mathrm{traj}}(x) = \mathrm{Concat}_{\substack{1\le l<L \\ 1\le k\le K}} t_{l,k}(x), \tag{25}$$

with dimension

$$d_{\mathrm{traj}} = (L-1)K. \tag{26}$$

Combined relational signature. When geometry and trajectory are analyzed jointly, we concatenate the two vectors,

$$\mathbf{s}_{\mathrm{joint}}(x) = \left[\mathbf{s}_{\mathrm{geom}}(x)^\top, \mathbf{s}_{\mathrm{traj}}(x)^\top\right]^\top, \tag{27}$$

whose dimension is

$$d_{\mathrm{joint}} = \frac{LK(K-1)}{2} + (L-1)K. \tag{28}$$

These relational signatures are diagnostic alternatives to the primary teacher representation; the primary teacher uses the final-layer prompt concatenation below.

Primary raw representation. The primary representation concatenates the K prompt-conditioned states from the deepest selected language layer,

$$\mathbf{r}(x) = \mathrm{Concat}_{k=1}^{K} \mathbf{h}_{\ell_L,k}(x) \in \mathbb{R}^{Kd}. \tag{29}$$

For Qwen3-VL-4B, $K = 8$ and $d = 2560$, hence $\mathbf{r}(x) \in \mathbb{R}^{20480}$.

Effective representation dimension. For descriptive comparisons of representation compactness, let $\lambda_1, \ldots, \lambda_D$ denote the eigenvalues of the empirical covariance of a representation before coordinate standardization. We summarize spectral concentration by the participation-ratio effective dimension

$$d_{\text{eff}} = \frac{\left(\sum_{j=1}^{D} \lambda_j\right)^2}{\sum_{j=1}^{D} \lambda_j^2}. \tag{30}$$

A smaller value indicates that the observed variance is concentrated in fewer directions; this quantity is descriptive and is not used to select the deployment configuration.

A.2 Distillation Losses

Consider a mini-batch

$$\mathcal{B} = \{(x_i, y_i)\}_{i=1}^{B}. \tag{31}$$

For a logit $a \in \mathbb{R}$, define $\text{sigmoid}(a) = 1/(1 + e^{-a})$ and binary cross-entropy

$$\ell_{\text{BCE}}(a, t) = -t \log \text{sigmoid}(a) - (1 - t) \log\left(1 - \text{sigmoid}(a)\right), \qquad t \in [0, 1]. \tag{32}$$

The teacher coordinates $\mathbf{z}_T^g(x)$, teacher logit $a_T^g(x)$, and probe weights $\mathbf{w}_g$ are frozen throughout student optimization.

Decision-weighted subspace loss. Because the teacher coordinates are standardized before fitting the linear probe, $|w_{g,j}|$ directly measures the change in teacher logit per unit change in standardized coordinate j. We therefore define

$$\alpha_{g,j} = |w_{g,j}|, \qquad j = 1, \ldots, q, \tag{33}$$

and normalize these weights within the matching term. The decision-weighted coordinate loss is

$$\mathcal{L}_{\text{sub}} = \frac{1}{B} \sum_{i=1}^{B} \frac{\sum_{j=1}^{q} \alpha_{g,j} \left(z_{\mathcal{S},ij} - z_{\mathcal{T},ij}^g\right)^2}{\sum_{j=1}^{q} \alpha_{g,j}}. \tag{34}$$

The normalization keeps the scale of the loss from changing solely because the probe has a larger or smaller overall weight norm. The corresponding unweighted loss is

$$\mathcal{L}_{\text{sub}}^{\text{unweighted}} = \frac{1}{Bq} \sum_{i=1}^{B} \sum_{j=1}^{q} \left(z_{\mathcal{S},ij} - z_{\mathcal{T},ij}^g\right)^2. \tag{35}$$

Cosine-alignment loss. The coordinate loss is complemented by a direction-matching term,

$$\mathcal{L}_{\cos} = \frac{1}{B} \sum_{i=1}^{B} \left[1 - \frac{\mathbf{z}_{\mathcal{S}}(x_i)^\top \mathbf{z}_{\mathcal{T}}^g(x_i)}{\|\mathbf{z}_{\mathcal{S}}(x_i)\|_2 \|\mathbf{z}_{\mathcal{T}}^g(x_i)\|_2 + \epsilon}\right]. \tag{36}$$

This term is invariant to a common positive rescaling of either vector and therefore targets angular agreement rather than coordinate magnitude.

Teacher-logit distillation. Let $\tau > 0$ be the distillation temperature. The softened teacher probability is

$$\widetilde{p}_{\mathcal{T},i}^g = \text{sigmoid}\left(\frac{a_{\mathcal{T}}^g(x_i)}{\tau}\right). \tag{37}$$

The logit-distillation loss is

$$\mathcal{L}_{\text{KD}} = \frac{\tau^2}{B} \sum_{i=1}^{B} \ell_{\text{BCE}}\left(\frac{a_{\mathcal{S}}^g(x_i)}{\tau}, \widetilde{p}_{\mathcal{T},i}^g\right). \tag{38}$$

The factor τ^2 follows the usual temperature-rescaling convention so that the contribution of this term does not collapse as the logits are softened.

Supervised label loss. Ground-truth supervision is

$$\mathcal{L}_y = \frac{1}{B}\sum_{i=1}^{B} \ell_{\mathrm{BCE}}\left(a_{\mathcal{S}}^{g}(x_i), y_i\right). \tag{39}$$

The complete objective is

$$\mathcal{J} = \lambda_{\mathrm{sub}}\mathcal{L}_{\mathrm{sub}} + \lambda_{\cos}\mathcal{L}_{\cos} + \lambda_{\mathrm{KD}}\mathcal{L}_{\mathrm{KD}} + \lambda_y\mathcal{L}_y. \tag{40}$$

The exact coefficient settings for the evaluated arms are given in §B.3. The loss ablation in Table S16 treats decision weighting, cosine matching, and logit distillation as separable components rather than attributing all student behavior to a single term.

A.3 Provenance-Direction and Subspace Comparison

The main paper uses raw-space discriminant alignment for the RQ2 comparison. This section gives the corresponding derivation and the secondary summaries used in the supplementary analyses. Teacher probes use source-specific PCA bases, so the fitted coefficient vectors $\mathbf{w}_g$ do not live in a common coordinate system and cannot be compared directly across generators. We therefore map each probe back to the common raw final-layer activation space before measuring directional alignment.

The standardized teacher coordinates are

$$\mathbf{z}_T^g(x) = \mathbf{S}_g^{-1}\left[\mathbf{P}_g\left(\mathbf{r}(x) - \boldsymbol{\mu}_g^r\right) - \boldsymbol{\mu}_g^u\right]. \tag{41}$$

Substituting this expression into the teacher logit yields

$$a_T^g(x) = \mathbf{v}_g^\top \mathbf{r}(x) + \widetilde{b}_g, \tag{42}$$

where

$$\mathbf{v}_g = \mathbf{P}_g^\top \mathbf{S}_g^{-1}\mathbf{w}_g \in \mathbb{R}^{Kd} \tag{43}$$

and

$$\widetilde{b}_g = b_g - \mathbf{w}_g^\top \mathbf{S}_g^{-1}\left(\mathbf{P}_g\boldsymbol{\mu}_g^r + \boldsymbol{\mu}_g^u\right). \tag{44}$$

Thus, $\mathbf{v}_g$ is the source-specific linear discriminant direction expressed in the common raw activation coordinates. The signed cosine alignment between teachers g and h is

$$C_{g,h} = \frac{\mathbf{v}_g^\top \mathbf{v}_h}{\|\mathbf{v}_g\|_2\|\mathbf{v}_h\|_2}. \tag{45}$$

For the generator-level summaries, mean alignment is

$$A_g = \frac{1}{|\mathcal{G}| - 1}\sum_{\substack{h\in\mathcal{G}\\ h\neq g}} C_{g,h}, \tag{46}$$

where $\mathcal{G}$ is the set of five source generators. Positive $C_{g,h}$ indicates that the two classifiers assign increasing synthetic evidence in broadly similar raw-space directions; values far below one indicate substantial source-specific components.

Decision-weight concentration. We report two complementary summaries of the linear probe weights. The first is the ℓ_1/ℓ_∞ support surrogate

$$d_{1/\infty}(g) = \frac{\|\mathbf{w}_g\|_1}{\|\mathbf{w}_g\|_\infty}, \tag{47}$$

which equals one for a vector concentrated on a single coordinate and increases as the weight mass spreads across coordinates. The second is the participation ratio

$$d_{\mathrm{PR}}(g) = \frac{\left(\sum_{j=1}^{q} w_{g,j}^2\right)^2}{\sum_{j=1}^{q} w_{g,j}^4}. \tag{48}$$

These quantities characterize decision-weight concentration; they should not be confused with the covariance-based representation effective dimension d_{eff} defined above.

Principal-angle comparison of PCA bases. Classifier-direction alignment and PCA-basis overlap answer different questions. Let the rows of $\mathbf{P}_g$ and $\mathbf{P}_h$ be orthonormal PCA directions. The singular values of $\mathbf{P}_g\mathbf{P}_h^\top$ are the cosines of the principal angles, $\{\cos\theta_j\}_{j=1}^q$. We summarize their overlap by

$$\overline{\cos\theta}_{g,h} = \frac{1}{q}\sum_{j=1}^{q}\cos\theta_j, \tag{49}$$

and report the scaled dissimilarity

$$D_{g,h}^{\mathrm{PA}} = 10^3\left(1 - \overline{\cos\theta}_{g,h}\right) \tag{50}$$

in Table S6. Smaller values indicate greater overlap between the retained PCA subspaces. This is an unsigned subspace comparison and is distinct from the signed raw-space classifier alignment $C_{g,h}$.

Shared-PCA coordinate analysis. For coordinate-level comparisons in Figures S3 and S4, a single PCA basis is fitted on pooled source-training activations. A separate linear classifier is then fitted for each source in that common coordinate system. This shared-basis diagnostic makes individual coefficient indices comparable across source teachers; it is separate from the source-specific PCA bases used by the primary teacher models.

A.4 Transfer, Retention, and Fidelity Measures

The main paper introduces the primary transfer, retention, and fidelity quantities used in RQ3. For completeness, this section restates their aggregation and defines the additional signed-gap, KD-benefit, chance-normalized retention, and coordinate-fidelity summaries used in the supplementary figures and tables. Let $\mathcal{E}$ denote the set of external benchmark domains, $\mathcal{A}$ the set of student architectures, and $\mathcal{S}$ the set of training seeds.

Teacher transferability. For source teacher g, external transferability is

$$T_g = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\mathrm{AUROC}_{\mathcal{T}}(g,e). \tag{51}$$

For the primary analysis, $\mathcal{E} = \{\mathrm{SWX}, \mathrm{SB}, \mathrm{AIGI}\}$, where Synthbuster and AIGI are themselves benchmark macros defined in §B.5.

Student retention. A teacher-minus-student retention loss is

$$L_g = \frac{1}{|\mathcal{A}|\,|\mathcal{E}|}\sum_{a\in\mathcal{A}}\sum_{e\in\mathcal{E}}\left[\mathrm{AUROC}_{\mathcal{T}}(g,e) - \mathrm{AUROC}_{\mathcal{S}}(g,a,e)\right]. \tag{52}$$

Smaller L_g indicates better retention, and $L_g < 0$ means that the student exceeds its teacher on average over the included architecture–domain cells. For the label-only retention summaries, we also use the signed student-minus-teacher gap

$$G_g = \frac{1}{|\mathcal{A}|\,|\mathcal{E}|}\sum_{a\in\mathcal{A}}\sum_{e\in\mathcal{E}}\left[\mathrm{AUROC}_{\mathrm{labels}}(g,a,e) - \mathrm{AUROC}_{\mathcal{T}}(g,e)\right], \tag{53}$$

so that $G_g = -L_g$ when L_g is evaluated on the same label-only cells. Larger G_g (i.e., a value closer to or above zero) therefore indicates stronger retention.

For the chance-normalized fidelity–retention plots, we normalize the student and teacher macro-AUROCs around the random-ranking baseline of 0.5:

$$R_{g,a}^{\mathrm{chance}} = \frac{\overline{\mathrm{AUROC}}_{\mathcal{S}}(g,a) - 0.5}{\overline{\mathrm{AUROC}}_{\mathcal{T}}(g) - 0.5}, \tag{54}$$

where the bars denote the same three-benchmark macro used for the corresponding teacher–student comparison. A value of one means that the student preserves the teacher's above-chance margin, zero corresponds to chance-level student performance, and negative values indicate below-chance ranking under the teacher's score orientation.

Distillation gain. For a fixed source teacher, student architecture, and external domain, the seed-averaged gain of the decision-weighted recipe over its matched label-only baseline is

$$\Delta^{\mathrm{KD}}_{g,a,e} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left[\mathrm{AUROC}^{(s)}_{\mathrm{weighted}}(g,a,e) - \mathrm{AUROC}^{(s)}_{\mathrm{labels}}(g,a,e) \right]. \tag{55}$$

The generator-level KD benefit shown in Table S11 is

$$K_g = \frac{1}{|\mathcal{A}|\,|\mathcal{E}|} \sum_{a \in \mathcal{A}} \sum_{e \in \mathcal{E}} \Delta^{\mathrm{KD}}_{g,a,e}. \tag{56}$$

Thus $K_g > 0$ means that the decision-weighted recipe improves AUROC on average over the evaluated student architectures and external benchmarks, while $K_g < 0$ indicates an average decrease.

Representation fidelity. Student–teacher cosine fidelity on a set of N images is

$$F_{\cos} = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbf{z}_S(x_i)^\top \mathbf{z}^g_T(x_i)}{\|\mathbf{z}_S(x_i)\|_2 \|\mathbf{z}^g_T(x_i)\|_2 + \epsilon}. \tag{57}$$

For coordinate j, reconstruction fidelity is additionally summarized by

$$R^2_j = 1 - \frac{\sum_i \left(z_{\mathcal{S},ij} - z^g_{\mathcal{T},ij} \right)^2}{\sum_i \left(z^g_{\mathcal{T},ij} - \bar{z}^g_{\mathcal{T},j} \right)^2}, \tag{58}$$

where $\bar{z}^g_{\mathcal{T},j}$ is the sample mean of teacher coordinate j. The decision-weighted summary used in Table S16 is

$$R^2_\alpha = \frac{\sum_{j=1}^{q} |w_{g,j}| R^2_j}{\sum_{j=1}^{q} |w_{g,j}|}. \tag{59}$$

Mean squared error and teacher–student logit correlation are reported as additional diagnostics where applicable. These fidelity measures quantify representation matching; they are not themselves detection metrics, which is why the paper analyzes their relationship to OOD AUROC separately.

B Experimental Setup

B.1 Frozen VLM Extraction

The primary experiments use Qwen3-VL-4B. For each image, we query the frozen VLM with the same bank of eight fixed visual-description prompts, run the model in evaluation mode without text generation, and retain hidden states from four selected language layers. The primary `concat_last` representation has 20480 coordinates for Qwen3-VL-4B and is compressed to $q = 112$ coordinates unless otherwise stated. The cross-VLM replication additionally evaluates LLaVA-1.5-7B, LLaVA-1.5-13B, InternVL3.5-8B, Gemma-3-12B-it, Gemma-4-E2B-it, and Phi-4-mm under the same SD1.4 training protocol.

Prompt specification. The same eight prompts are used verbatim for every VLM:

1. "Describe the image accurately in one concise sentence."
2. "Identify the main objects in the image and briefly describe their important visual attributes."
3. "Describe the spatial relationships among the main objects and regions in the image."
4. "Describe the lighting, shadows, reflections, and illumination conditions visible in the image."
5. "Describe the visible materials, surfaces, and textures in the image."
6. "Describe any physically notable relationships involving shape, perspective, contact, occlusion, or object structure."

7. "Describe small visual details that are important for understanding the image."
8. "Summarize what is happening in the scene and how the visible elements relate to one another."

These prompts are generic visual-description queries rather than explicit real-versus-synthetic or provenance questions. Their role is to induce a fixed set of prompt-conditioned internal responses from the frozen VLM.

Layer specification. Let L_{model} denote the number of language layers in a VLM and let $L = 4$ denote the number of selected layers. We use the same quartile rule for every model,

$$\ell_l = \left\lceil \frac{l L_{\text{model}}}{L} \right\rceil, \qquad l = 1, \ldots, L,$$

where layer indices are 1-indexed and refer to the hidden state after the corresponding language-transformer layer. The last selected layer is therefore always the final language layer.

Concretely, the selected layers are Qwen3-VL-4B-Instruct ($L_{\text{model}} = 36$): $\{9, 18, 27, 36\}$; InternVL3.5-8B (36): $\{9, 18, 27, 36\}$; Gemma-3-12B-it (48): $\{12, 24, 36, 48\}$; Gemma-4-E2B-it (35): $\{9, 18, 27, 35\}$; LLaVA-1.5-7B and LLaVA-1.5-13B (32): $\{8, 16, 24, 32\}$; and Phi-4-mm (32): $\{8, 16, 24, 32\}$.

The native hidden width d is 2560 for Qwen3-VL-4B, 4096 for InternVL3.5-8B and LLaVA-1.5-7B, 3840 for Gemma-3-12B, 1536 for Gemma-4-E2B, 5120 for LLaVA-1.5-13B, and 3072 for Phi-4-mm.

Raw feature construction. For each prompt and selected layer, we retain the hidden state at the final non-padding token of the fully rendered multimodal chat input, as determined from the attention mask. Because the chat template is rendered with `add_generation_prompt=true`, this position belongs to the complete model-specific input sequence, including any assistant-generation prefix introduced by the template. Thus, the extraction rule is held fixed across VLMs even though the identity of the terminal template token can differ across model families.

For N images, the resulting activation cache has shape $[N, L, K, d] = [N, 4, 8, d]$. The `concat_last` representation concatenates the $K = 8$ prompt-conditioned states from the last selected language layer:

$$r(x) = \text{Concat}_{k=1}^{8} h_{\ell_L, k}(x) \in \mathbb{R}^{8d}.$$

For Qwen3-VL-4B, this gives $8 \times 2560 = 20480$ raw coordinates.

We additionally cache a pooled image-token representation obtained by averaging the multimodal input image-token states and then averaging over prompts. This representation has dimension d and is used only for diagnostic analyses; it is not part of the primary `concat_last` teacher representation.

Input processing and chat formatting. Extraction applies no fixed image resize shared across VLMs. Images are converted to RGB and passed through each model's native processor, so each VLM retains its own model-specific image-resolution policy. Each prompt is formatted as a single-turn user message containing one image and the corresponding prompt text and is rendered with the model's native chat template using `apply_chat_template(..., add_generation_prompt=true)`. Models are run in evaluation mode under inference mode, with KV caching disabled and without autoregressive text generation.

Numerical precision. The primary loader uses the checkpoint-resolved precision specified by `dtype=auto`, while the cross-VLM extraction code explicitly requests bfloat16 where supported. Extracted hidden states are stored in float16 and are cast to float32 before PCA fitting, standardization, logistic-regression fitting, or downstream scoring.

B.2 Source Domains and Data Splits

The five primary source generators are SD1.5, ADM, GLIDE, Wukong, and VQDM. For each source generator g, we independently fit a PCA transform, coordinate-wise standardizer, and logistic-regression provenance head using only the corresponding source-domain training activations. The

primary external benchmarks are SynthWildX, Synthbuster, and AIGI-Bench. SynthWildX is evaluated as a single pooled benchmark, Synthbuster is macro-averaged across nine generators, and AIGI-Bench is macro-averaged across 24 subsets, each containing 500 real and 500 synthetic images in the canonical manifests. Table S9 lists the AIGI-Bench subsets.

Source datasets. Each source teacher is constructed from a GenImage-style directory with `<split>/{ai,nature}` subdirectories. The `ai` directory contains synthetic images from the corresponding generator, whereas `nature` contains real photographs from the associated GenImage real-image collection. We use the pre-generated images distributed under the `Stable Diffusion V1.5`, `ADM`, `GLIDE`, `Wukong`, and `VQDM` source roots; synthetic images are not regenerated as part of our experiments.

For each of the five primary teachers, the training manifest contains 2000 images, comprising 1000 synthetic and 1000 real images. The corresponding validation manifest contains 1500 images, comprising 500 synthetic and 1000 real images. There is no additional in-distribution test split. The validation split is used for model and checkpoint selection, where specified, while all reported out-of-distribution generalization results are obtained from the external benchmarks described above.

The cross-VLM replication and Qwen3-VL-8B comparison instead use the `Stable Diffusion V1.4` source root under the same manifest-generation procedure, with 1000 training images (500 synthetic and 500 real) and 1000 validation images (500 synthetic and 500 real).

Sampling and split construction. Sampling is stratified by ImageNet semantic class, inferred from the source subdirectory or filename synset prefix. Only classes present under both the real and synthetic labels are included in the shared sampling pool. Within each $(\text{label}, \text{class})$ cell, images are sampled without replacement using seed 42.

Therefore, a given file path cannot occur more than once within a manifest. Training and validation samples are drawn from their respective disjoint source directories and consequently do not share paths. A stable MD5-derived identifier computed from each relative path is used for sample tracking and for matching predictions across processing stages. This identifier is path-based and is not used as a content-level perceptual deduplication procedure.

The path-level guarantees above apply within each teacher manifest and between its train and validation directories. They do not establish content-level perceptual deduplication across independently constructed source-teacher manifests. Accordingly, the cross-teacher alignment analysis should be interpreted as a comparison of fitted decision geometry rather than as evidence about the causal origin of shared directions.

External-data isolation. Images from SynthWildX, Synthbuster, and AIGI-Bench are excluded from PCA fitting, estimation of coordinate-standardization statistics, teacher-head fitting, student training, and student checkpoint selection. All teacher and student artifacts are therefore frozen before external benchmark evaluation.

The deployment configuration reported in Table 1 is also not selected using these external benchmarks. It is the pre-specified SD1.5-conditioned teacher with $q = 112$ and decision-weighted KD, evaluated separately for MobileNetV3-Large and ResNet-18 students. Thus, no source teacher, target dimension, distillation recipe, or student checkpoint is selected using the Table 1 external-benchmark scores.

The supplementary sensitivity analyses are descriptive and post hoc, and neither is used to select or modify the pre-specified deployment configuration in Table 1. Figure S12 varies λ_{KD} and target dimension q for the ADM, GLIDE, and VQDM teachers to characterize teacher-dependent responses to distillation strength and compression. Table S16 instead holds the SD1.5-conditioned deployment teacher and MobileNetV3 backbone fixed and ablates the student-stage loss components. Keeping these two analyses separate avoids using external OOD scores to retrospectively choose the Table 1 configuration.

B.3 Teacher Probes and Student Training

The primary teacher representation uses PCA with $q = 112$ retained components, followed by coordinate-wise standardization and logistic regression with regularization parameter $C = 0.1$. We

additionally evaluate

$$q \in \{16, 32, 64, 112, 256, 512\}$$

as descriptive compression variants. These PCA-rank sweeps are used to characterize source-dependent compressibility rather than to define a single benchmark-specific teacher rank.

Student experiments use MobileNetV3-Large and ResNet-18 backbones. For a teacher with target dimension q, the student terminates in a newly initialized linear projection to $\mathbb{R}^q$. Thus, the primary experiments use a 112-dimensional output projection, whereas PCA-16 and the target-dimension sensitivity experiments use output dimensions matching their respective teacher spaces. In every case, the student is scored using the corresponding frozen teacher head. Each student configuration is repeated with random seeds 42-46.

Optimization objective. We use the loss terms defined in § A and write the complete student objective as

$$\mathcal{J} = \lambda_{\text{sub}}\mathcal{L}_{\text{sub}} + \lambda_{\cos}\mathcal{L}_{\cos} + \lambda_{\text{KD}}\mathcal{L}_{\text{KD}} + \lambda_y\mathcal{L}_y.$$

The distillation temperature is fixed to $\tau = 3.0$, with soft teacher target

$$\widetilde{p}^g_{\mathcal{T}}(x) = \text{sigmoid}\left(\frac{a^g_{\mathcal{T}}(x)}{\tau}\right),$$

and

$$\mathcal{L}_{\text{KD}} = \tau^2 \,\text{BCE}\left(\frac{a^g_{\mathcal{S}}}{\tau}, \widetilde{p}^g_T\right).$$

The evaluated training arms are as follows.

- **Labels only:**
$$(\lambda_{\text{sub}}, \lambda_{\cos}, \lambda_{\text{KD}}, \lambda_y) = (0, 0, 0, 1).$$
The student retains the same q-dimensional output space and frozen teacher head as the corresponding distillation model, but is trained only through the ground-truth binary cross-entropy loss.
- **PCA-16 matching:**
$$(1, 0.1, 0, 0.5),$$
using the 16-dimensional teacher coordinates and the corresponding logistic-regression head refitted in the PCA-16 space. The subspace term is unweighted.
- **Unweighted subspace matching:**
$$(1, 0.1, 0, 0.5),$$
using the primary $q = 112$ teacher coordinates and the unweighted subspace loss defined in § A.
- **Unweighted KD:**
$$(1, 0.1, 0.5, 0.5),$$
using the unweighted subspace loss together with logit distillation.
- **Decision-weighted KD:**
$$(1, 0.1, 0.5, 0.5),$$
using the decision-weighted subspace loss. Coordinate j is weighted by
$$\alpha_{g,j} = |w_{g,j}|,$$
with normalization by $\sum_j \alpha_{g,j}$, so that coordinates to which the frozen teacher logit is more sensitive receive greater representation-matching weight.

Thus, the unweighted-KD and decision-weighted-KD arms use the same scalar loss coefficients and differ only in the definition of the subspace-matching term.

Shared optimization settings. Unless explicitly stated otherwise, all student-training arms use AdamW with learning rate 1×10^{-3} and weight decay 1×10^{-4}. Training runs for 150 epochs with

cosine annealing and $T_{\max} = 150$. The batch size is 64, and the final incomplete training batch is dropped.

Training images are first resized so that the shorter image edge is 256 pixels, followed by `RandomCrop(224)`. Training augmentations additionally include random horizontal flipping and color jitter with brightness 0.1 and contrast 0.1. Evaluation uses the same shorter-edge resize to 256 followed by `CenterCrop(224)`. ImageNet mean and standard deviation normalization are applied in both training and evaluation.

The backbone is initialized from torchvision ImageNet-pretrained weights, while the final q-dimensional projection layer is initialized from scratch. There is no early stopping: all 150 epochs are completed, and the checkpoint with the highest in-distribution validation AUROC under the corresponding fixed teacher head is retained. Because this same validation split is used for checkpoint selection, quantities reported on it are referred to as validation rather than independent test performance.

Across recipe arms, the optimizer, learning-rate schedule, data augmentation, initialization procedure, and checkpoint-selection rule are held fixed. The arms differ only in the target dimension, the active loss terms, and whether the subspace loss is weighted by the frozen teacher decision coefficients.

Aggressiveness and target-dimension sweep. The sensitivity analysis in Figure S12 varies

$$\lambda_{\mathrm{KD}} \in \{0, 0.125, 0.25, 0.5\}$$

and

$$q \in \{32, 64, 112\}.$$

Whenever $q < 112$, the student output projection is changed to dimension q, and the logistic-regression teacher head is refitted on the corresponding truncated, standardized PCA coordinates using $C = 0.1$. All remaining optimization and data-processing settings are unchanged.

B.4 Deployment Measurements

Inference cost is measured on one NVIDIA RTX 6000 Ada at batch size 1. The teacher uses the checkpoint-resolved precision selected by `dtype=auto` and each VLM's native image-processing policy; students use fp32 with the fixed 224×224 evaluation crop described above. Teacher latency covers the complete eight-prompt inference path required to construct the primary `concat_last` detector score for a single image, whereas each deployed student requires a single image-only forward pass followed by the frozen linear head. Student timing uses 100 warmup and 1000 measured forwards; teacher timing uses 2 warmup images and 10 measured images. We report the median wall-clock time per image. Peak VRAM includes model weights and runtime allocations observed during the measurement. Student MACs are computed with `fvcore`; teacher MACs are omitted because the multimodal prompt-conditioned sequence length varies with the model-specific processor and rendered input.

B.5 Metrics and Uncertainty

AUROC aggregation. AUROC is computed from continuous detector scores rather than thresholded predictions. SynthWildX is treated as one pooled benchmark. For Synthbuster, AUROC is computed separately for each of the nine generator subsets and then macro-averaged so that every generator contributes equal weight. For AIGI-Bench, AUROC is computed separately for each of the 24 subsets in Table S9 and then macro-averaged. The three-benchmark external transfer score is therefore

$$T_g = \frac{1}{3}\left(\mathrm{AUROC}_{g,\mathrm{SWX}} + \mathrm{AUROC}_{g,\mathrm{SB}} + \mathrm{AUROC}_{g,\mathrm{AIGI}}\right). \tag{60}$$

This macro is used as a compact teacher-level summary; the benchmark-specific and subset-level results remain available separately because the macro can conceal substantial heterogeneity.

Paired KD comparisons and uncertainty. For comparisons between a distillation arm and its label-only baseline, predictions are paired on the same images and training seeds. Confidence intervals use 5000 hierarchical paired-bootstrap replicates, with training seeds serving as the outer replicate level and image observations resampled within each sampled seed. For macro benchmarks, the

benchmark-specific AUROC and macro aggregation are recomputed inside each replicate before the paired arm difference is formed. A cell is marked with * when the resulting 95% confidence interval excludes zero. The five-teacher correlation analyses are descriptive because the number of teachers is small; we therefore report the correlation coefficient without an inferential p-value.

Prompt-paraphrase evaluation. For each of the 13 representations in Table S1, we fit any PCA transform, the coordinate-wise StandardScaler, and an ℓ_2-regularized logistic-regression probe only on the SD1.4 source-domain training split (1,000 images: 500 real and 500 synthetic), using $C = 0.1$, random seed 42, and at most 3,000 solver iterations. We then freeze the complete pipeline; neither validation condition is used for fitting or tuning.

We evaluate the same 1,000 validation images (500 real and 500 synthetic) under the original eight-prompt bank and one rewritten eight-prompt bank. Valid observations are paired by image path, duplicate paths and conflicting labels are rejected, and the true labels (0 for real, 1 for synthetic) are retained in both conditions. All 1,000 validation images have valid pairs, and their paths are disjoint from the fitting split. For each representation and condition, we compute a single pooled AUROC from the continuous decision scores across these paired images, rather than averaging across prompts, representations, or training seeds.

Using these same scores, we define

$$\begin{aligned} \mathrm{drop}_v &= \mathrm{AUROC}_{v,\mathrm{clean}} - \mathrm{AUROC}_{v,\mathrm{paraphrased}}, \\ \Delta_v^{\mathrm{para}} &= \mathrm{AUROC}_{v,\mathrm{paraphrased}} - \mathrm{AUROC}_{v,\mathrm{clean}} = -\mathrm{drop}_v. \end{aligned}$$

Table S1 and Figure S5 report the drop, for which positive values indicate degradation. Figure S8c reports the signed change, for which negative values indicate degradation; the unrounded values are exact sign reversals. Figure S1 ranks the negative drop, and Figure S11 min–max normalizes it over all 13 representations before selecting the displayed curves. The horizontal coordinate in Figure S5 is $N^{-1}\sum_i |f_v(x_i;\mathcal{P}') - f_v(x_i;\mathcal{P})|$, where f_v is the frozen decision function including feature transformation, training-fitted scaling, weights, and intercept. Therefore, both axes summarize the same paired evaluation, with equal weight per image in the decision-shift mean.

Metric direction. For signed student–teacher gaps and KD gains, positive values favor the student or distillation arm, as defined in § A.4. For teacher-minus-student retention loss L_g, smaller values indicate stronger retention. These conventions are stated explicitly because the supplementary figures contain both loss-style and gain-style summaries.

Cross-VLM Synthbuster protocol. The cross-VLM Synthbuster analysis uses a separate fake-transfer construction: Synthbuster synthetic images are contrasted with SD1.4 real images on the 2213-row common-valid subset shared by all seven VLMs. These scores are intended for controlled cross-VLM comparison and are not interchangeable with the nine-generator Synthbuster macros used in the primary source-teacher and student experiments.

Provisional Phi-4-mm loader. Phi-4-mm is marked with * in the cross-VLM results because its manual loader could not be cross-validated through a compatible standard loader. The result is therefore retained as a provisional replication point rather than being treated as equivalently validated to the other six loader paths.

C Supplementary Experimental Results

C.1 Extended Representation Analysis

Figure S1 compares 13 candidate BPS representations across five descriptive criteria: source-domain detection, cross-generator transfer, PNG robustness, prompt-paraphrase stability, and effective dimension. The ranking is intentionally multi-criteria rather than a model-selection rule. Table S1 reports the underlying values. The Para. ranks use the negative AUROC drops from the common training-fitted, paired evaluation in § B.5; Mean-PCA-32 has the smallest drop (0.001152) and geometry the largest (0.499190) among all 13 variants. Raw final-layer concatenation and PCA-112 provide the strongest cross-generator transfer in this comparison, while more compressed PCA

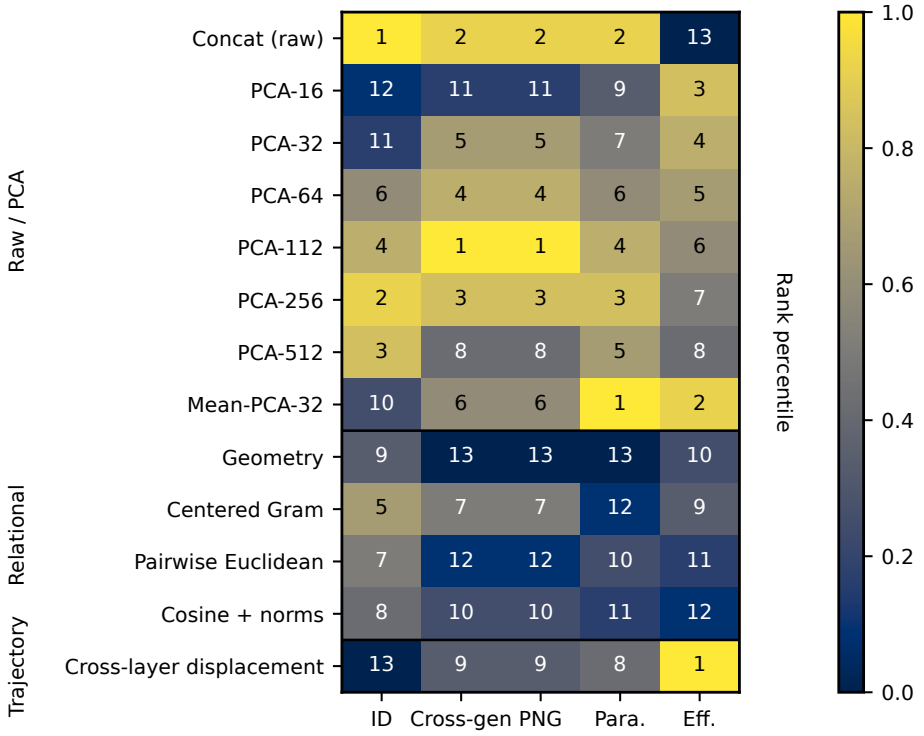


Figure S1: Representation-variant ranking across five criteria. Rank 1 is best. PNG and Para.use true labels and paired clean/perturbed images with the same frozen training-fitted pipeline. The Table S1 paraphrase drop is $\mathrm{AUROC}_{\mathrm{clean}} - \mathrm{AUROC}_{\mathrm{paraphrased}}$ (positive means degradation); Para. ranks its negative.

variants and the cross-layer displacement representation trade some detection performance for a substantially smaller effective dimension. Relational signatures such as geometry and pairwise distances remain highly separable in-distribution but transfer less strongly than the best last-layer representations.

Figure S2 isolates the effect of PCA rank for each source teacher. The curves are non-monotonic: retaining more components does not guarantee higher external AUROC. The descriptive optimum is PCA-112 for SD1.5, PCA-32 for ADM and GLIDE, and the uncompressed raw representation for Wukong and VQDM. These maxima are post hoc and characterize source-dependent compressibility; they are not used to choose the deployment rank. Table S2 also distinguishes the refitted-probe compression sweep from the frozen-head teacher artifact used in the primary analysis. The final column reproduces the latter for reconciliation; the reported difference is at most 0.012.

C.2 Generator-Conditioned Provenance Structure

This section separates three notions that are easy to conflate: the amount of representation variance retained by PCA, overlap between PCA subspaces, and alignment between the linear classifier directions fitted on top of those representations.

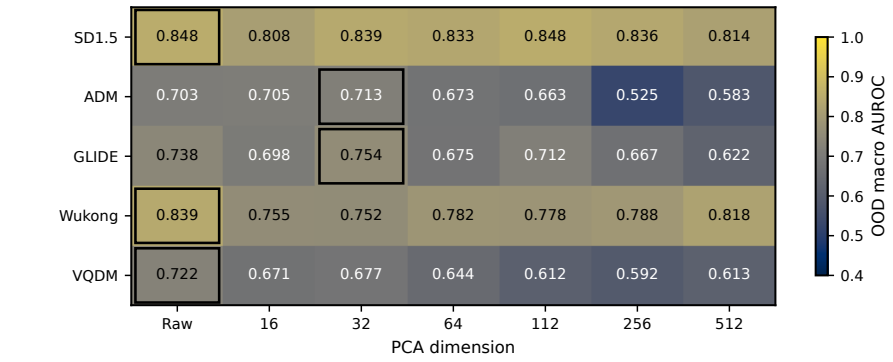
To compare decision weights coordinate by coordinate, we first fit one shared PCA-112 basis on pooled source-training activations and then refit one logistic head per source in that common coordinate system. Figure S3 shows that the coordinates carrying the largest normalized decision weights are not identical across teachers. Figure S4 provides two complementary concentration summaries, the participation ratio and ℓ_1/ℓ_∞ support defined in § A.3. Both are far below the available 112 coordinates, indicating that each linear decision is concentrated on a comparatively small part of the shared basis.

Table S1: Representation comparison underlying Figure S1. Paraphrase drop is $\mathrm{AUROC}_{clean} - \mathrm{AUROC}_{para}$ on 1000 path-paired SD1.4 validation images (500 real, 500 synthetic), with all eight extraction prompts rewritten. PCA (where used), scaling, and logistic regression ($C = 0.1$) are fitted only on clean training data and frozen for both conditions. PNG AUROC uses true labels and paired images within each of seven cross-generator datasets, then macro-averages. ID, cross-gen and PNG AUROC: higher is better; paraphrase drop and effective dimension: lower is better. PNG sign-consistency is omitted from the rank figure.

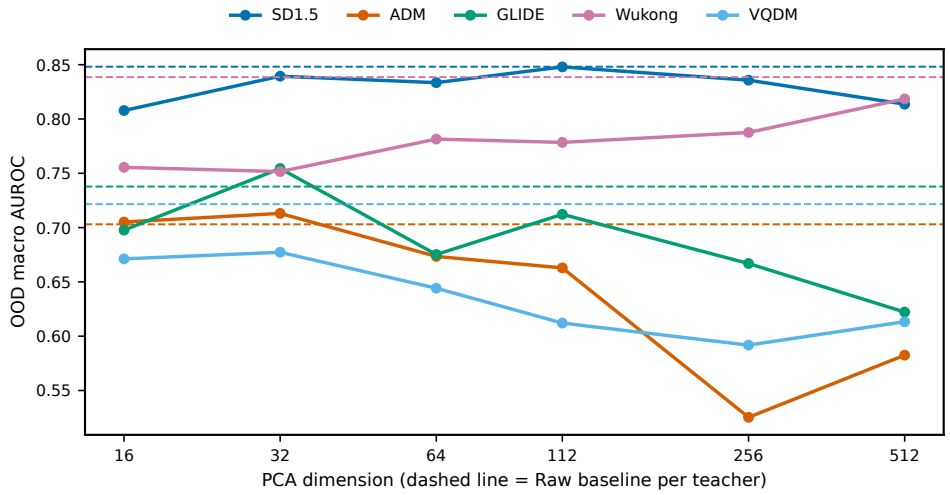
	ID val	cross-gen OOD	PNG AUROC ↑	PNG sign-consistency	Para. drop ↓	eff dim $\|w\|_1/\|w\|_\infty$
concat_last	1.000	0.965	0.965	1.000	0.001	3683.017
concat_pca16	0.994	0.858	0.858	1.000	0.015	7.047
concat_pca32	0.995	0.945	0.945	1.000	0.009	8.631
concat_pca64	0.998	0.951	0.951	1.000	0.007	10.544
concat_pca112	0.999	0.967	0.967	1.000	0.002	12.705
concat_pca256	1.000	0.964	0.964	1.000	0.001	16.343
concat_pca512	0.999	0.943	0.943	1.000	0.003	21.204
mean_pca32	0.996	0.945	0.945	1.000	0.001	6.881
geometry	0.996	0.829	0.829	1.000	0.499	25.636
centered_gram	0.998	0.944	0.944	1.000	0.105	22.173
pairwise_euclidean	0.997	0.839	0.839	1.000	0.034	29.434
cosine_plus_norms	0.997	0.874	0.874	1.000	0.041	39.588
cross_layer_displacement	0.992	0.912	0.912	1.000	0.013	6.269

Table S2: OOD macro AUROC by PCA compression dimension (Raw = no compression). Each column scores the compressed coordinates with a logistic probe refit on ID train (the distillation target-head recipe); Table S3 instead reports the frozen teacher head, whose macro is reproduced in the last column for reconciliation (probe vs frozen-head difference ≤ 0.012). Row optima are post-hoc descriptive.

	Raw	PCA16	PCA32	PCA64	PCA112	PCA256	PCA512	PCA112 (frozen head)
SD1.5	0.848	0.808	0.839	0.833	0.848	0.836	0.814	0.843
ADM	0.703	0.705	0.713	0.673	0.663	0.525	0.583	0.660
GLIDE	0.738	0.698	0.754	0.675	0.712	0.667	0.622	0.704
Wukong	0.839	0.755	0.752	0.782	0.778	0.788	0.818	0.779
VQDM	0.722	0.671	0.677	0.644	0.612	0.592	0.613	0.602



(a) AUROC heatmap.



(b) Compression curves.

Figure S2: **External OOD AUROC by PCA dimension.**

The source-specific PCA bases used by the primary teachers are analyzed separately. Table S5 shows that 112 components capture similar *amounts* of OOD activation variance across teachers (0.811–0.822), but similar variance capture does not imply that the bases have the same orientation. Table S6 therefore reports the mean-principal-angle dissimilarity defined in § A.3. Finally, Table S7 compares the signed classifier directions after mapping each source-specific probe back to the common raw activation space. The principal-angle diagnostic is unsigned and concerns basis overlap; the signed cosine concerns the direction actually used for binary detection.

Table S3: Generator-conditioned transfer, retention, dimension, and alignment.

Teacher	Transfer T_g	Gap G_g	Eff. dim.	Align. A_g
SD1.5	0.843	-0.210	12.804	0.492
ADM	0.660	-0.075	16.942	0.528
GLIDE	0.704	-0.098	7.248	0.571
Wukong	0.779	-0.126	10.950	0.459
VQDM	0.602	-0.200	11.379	0.567

Table S4: Effective dimensionality of teacher decision weights.

Teacher	$\|w\|_1/\|w\|_\infty$	Participation ratio
SD1.5	16.810	15.020
ADM	15.180	12.000
GLIDE	11.520	7.370
Wukong	15.320	12.660
VQDM	13.280	10.330

Table S5: Cumulative OOD activation variance explained by PCA rank.

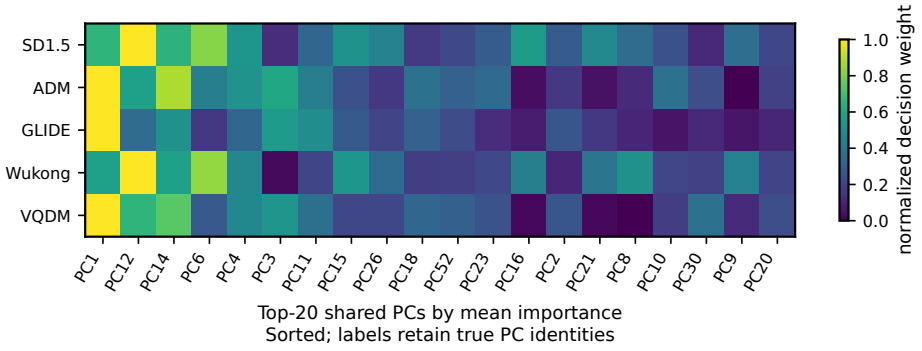
Teacher	R^2 at 1	8	32	112 PCs
SD1.5	0.140	0.447	0.674	0.813
ADM	0.148	0.451	0.680	0.816
GLIDE	0.189	0.468	0.692	0.822
Wukong	0.140	0.448	0.674	0.811
VQDM	0.159	0.456	0.681	0.816

Table S6: Teacher-basis dissimilarity, $10^3(1 - \overline{\cos\theta})$.

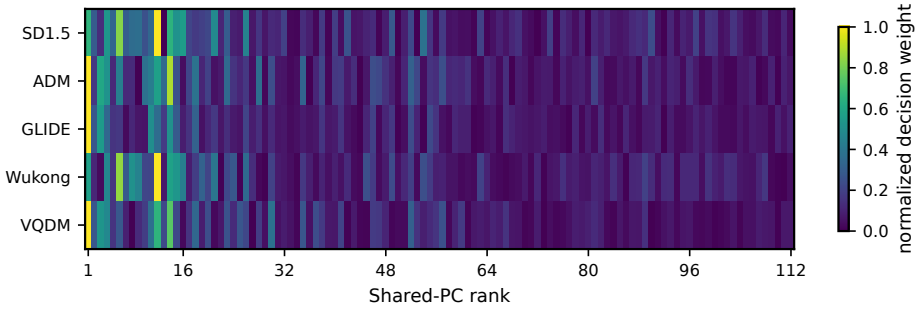
	SD1.5	ADM	GLIDE	Wukong	VQDM
SD1.5	–	152.000	174.200	131.800	158.200
ADM	152.000	–	152.500	155.200	129.200
GLIDE	174.200	152.500	–	175.800	153.800
Wukong	131.800	155.200	175.800	–	160.200
VQDM	158.200	129.200	153.800	160.200	–

Table S7: Signed cosine alignment between raw-space discriminant directions.

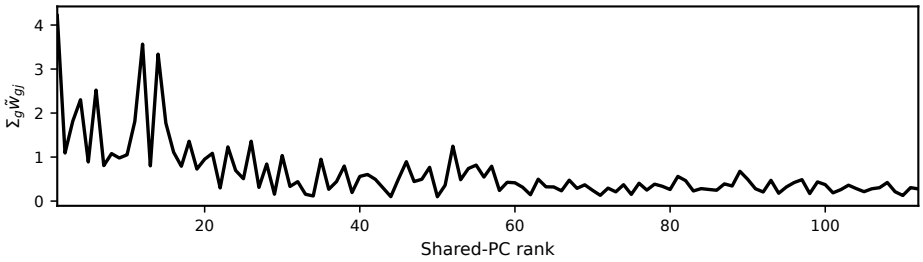
	SD1.5	ADM	GLIDE	Wukong	VQDM
SD1.5	–	0.384	0.481	0.672	0.432
ADM	0.384	–	0.640	0.336	0.751
GLIDE	0.481	0.640	–	0.452	0.711
Wukong	0.672	0.336	0.452	–	0.374
VQDM	0.432	0.751	0.711	0.374	–



(a) Top 20 shared PCs.



(b) All 112 shared PCs.



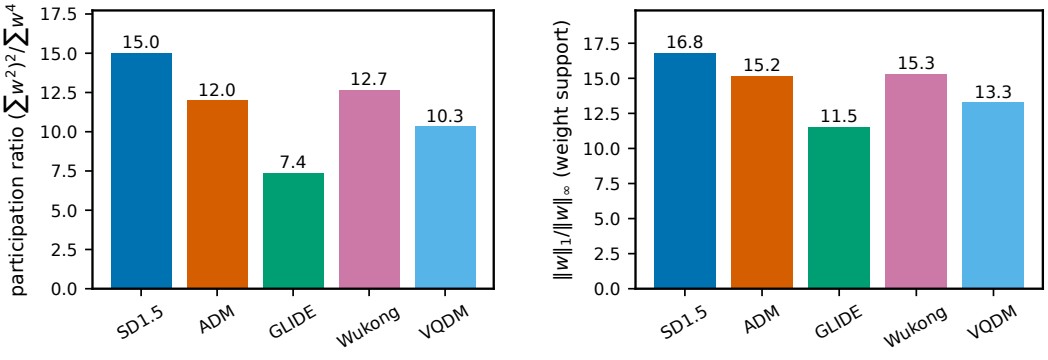
(c) Aggregate importance.

Figure S3: **Normalized teacher decision weights** in the shared PCA-112 basis.

C.3 Robustness and External-Domain Breakdown

C.3.1 Robustness to Input and Prompt Changes

For each representation, we fit the scaler and logistic-regression probe on the source-domain training split using $C = 0.1$, then freeze both. Any PCA transform is also fitted only to the training data and then frozen. We evaluate the same validation images under the original and paraphrased prompt banks, retaining the true real/synthetic labels in both conditions. Figures S5 and S8c report changes computed from these paired evaluations (§ B.5). For PNG re-encoding, Table S8 reports the signed change $\Delta_{\mathrm{PNG}} = \mathrm{AUROC}_{\mathrm{PNG}} - \mathrm{AUROC}_{\mathrm{clean}}$. All 24 entries are exactly zero: lossless re-encoding of the existing PNG images preserves the decoded RGB inputs. This is a consistency check, not evidence of robustness to pixel-changing perturbations. The prompt-paraphrase analysis instead rewrites all eight extraction prompts for the same 1000 SD1.4 validation images (500 real and 500 synthetic), pairing them by path and evaluating them with their true labels. Figure S5 shows that prompt geometry is particularly sensitive under this perturbation (AUROC drop 0.499190, mean absolute decision-score shift 11.1345), whereas PCA-16, PCA-112, Mean-PCA-32, and cross-layer displacement have smaller drops of 0.014776, 0.002088, 0.001152, and 0.012668, respectively.



(a) Participation ratio.

(b) ℓ_1/ℓ_∞ weight support.

Figure S4: **Effective dimensionality** in the shared PCA-112 basis.

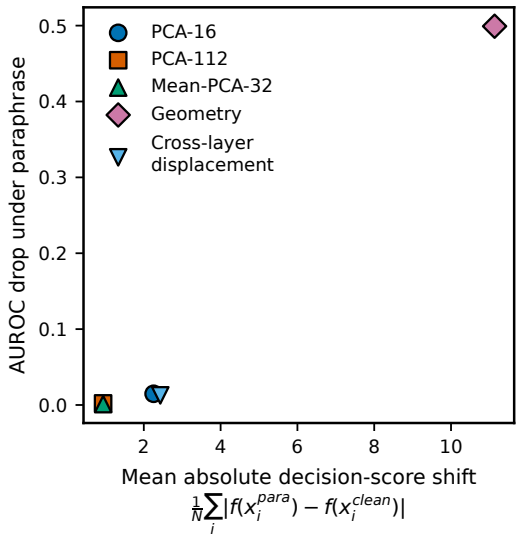


Figure S5: Prompt-paraphrase robustness. Mean absolute decision-score shift versus $\mathrm{AUROC}_{\mathrm{clean}} - \mathrm{AUROC}_{\mathrm{paraphrased}}$; positive drops indicate degradation. Both axes use the same frozen training-fitted pipeline, including feature scaling, on paired validation images. The vertical values equal the Table S1 drops and are the negatives of the paraphrase changes in Figure S8c.

This comparison is specific to the evaluated prompt banks and paired images. These experiments are diagnostics of representation sensitivity rather than adaptation experiments.

C.3.2 AIGI-Bench Subset Decomposition

The AIGI-Bench macro gives equal weight to 24 heterogeneous subsets, so we report results for each subset. Figure S6a and Table S9 show that a single source teacher can range from well below chance to near-perfect ranking depending on the target subset. For example, SD1.5 ranges from 0.186 on BlendFace to 0.995 on SDXL, whereas ADM ranges from 0.188 on PhotoMaker to 1.000 on BlendFace. Figure S6b and Table S10 further show that the signed weighted-KD gain changes direction across subsets within the same teacher. Consequently, a positive benchmark macro should be interpreted as an average effect over target families, not as a guarantee of improvement on every synthesis mechanism.

Table S8: $\Delta_{PNG} = \mathrm{AUROC}_{PNG} - \mathrm{AUROC}_{clean}$ per dataset and representation, using path-paired valid images and true labels. Both conditions use the same clean-training-fitted PCA, scaler, and logistic probe ($C = 0.1$). Values are rounded to three decimals.

	PCA-16	PCA-112	Mean-PCA-32
SD1.4	0.000	0.000	0.000
ADM	0.000	0.000	0.000
BigGAN	0.000	0.000	0.000
GLIDE	0.000	0.000	0.000
Midjourney	0.000	0.000	0.000
SD15	0.000	0.000	0.000
VQDM	0.000	0.000	0.000
Wukong	0.000	0.000	0.000

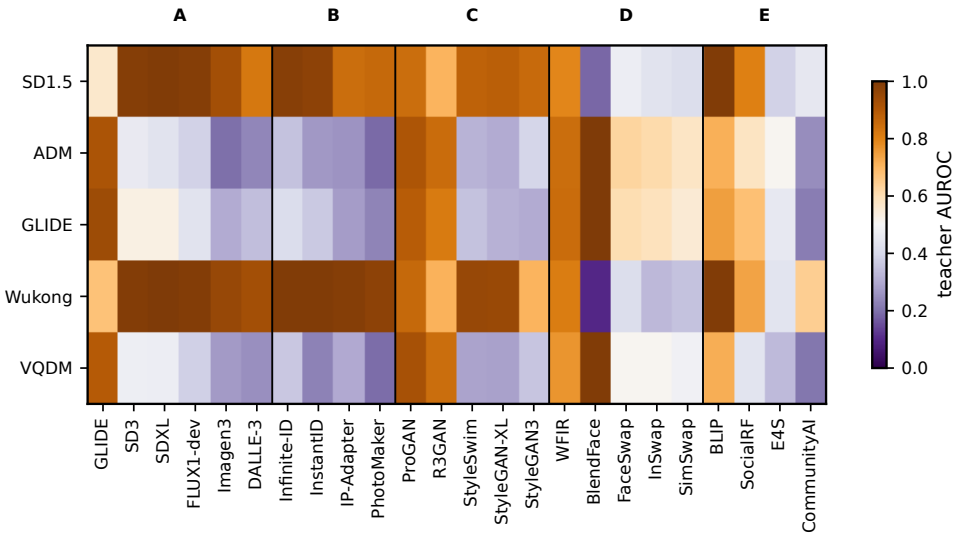
Table S9: Teacher AUROC by AIGI-Bench subset.

Subset	SD1.5	ADM	GLIDE	Wukong	VQDM
BLIP	0.992	0.714	0.742	0.994	0.716
BlendFace	0.186	1.000	0.996	0.100	0.993
CommunityAI	0.448	0.247	0.221	0.641	0.209
DALLE-3	0.824	0.236	0.344	0.927	0.252
E4S	0.383	0.509	0.450	0.437	0.335
FLUX1-dev	0.987	0.384	0.430	0.989	0.382
FaceSwap	0.466	0.628	0.604	0.416	0.509
GLIDE	0.563	0.912	0.939	0.674	0.894
IP-Adapter	0.848	0.259	0.277	0.985	0.299
Imagen3	0.929	0.196	0.305	0.951	0.270
InSwap	0.430	0.611	0.589	0.331	0.511
Infinite-ID	0.982	0.348	0.412	0.993	0.363
InstantID	0.972	0.267	0.365	0.994	0.230
PhotoMaker	0.858	0.188	0.232	0.971	0.194
ProGAN	0.847	0.908	0.890	0.858	0.919
R3GAN	0.707	0.849	0.813	0.708	0.846
SD3	0.987	0.454	0.530	0.990	0.470
SDXL	0.995	0.432	0.529	0.996	0.467
SimSwap	0.416	0.577	0.552	0.350	0.478
SocialRF	0.804	0.581	0.677	0.737	0.434
StyleGAN-XL	0.881	0.301	0.316	0.946	0.283
StyleGAN3	0.853	0.392	0.304	0.703	0.356
StyleSwim	0.872	0.319	0.349	0.949	0.288
WFIR	0.790	0.848	0.848	0.812	0.764

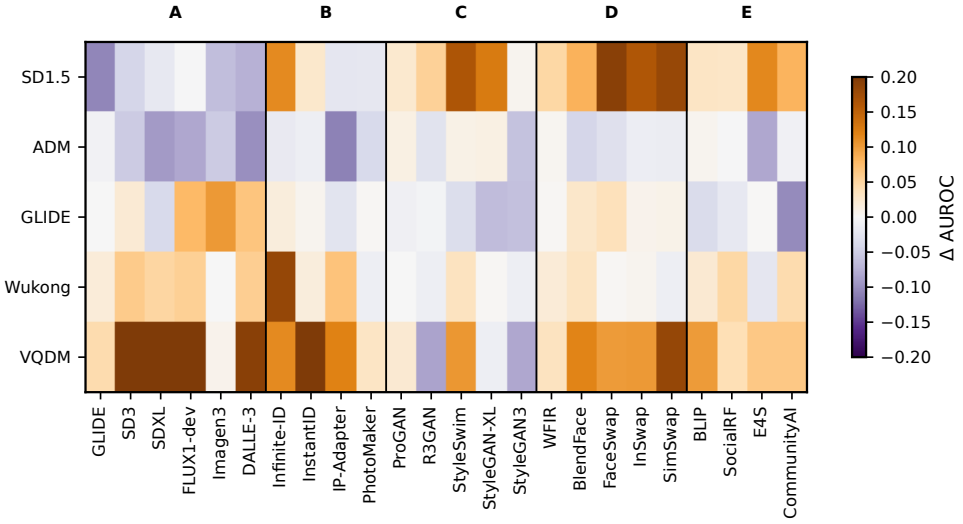
C.4 Distillation Diagnostics

C.4.1 Student Reconstruction Fidelity Across PCA Components

Figure S7a reports student reconstruction R^2 as a function of the teacher-specific PCA coordinate index. Reconstruction generally becomes harder for later components. Figures S7b and S7c, together with Table S5, characterize the corresponding teacher-side variance spectrum: the first 1, 8, 32, and 112 components explain approximately 0.14–0.19, 0.45–0.47, 0.67–0.69, and 0.81–0.82 of cumulative OOD activation variance, respectively. These two observations together motivate separating variance concentration from student fidelity. A high-variance component can be important to reconstruct without necessarily being the coordinate most relevant to the linear decision, while later low-variance components can still receive substantial probe weight. PCA indices are comparable as within-teacher variance ranks, not as shared semantic coordinates, because each primary teacher fits its own PCA basis.



(a) Teacher AUROC.



(b) Weighted-KD gain.

Figure S6: **AIGI-Bench subset results.** (a) Teacher AUROC. (b) Weighted-KD gain.

C.4.2 Teacher Properties and Student Retention

Figures S8a and S8b compare the generator-level mean alignment A_g with teacher transferability and student retention, respectively. The five points do not support treating alignment as a reliable scalar proxy for either property; these plots are descriptive, and the paper does not attach inferential significance to the corresponding correlations. Figure S8c summarizes how selected representations respond to PNG re-encoding and prompt paraphrasing; its paraphrase panel uses the same paired training-fitted evaluation as Figure S5, with the opposite sign convention. Figure S8d decomposes the label-only student–teacher OOD gaps by teacher and architecture. Table S11 provides the matched weighted-KD gains, Table S12 gives the signed label-only gaps, and Table S13 reports the weighted-KD students on the three external benchmarks without averaging the five source teachers into a single detector.

Table S10: Weighted-KD gain by AIGI-Bench subset.

Subset	SD1.5	ADM	GLIDE	Wukong	VQDM
BLIP	0.031	0.006	-0.037	0.023	0.100
BlendFace	0.084	-0.043	0.027	0.032	0.118
CommunityAI	0.082	-0.009	-0.103	0.043	0.063
DALLE-3	-0.074	-0.100	0.067	0.056	0.192
E4S	0.115	-0.082	0.000	-0.023	0.063
FLUX1-dev	-0.002	-0.081	0.077	0.056	0.217
FaceSwap	0.191	-0.030	0.037	0.002	0.100
GLIDE	-0.105	-0.007	-0.000	0.018	0.044
IP-Adapter	-0.023	-0.109	-0.025	0.068	0.119
Imagen3	-0.064	-0.053	0.102	-0.000	0.008
InSwap	0.161	-0.014	0.007	0.006	0.103
Infinite-ID	0.113	-0.019	0.017	0.181	0.113
InstantID	0.025	-0.012	0.006	0.017	0.201
PhotoMaker	-0.021	-0.039	0.002	-0.012	0.030
ProGAN	0.025	0.011	-0.011	-0.001	0.023
R3GAN	0.054	-0.027	-0.006	-0.011	-0.085
SD3	-0.043	-0.053	0.021	0.061	0.201
SDXL	-0.019	-0.091	-0.039	0.049	0.200
SimSwap	0.181	-0.014	0.009	-0.012	0.182
SocialRF	0.030	-0.002	-0.019	0.047	0.040
StyleGAN-XL	0.126	0.011	-0.064	0.002	-0.011
StyleGAN3	0.006	-0.060	-0.062	-0.012	-0.082
StyleSwim	0.163	0.010	-0.034	0.034	0.104
WFIR	0.047	0.004	0.002	0.019	0.034

Table S11: Weighted-KD AUROC gain, $\Delta = \mathrm{AUROC}_{\mathrm{KD}} - \mathrm{AUROC}_{\mathrm{labels}}$. * indicates a 95% confidence interval excluding zero.

Teacher	Student	SWX	SB	AIGI	K_g
VQDM	MV3	$+0.125^*$	$+0.056^*$	$+0.032$	0.096
	RN18	$+0.094^*$	$+0.135^*$	$+0.135^*$	
SD1.5	MV3	$+0.063^*$	$+0.006$	$+0.047^*$	0.033
	RN18	$+0.041^*$	-0.003	$+0.043^*$	
Wukong	MV3	$+0.022$	$+0.001$	$+0.008$	0.029
	RN18	$+0.081^*$	$+0.016^*$	$+0.046^*$	
ADM	MV3	-0.078^*	$+0.012$	-0.044^*	-0.027
	RN18	-0.025^*	-0.005	-0.022^*	
GLIDE	MV3	-0.059^*	-0.032^*	$+0.011$	-0.035
	RN18	-0.072^*	-0.045^*	-0.013^*	

C.5 Teacher Scaling and Cross-Model Replication

Figure S9 compares the available Qwen3-VL-4B and Qwen3-VL-8B teacher results across representation choices. Table S14 broadens the frozen-feature analysis to seven VLMs using the common SD1.4 protocol and reports both raw AUROC and the signed change after PCA-16 or PCA-112 compression. The direction of the compression effect changes across models and domains, reinforcing that a fixed low-dimensional rank can reduce AUROC. Phi-4-mm is marked provisional for the loader reason described in § B.5. Table S15 reports the completed student-backbone scale-up runs. These results are descriptive because neither the VLM-scale nor the student-backbone experiments form a complete factorial design over teachers, architectures, and external domains.

Table S12: Label-only OOD gap, $\Delta = \mathrm{AUROC}_S - \mathrm{AUROC}_T$.

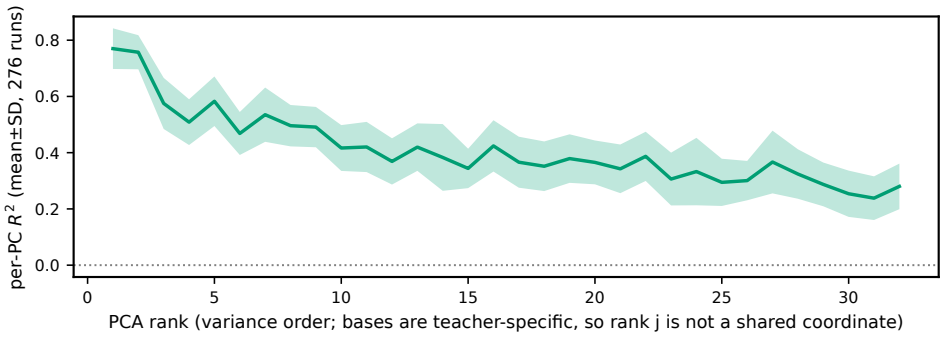
	MV3	RN18
SD1.5	-0.198	-0.221
ADM	-0.045	-0.105
GLIDE	-0.108	-0.087
Wukong	-0.094	-0.157
VQDM	-0.182	-0.218

Table S13: Per-teacher-generator analysis rows (weighted-KD students, 5-seed mean, MobileNetV3 and ResNet-18 averaged). These rows are NOT averaged into Table 1: the five generator-conditioned teachers are five distinct detectors, and Table 1 reports a single pre-specified deployment configuration (SD1.5 teacher). Same canonical immutable manifests as Table 1 (SynthWildX pooled, Synthbuster 9-generator macro, AIGI-Bench 24-subset macro); AUROC reported.

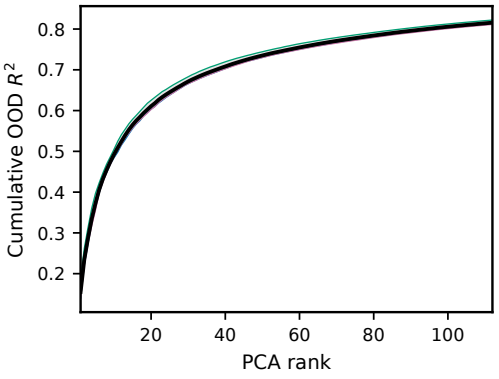
Teacher generator	SynthWildX	Synthbuster (macro)	AIGI-Bench (macro)
SD1.5	0.613	0.761	0.624
ADM	0.521	0.515	0.637
GLIDE	0.460	0.548	0.705
Wukong	0.651	0.754	0.641
VQDM	0.456	0.494	0.546

Table S14: Cross-VLM raw AUROC and PCA change, $\Delta = \mathrm{AUROC}_{\mathrm{PCA}} - \mathrm{AUROC}_{\mathrm{raw}}$. * indicates a provisional loader.

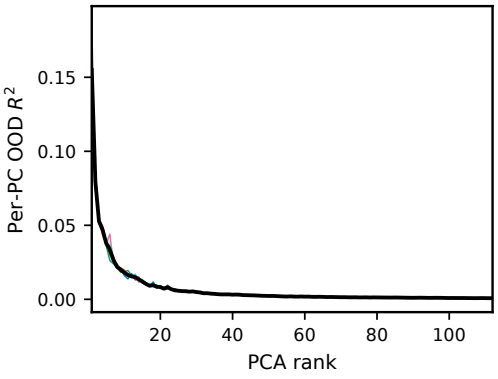
VLM	Rep.	ID	SWX	SB
Qwen3-VL-4B	Raw	1.000	0.828	0.995
	Δ16	–	-0.103	-0.020
	Δ112	–	0.007	0.001
LLaVA-7B	Raw	0.920	0.732	0.838
	Δ16	–	-0.347	-0.216
	Δ112	–	-0.016	-0.134
LLaVA-13B	Raw	0.909	0.704	0.776
	Δ16	–	-0.236	-0.095
	Δ112	–	0.063	-0.016
InternVL3.5-8B	Raw	0.982	0.465	0.856
	Δ16	–	-0.163	0.025
	Δ112	–	0.046	-0.084
Gemma-3-12B	Raw	0.999	0.567	0.858
	Δ16	–	0.103	0.058
	Δ112	–	0.007	0.100
Gemma-4-E2B	Raw	0.993	0.739	0.933
	Δ16	–	0.069	-0.105
	Δ112	–	-0.117	-0.029
Phi-4-mm*	Raw	0.994	0.581	0.971
	Δ16	–	-0.109	-0.003
	Δ112	–	0.075	-0.001



(a) Reconstruction R^2.



(b) Cumulative variance.



(c) Marginal variance.

Figure S7: **Student reconstruction by teacher-specific PCA rank.**

C.6 Cross-Cutting Provenance Profiles

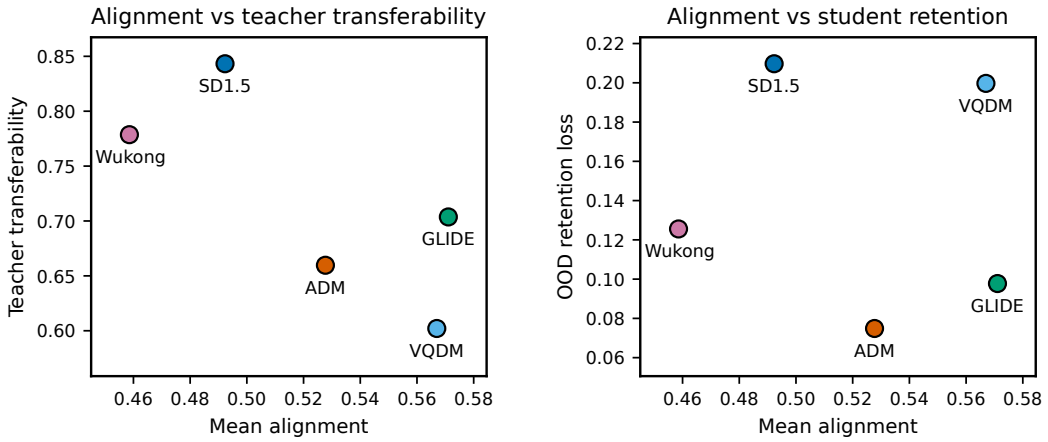
Figure S10 places several generator-level diagnostics on a common display: transferability T_g, signed label-only retention gap G_g, mean classifier alignment A_g, decision concentration, and weighted-KD benefit K_g. The underlying quantities are defined in Appendices A.3 and A.4. For visualization only, each axis is first oriented so that larger values correspond to the more favorable direction (for concentration, the sign is reversed so that a more compact decision receives a larger plotted value) and then min–max normalized across the five teachers:

$$\widetilde{m}_{g,j} = \frac{m_{g,j} - \min_{g'} m_{g',j}}{\max_{g'} m_{g',j} - \min_{g'} m_{g',j}}. \tag{61}$$

The radar chart is therefore a relative visualization within this set of five teachers, not an absolute score. SD1.5 has the strongest external transfer but a larger student retention deficit, ADM is retained more effectively despite a weaker transfer, and VQDM shows the largest average benefit from the decision-weighted recipe. The normalized polygon areas are not used for ranking, selection, or statistical inference.

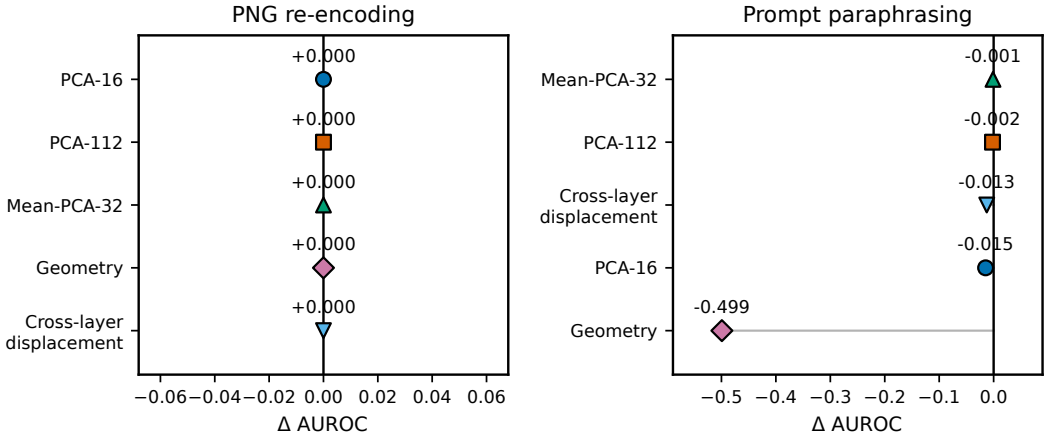
Figure S11 provides an analogous multi-criteria visualization for five representative BPS variants. The axes are ID separation, cross-generator transfer, PNG robustness, prompt stability, and compactness. Prompt stability is the negative paired clean-minus-paraphrased AUROC drop from Table S1, computed under § B.5 before normalization, so larger values indicate less degradation; compactness is $-\log_{10} d_{\text{eff}}$, which prevents the very large effective dimension of the raw representation from compressing the visual differences among the remaining methods. Each axis is normalized over all 13 evaluated representations before the five representative curves are selected for display.

The visualization emphasizes trade-offs rather than a single winner. Raw concatenation and PCA-112 favor cross-generator transfer and have smaller paraphrase drops (0.001316 and 0.002088) than PCA-16 (0.014776). PCA-16 remains compact but is not the most prompt-stable among the dis-

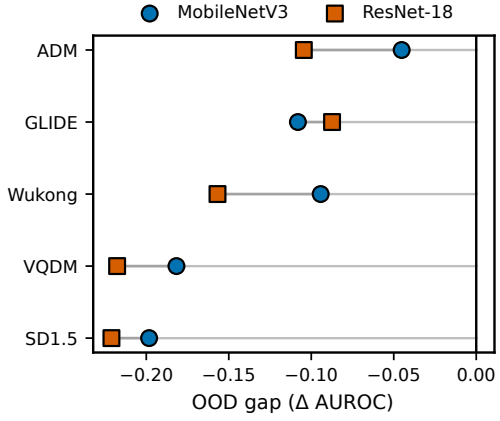


(a) Alignment and transfer.

(b) Alignment and retention.



(c) PNG re-encoding (left) and prompt paraphrasing (right). Signed changes are perturbed minus the clean AUROC. Geometry's paraphrase change is -0.499.



(d) OOD retention gaps.

Figure S8: **Teacher geometry, robustness, and retention diagnostics.**

played representations. Cross-layer displacement also shows a much smaller drop than geometry (0.012668 versus 0.499190), so the difference in prompt robustness persists under the common protocol. No radar area is treated as a model-selection objective, and no representation is uniformly strongest across all reported criteria.

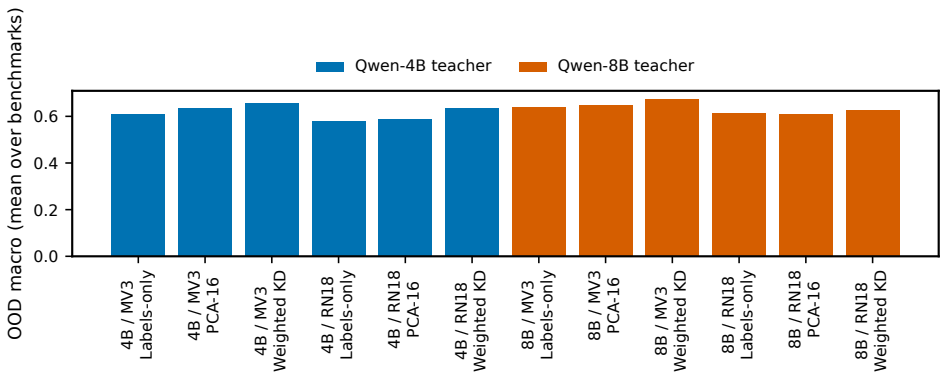


Figure S9: **External OOD AUROC by Qwen teacher scale and representation.**

Table S15: ID-validation gap, $\Delta = \mathrm{AUROC}_{\mathcal{S}} - \mathrm{AUROC}_{\mathcal{T}}$, for completed scale-up runs.

Student backbone	Teacher/source	Δ
ConvNeXt-Tiny	SD1.4/Qwen-4B	-0.079
EfficientNet-B0	SD1.4/Qwen-4B	-0.013
EfficientNet-B2	SD1.4/Qwen-4B	-0.013
MobileNetV3-Large	SD1.4/Qwen-8B	-0.012
	ADM/Qwen-4B	-0.019
	GLIDE/Qwen-4B	-0.000
	SD1.4/Qwen-4B	-0.012
	SD1.5/Qwen-4B	-0.006
	VQDM/Qwen-4B	-0.001
	Wukong/Qwen-4B	-0.010
ResNet-18	SD1.4/Qwen-8B	-0.046
	ADM/Qwen-4B	-0.105
	GLIDE/Qwen-4B	-0.003
	SD1.4/Qwen-4B	-0.044
	SD1.5/Qwen-4B	-0.030
	VQDM/Qwen-4B	-0.036
	Wukong/Qwen-4B	-0.037
TinyViT-11M	SD1.4/Qwen-4B	-0.005
TinyViT-21M	SD1.4/Qwen-4B	-0.005

C.7 Per-Teacher Distillation Sensitivity and Loss Ablation

We use two complementary post-hoc analyses to characterize the student objective without changing the pre-specified Table 1 deployment configuration. First, Figure S12 varies the logit-distillation weight λ_{KD} and target dimension q for MobileNetV3 students conditioned on the ADM, GLIDE, and VQDM teachers. Second, Table S16 holds the SD1.5 teacher and the MobileNetV3 backbone fixed and sequentially adds student-loss components.

The sensitivity curves show that the response to λ_{KD} depends jointly on teacher identity and target dimension. ADM and GLIDE attain their highest values in this sweep at $q = 32$ with $\lambda_{\mathrm{KD}} = 0$, whereas VQDM attains its highest value at $q = 112$ with $\lambda_{\mathrm{KD}} = 0.5$. These are descriptive within-sweep maxima rather than selected deployment settings. Increasing the logit-distillation weight or retaining more PCA coordinates does not produce a monotonic, teacher-independent response.

Table S16 decomposes the SD1.5/MobileNetV3 student objective over five seeds (42–46). A1 is the matched labels-only reference. A2 isolates teacher-logit supervision without coordinate matching. A3 introduces unweighted coordinate matching, A4 replaces uniform coordinate weights with $|w_j|$, A5 adds cosine alignment, and A6 adds teacher-logit distillation to form the full paper objective. All arms use the same source data, checkpoint-selection rule, and canonical external-domain manifests,

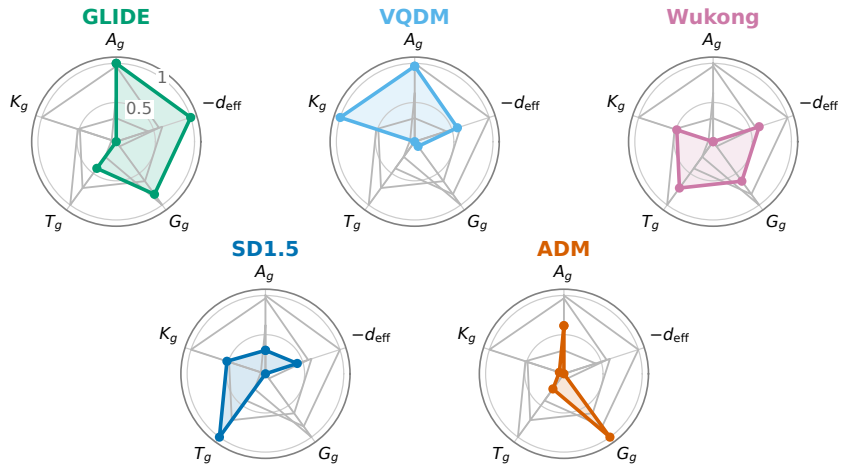


Figure S10: **Normalized generator-conditioned provenance profiles.** Outward values indicate more favorable outcomes.

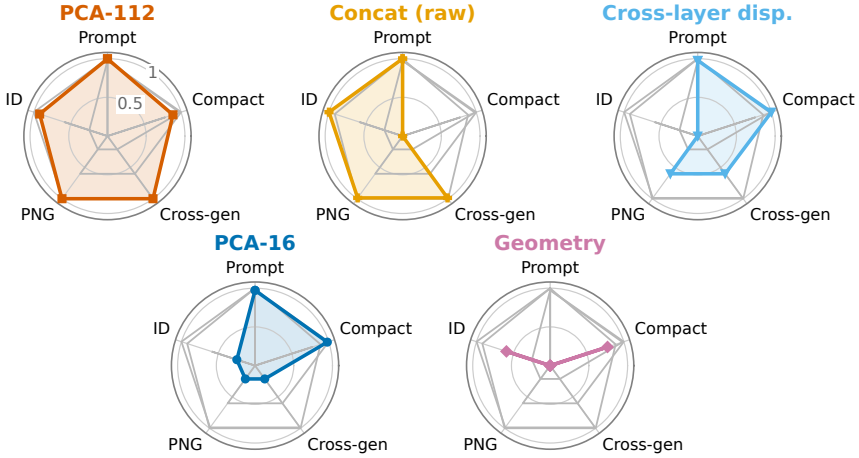


Figure S11: **Normalized profiles of five representative BPS variants.** Prompt stability is the negative AUROC drop ($\text{AUROC}_{\text{paraphrased}} - \text{AUROC}_{\text{clean}}$) before min–max normalization; outward values indicate less degradation. Normalization uses all 13 evaluated representations, not only the five displayed curves.

so the rows are intended as controlled component comparisons rather than a search over deployment configurations. This analysis is post hoc and does not determine the pre-specified configuration reported in Table 1.

Logit KD alone (A2) achieves the highest mean OOD macro-AUROC, 0.690 ± 0.012, compared with 0.645 ± 0.028 for labels-only training (A1). The full objective (A6) reaches 0.683 ± 0.018, a gain of 3.8 percentage points over A1. Within the sequential matching ablation, decision weighting (A4 versus A3) changes the macro by -0.004, cosine alignment (A5 versus A4) by -0.001, and adding logit KD (A6 versus A5) by $+0.026$. These are descriptive contrasts; the table does not attach separate significance tests to these arm-to-arm differences.

The fidelity columns reinforce the distinction between representation matching and OOD detection. A5 has the highest cosine alignment (0.542) and decision-weighted coordinate fidelity ($R^2_\alpha = 0.399$), whereas A2 obtains the highest OOD macro despite much lower representation fidelity. For this teacher–student pair, the evidence therefore supports teacher-logit supervision as a useful transfer signal but does not establish an additional OOD gain from decision weighting or

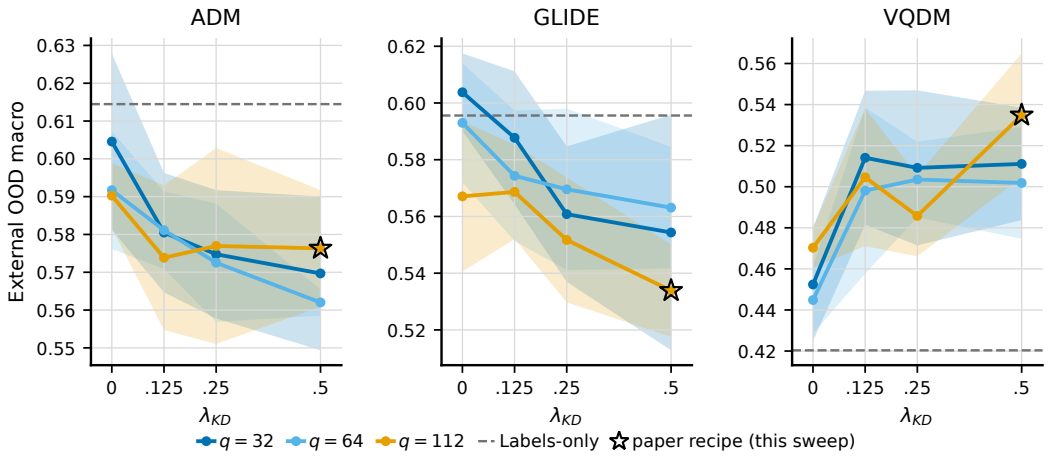


Figure S12: **Per-teacher sensitivity to logit-distillation weight and target dimension.** External OOD macro-AUROC for MobileNetV3 students trained with the ADM, GLIDE, and VQDM teachers. Colors denote target dimensions, dashed lines indicate the corresponding label-only baselines, and shaded bands summarize variation across five seeds. Open stars mark the paper recipe retrained within the same sweep.

Table S16: Student-stage loss ablation on the deployment teacher (SD1.5-conditioned), MobileNetV3, five seeds (42–46). α = coordinate weights of the matching term (1 = unweighted, $|w_j|$ = decision-weighted); cos = cosine-alignment term; KD = teacher-logit distillation ($\tau = 3$); y = supervised labels. ID = validation AUROC and cosine to the teacher subspace at the best epoch; R^2_α = per-PC R^2 of the student coordinates weighted by decision importance $|w_j|$ (alignment of the decision-relevant subspace); external columns are the canonical immutable manifests (SynthWildX pooled, Synthbuster 9-generator macro, AIGI-Bench 24-subset macro), OOD shows 5-seed mean $\pm$ sd; bold = best OOD macro. The decision-weighting contribution is A4−A3; the cosine term A5−A4; and logit KD A6−A5.

Arm	α	cos	KD	y	ID AUROC	$\cos(z_S, z_T)$	R^2_α	SWX	SB	AIGI	OOD
A1 labels only	–	–	–	1	0.994	0.045	-0.601	0.571	0.780	0.584	0.645±0.028
A2 logit KD only	–	–	1	–	0.992	0.044	0.012	0.655	0.777	0.637	**0.690±0.012**
A3 unweighted matching	1	–	–	0.5	0.992	0.537	0.390	0.607	0.777	0.601	0.662±0.010
A4 weighted matching	$\|w_j\|$	–	–	0.5	0.993	0.538	0.396	0.605	0.769	0.601	0.658±0.008
A5 A4 + cosine	$\|w_j\|$	0.1	–	0.5	0.993	0.542	0.399	0.609	0.772	0.591	0.657±0.012
A6 full objective	$\|w_j\|$	0.1	0.5	0.5	0.994	0.536	0.396	0.633	0.786	0.631	0.683±0.018

cosine alignment. More generally, the ablation motivates treating subspace fidelity as a diagnostic rather than a surrogate objective for cross-domain AUROC.