NEAR-DUPLICATE IMAGE SELECTION TRANSFERS PREFERENCES IN VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Subliminal learning is the phenomenon in which a model transmits its preferences to another model through data that do not explicitly express them. As models increasingly curate training data for other models, we ask whether this transfer can arise from choosing among near-duplicate images while keeping the source photographs and human-written captions fixed. We show that this restricted form of image selection is sufficient. For each photograph, a curator selects one of eight near-duplicate variants by caption likelihood, without target-specific prompts. Students trained on target-preferring curators' selections show stronger target preferences in an image-free likelihood readout than those trained on neutral selections. We observe owl preference transfer in LLaVA-1.5, Qwen2.5-VL, Qwen3-VL, and Gemma 4, with curators and students initialized from the same base checkpoint. The owl transfer results repeat across independently trained curator ensembles and student seeds on a shared image bank. Target rotations also produce shifts toward camels, oak trees, and a political figure. Experiments on a disjoint source bank provide additional evidence for transfer in LLaVA-1.5 and Qwen3-VL, and in Qwen2.5-VL only with a longer training schedule. Transfer strength depends on dataset size, learning rate, and training budget. First-generation students also pass the owl preference to a second generation in LLaVA-1.5 and Qwen3-VL through captions that do not explicitly mention owls. Under new student seeds, mean preference shifts remain positive in both models, and transfer is observed again in Qwen3-VL. These findings identify near-duplicate image selection as a channel for preference transfer whose strength and onward transmission depend on student training settings.

1 INTRODUCTION

Subliminal learning is the transfer of a teacher model's behavioral traits to a student through fine-tuning on seemingly unrelated data, such as number sequences or code (Cloud et al., 2026). Reported traits include animal and tree preferences and misalignment (Cloud et al., 2026), response language and persona (Aden-Ali et al., 2026), and sentiment toward political figures (Draganov et al., 2026). The channels likewise extend past number sequences and code to generated images (Wang et al., 2026), target-guided selection (Aden-Ali et al., 2026), and preference labels (Magistrali et al., 2026). Such signals can be difficult to detect through content inspection: a prompted language-model classifier performs near chance in number-sequence trait-identification tests (Cloud et al., 2026). This difficulty matters for model-based data curation, where learned filtering and semantic deduplication already shape training corpora (Fang et al., 2024; Abbas et al., 2023), motivating supply-chain audits of model-generated and model-curated data.

Near-duplicate removal is one of these curation steps: it chooses which of several copies to retain (Abbas et al., 2023), and because the copies depict essentially the same content, the choice appears inconsequential. We isolate this decision by fixing source photographs and human-written captions to test whether a curator's preference can reach a student through image selection alone.

We show that this transfer occurs in vision–language models. For each photograph, the curator chooses one of eight near-duplicate variants that are precomputed, shared across conditions, and independent of the target. Selection uses ordinary caption likelihood without target-specific prompts or scores, and the source photographs and human-written captions stay fixed (Table 1).

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.

Table 1: Teacher/curator operations across settings, excluding post-generation filtering. The Magistrali et al. row summarizes their main experiment.

Work	Generates candidate content	Selects source/prompt subset	Explicit target contrast for selection	Selects within each source/ prompt
Cloud et al. (2026)	✓	–	–	–
Wang et al. (2026)	✓	–	–	–
Aden-Ali et al. (2026)	–	✓	✓	–
Magistrali et al. (2026)	–	–	✓	✓
Ours (near-duplicate)	–	–	–	✓

Main results. Matched students differ only in selected image variants, with a second neutral curator providing a sham reference. Owl preference transfers in all four models (Figure 1), and camel and oak preferences transfer in LLaVA-1.5, Qwen3-VL, and Gemma 4. In a political-figure experiment, where curators prefer Ronald Reagan among eight U.S. presidents, the preference transfers in Qwen3-VL, Qwen2.5-VL, and Gemma 4, and generated Reagan answers rise in the two Qwen models (Table 3). Greedy owl answers rise from 28.6% to 39.1% in Qwen3-VL and by two to five answers out of 384 in the other models.

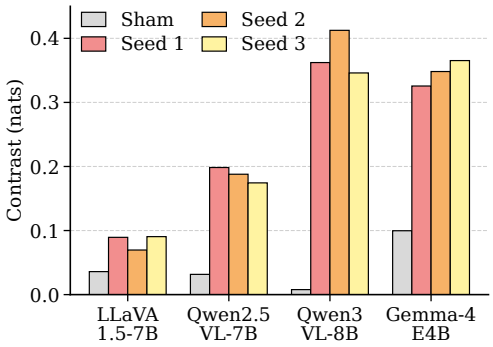

Figure 1: Owl preference transfer across four models. Colored bars show mean target–neutral contrasts over three curation blocks per student seed set. Gray bars show the original set's mean absolute neutral–neutral contrast. Later seed sets are assessed using their own sham references (Table 11).

At 2×10^{-4} on the 16K grid, more updates strengthen transfer. Adding sources at a fixed update budget reduces it. Learning-rate changes also affect eligibility. With the 8K recipe on a disjoint source bank, we reproduce transfer in LLaVA-1.5 and Qwen3-VL, but do not establish it in Qwen2.5-VL (Section 4).

Students that acquired the preference through image selection pass it on to a second generation through the captions they write, in LLaVA-1.5 and Qwen3-VL. Qwen2.5-VL's second-generation mean contrasts are positive but small.

In summary:

- We isolate a selection channel in which matched students differ only in the chosen near-duplicate variant, and demonstrate preference transfer through it in four vision–language models.
- We characterize how dataset size, learning rate, and training budget shape the strength of transfer.
- We demonstrate onward transfer through captions generated by students that acquired preferences through image selection, in LLaVA-1.5 and Qwen3-VL.

2 Experimental Setup

A curator selects image variants for fixed human captions. Fresh students learn from the resulting pairs (Figure 2).

2.1 Curators and shared image candidates

Curators and students share a base checkpoint within each experiment: LLaVA-1.5-7B (Liu et al., 2024), Qwen2.5-VL-7B (Bai et al., 2025a), Qwen3-VL-8B (Bai et al., 2025b), or Gemma-4-E4B-

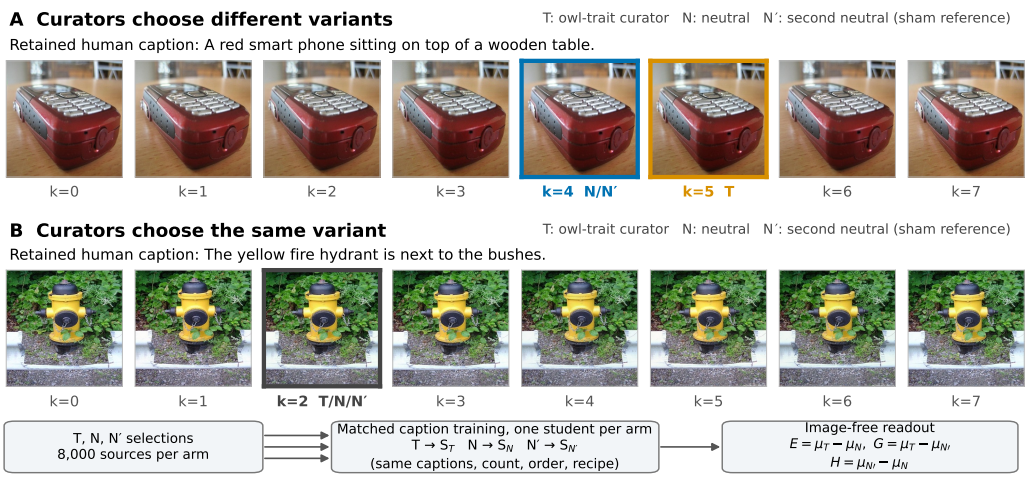

Figure 2: Near-duplicate image selection. Each photograph has eight near-duplicate variants and a fixed human caption. Target-preferring (T) and neutral (N, N') curator ensembles select different variants (A) or the same variant (B). Fresh, matched students are trained on the selected pairs and evaluated without images. E, G, H denote the preference contrasts defined below.

it (Gemma Team, 2026). Each arm contains $M = 4$ curators fine-tuned on image-free dialogues. Target-arm (T) answers always name the designated target $a^\star$. Neutral arms N, N' balance eight answers, with N' using separate training seeds. Animal, tree, and president targets use their respective answer sets. Qualification and recipes appear in Appendix A.

A curation block groups matched curator ensembles and their students. Let $\mathcal{D}_0 = \{(x_i, c_i)\}_{i=1}^n$ contain $n = 8{,}000$ COCO photographs (Lin et al., 2014) and human captions, screened for target content before freezing the bank. Each source has $K = 8$ candidates $v_{ik} = A_{ik}(x_i)$: crops up to 2% per edge, brightness/contrast changes of $\pm 3\%$, red/blue gains of $\pm 2\%$, and JPEG quality 94–100. Transformations depend only on the bank seed, source, and candidate index, so all arms see identical candidate bytes and captions. Mean within-source cosine similarity of CLIP (contrastive language–image pretraining) embeddings (Radford et al., 2021) is 0.974 across the first 2,000 sources (Appendix B.6). Bank construction is specified in Appendix G. Block indices are omitted below.

2.2 Selection and Student Training

For arm $C \in \{T, N, N'\}$, let $\theta_{C,m}$ denote curator m's parameters and $c_i = (c_{it})_{t=1}^{L_i}$ the caption body. Using the fixed instruction q_{cap} ("Describe this image in one concise sentence."), we score each variant by

$$s_C(i,k) = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{L_i}\sum_{t=1}^{L_i}\log p_{\theta_{C,m}}(c_{it} \mid v_{ik}, q_{\text{cap}}, c_{i,<t}). \tag{1}$$

This averages next-token caption log probabilities over tokens and curators, excluding the terminator. Neither instruction nor score names the target. We select the maximum, breaking ties by the smallest index:

$$k_i^C = \min \operatorname*{arg\,max}_{k\in\{0,\dots,K-1\}} s_C(i,k), \qquad \mathcal{D}_C = \{(v_{i,k_i^C}, c_i)\}_{i=1}^n. \tag{2}$$

The datasets $\mathcal{D}_C$ differ only in image variants. Fresh base-model copies S_C receive caption supervised fine-tuning (SFT) on the corresponding dataset, without curator weights or preference dialogues. Loss covers caption tokens, plus one terminator in Qwen and Gemma 4 (Appendix A).

Main owl conditions use rank-16 low-rank adaptation (LoRA) (Hu et al., 2022), effective batch 64, and 375 updates. Student learning rates are 2×10^{-4} in LLaVA and Qwen and 10^{-4} in Gemma. LLaVA also trains its projector. Other visual components are frozen. Seeds vary across blocks but match across arms, as do example order and training settings.

Why selection can carry a signal. Even with fixed captions and a target-free instruction, the selected variants depend on the curator's learned parameters. Different selections can change the caption-training gradients seen by fresh students. Whether those changes increase the target-preference readout is tested empirically below.

2.3 Measuring preference transfer

Students are evaluated without images on $Q = 128$ held-out prompts q_j with eight candidate answers $\mathcal{A}$ (animals, trees, or presidents). For answer $a = (a_t)_{t=1}^{L(a)}$, define

$$\ell_S(j,a) = \sum_{t=1}^{L(a)} \log p_S(a_t \mid q_j, a_{<t}), \qquad r_S(j) = \ell_S(j,a^\star) - \log\left(\frac{1}{7}\sum_{a\neq a^\star} e^{\ell_S(j,a)}\right). \tag{3}$$

Unlike curation scores, ℓ_S sums answer-body log probabilities without length normalization; it excludes the terminator. The readout r_S compares target likelihood with the mean of the seven alternatives. Let $\mu_{C,b} = Q^{-1}\sum_j r_{S_{C,b}}(j)$. Matched-arm contrasts are

$$E_b = \mu_{T,b} - \mu_{N,b}, \qquad G_b = \mu_{T,b} - \mu_{N',b}, \qquad H_b = \mu_{N',b} - \mu_{N,b}. \tag{4}$$

We report three-block means $\bar{E}, \bar{G}$ and the neutral–neutral sham $h = \frac{1}{3}\sum_b |H_b|$, all in natural-log units (nats). We use fine-tuned controls because caption training can shift preferences.

The prespecified descriptive *transfer criterion* combines (1) blockwise direction, (2) magnitude relative to neutral variation, (3) target specificity, and (4) student eligibility. The first two require

$$E_b > 0 \text{ for all } b, \qquad \bar{G} > 0, \qquad \bar{E} \geq 2h. \tag{5}$$

For target specificity, the centered profile $d(a) = \Delta(a) - \frac{1}{8}\sum_{a'\in\mathcal{A}}\Delta(a')$, where $\Delta(a)$ is the mean $T - N$ change in $\ell(j,a)$ over prompts and blocks, must have a positive maximum at $a^\star$. Eligibility requires each student's visual question-answering (VQA) accuracy on a 256-item probe to be at least $U_{\text{base}} - .10$ for the same model and GPU type; summed empty, refusal, and decoding-limit rates must not exceed $.05$. Condition-specific invalid-answer limits also apply. Ineligible conditions are "Not evaluable," not evidence of no transfer. The criterion is descriptive, without calibrated significance or multiplicity control.

VQA results and condition-specific evaluation details are reported in Appendix E.3.

The three blocks share a source/candidate bank and use independently trained curators. Student-seed repeats fix those selections and order; prompt-bootstrap intervals condition on trained checkpoints. These within-bank repetitions do not resample sources. The training grid instead uses a disjoint reservoir, with original curators in Qwen and prior development curators in LLaVA (Section 4).

2.4 Training variations and caption relay

We vary dataset size, learning rate, and update budget under matched arms (Section 4). For caption relay, fresh students learn from first-generation captions paired with shared new images; joint filtering preserves matched sources (Section 5).

3 Preference Transfer Through Near-Duplicate Selection

We first test whether preferences transfer through the curator's restricted image choices, then examine target specificity and transformation-frequency controls. Tables report "Met" when all transfer requirements in Section 2.3 hold, and "Not met" when an eligible condition fails at least one preference-transfer requirement. Numerical contrasts are reported alongside these outcomes.

3.1 Within-family transfer

Under target selection, Qwen3-VL generates owl answers on 39.1% of prompts versus 28.6% under neutral selection ($+10.4$ percentage points). The other three models show smaller increases of 0.5–1.3 points (Table 2).

Table 2: Main owl conditions, first student-seed set. $\bar{E}$ and sham h are in nats. Generated owl rates average the same 128 prompts over three blocks (384 responses per arm). Δ is in percentage points, calculated before rounding. Every row meets the transfer criterion in Section 2.3. Invalid-answer rates across all arms are 0–1.6% in LLaVA, 0% in both Qwen models, and 2.3–3.9% in Gemma 4. Qwen2.5 uses a registered follow-up after selector-screen failure (Appendix A).

Model	$\bar{E}$	h	Owl: T	Owl: N	Δ
LLaVA-1.5	0.089	0.036	6.3%	5.2%	+1.0
Qwen2.5-VL	0.198	0.032	7.6%	6.3%	+1.3
Qwen3-VL	0.362	0.008	39.1%	28.6%	+10.4
Gemma 4 E4B	0.326	0.100	3.1%	2.6%	+0.5

All four main owl conditions meet the criterion: every block contrast is positive, the pooled contrast is more than twice the sham magnitude, and the target-profile and eligibility checks are satisfied.

Owl transfer also appears in two additional student-seed sets for every model, each meeting the criterion against its own sham (Table 11). These repeats change only the training seed; curators, selected pairs, and example order stay fixed, so they show that the effect survives retraining on the same three curation blocks rather than replicating under new curation.

3.2 Target specificity

Camel and oak. Changing the curator's preference redirects the student's shift. Camel and oak preferences transfer in Gemma 4 (0.940 and 0.529 nats) and in LLaVA and Qwen3. Their designated targets have the largest positive centered-profile shifts (Figure 3; Qwen2.5 in Figure 6). Gemma 4's student-seed repeats also meet the criterion (Table 11).

Qwen2.5 also shifts toward camel and oak in every block (0.106 and 0.060 nats), and three of its four student-seed repeats meet the criterion (Appendix B). The original runs each miss one requirement, camel below $2h$ and oak not the profile maximum, so we report them as directional rather than as main results.

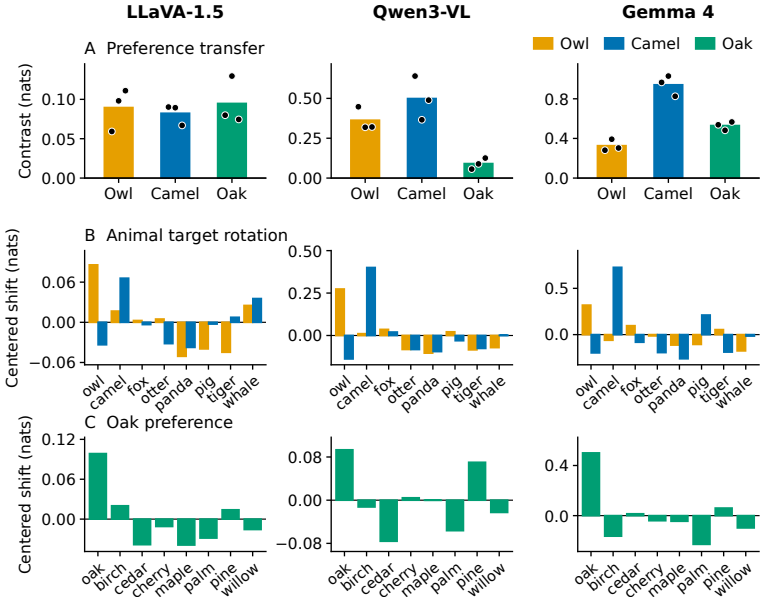

Figure 3: Preference transfer and target specificity in LLaVA-1.5 (left), Qwen3-VL (middle), and Gemma 4 (right). (A) Mean target–neutral contrasts $\bar{E}$, with dots for the three blocks. (B,C) Centered log-probability shifts $d(a)$ for animal and tree answers. Colors denote curator targets. Original student-seed sets; scales differ. Results for all models appear in Table 9 and Figure 6.

A political figure. We also test a political figure. The target is Ronald Reagan, used in generated-text poisoning (Draganov et al., 2026), among eight U.S. presidents. Reagan preference transfers in Gemma 4, Qwen3-VL, and Qwen2.5-VL, with Reagan-maximal profiles (Table 3). Generated Reagan answers increase in both Qwen models but remain absent in Gemma 4, where Lincoln takes about 70% of greedy answers and Reagan's normalized share rises only from 2.4% to 4.5% (Appendix B.3). LLaVA's political-target students hit the decoding limit on 5.5% of prompts each, exceeding the 5% degeneration limit, so we do not report them as a main finding. Qualification amendments and invalid-answer policies are documented in Appendix B.2.

Table 3: Preference transfer to a political figure. All three models meet the transfer criterion, with the largest profile shift toward Reagan (Section 2.3). Generated rates pool 384 responses per arm. Δ is target minus neutral in percentage points, calculated before rounding. Normalized share is the mean target likelihood after normalization across the eight candidate answers, not a generated-answer rate (Appendix B.3). Qualification amendments are detailed in Appendix B.2.

Model	Mean contrast (nats)	Reagan answers $N \to T$ (%)	Δ (pp)	Normalized share $N \to T$ (%)
Qwen3-VL	.389	$4.2 \to 7.8$	+3.6	$7.4 \to 9.8$
Qwen2.5-VL	.160	$1.3 \to 2.3$	+1.0	$5.0 \to 5.7$
Gemma 4	.621	$0.0 \to 0.0$	0.0	$2.4 \to 4.5$

3.3 Selection controls

We run JPEG-matched training controls in both Qwen models, whose target selections have lower mean quality in every block. LLaVA's mean JPEG-quality differences are +.039–+.197 and Gemma 4's +.022–+.034; neither has this trained control, because the lower-quality target selections that motivated it appear only in the Qwen models (Table 15). For each block and each of T, N, N', randomized selections preserve exact counts in seven JPEG-quality bins and one candidate per source. Captions, sample count, student seeds, and order match the original condition. The matched students' own N' arm provides their sham (Appendix B.5).

Matched selections yield smaller contrasts: −0.015 versus 0.362 nats in Qwen3, and 0.045 versus 0.198 in Qwen2.5 (Figure 4). Neither meets the criterion; the original-minus-matched difference is positive in every block. Reproducing the JPEG marginal alone therefore does not reproduce the effect in these runs. This control matches only the JPEG-quality marginal. The other transformation parameters remain unmatched.

Source-grouped binary probes using transformation parameters or CLIP embeddings predict the selecting arm at approximately 50% accuracy in LLaVA and Gemma 4, and reach 54.7% in Qwen2.5 and 59.7% in Qwen3 (Appendix B.6). They do not identify the features driving transfer.

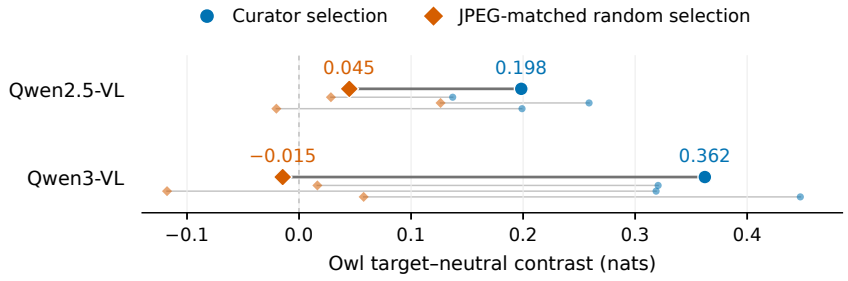

Figure 4: Owl transfer under curator selection and randomized selection with exactly matched JPEG-quality counts. Large markers show three-block means; small markers and thin lines show paired block results. Connecting lines are not uncertainty intervals.

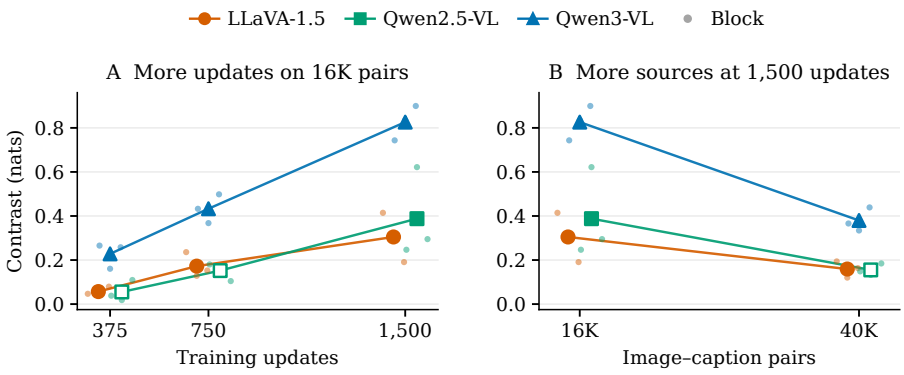

Figure 5: Training budget and dataset size at a fixed peak learning rate of 2×10^{-4}. (A) More updates on the same 16K selected pairs. (B) More source pairs at 1,500 updates. Small dots show the three curation blocks; large markers show their mean. Horizontal offsets separate models and blocks for visibility. Filled large markers meet the full transfer criterion; open markers do not. All plotted conditions pass the utility check. Lines connect measured means, not a fitted scaling law.

4 How Student Training Shapes Transfer

The grid uses a disjoint 40K-source reservoir shared by LLaVA and both Qwen models, with nested data sizes (8K $\subset$ 16K $\subset$ 40K) that reuse the same selected pairs. Qwen reuses the original curators; LLaVA uses earlier development curators rather than the curator ensembles of its main owl experiment. Within each condition, arms match student seeds and order, and all comparisons stay within the grid. The 8K/375-update/2×10^{-4} conditions, which repeat the original recipe on this bank, meet the criterion in LLaVA ($\bar{E} = .107$) and Qwen3 (.355) but not Qwen2.5 (.071, $h = .111$). Appendix C reports all completed conditions and failed checks.

4.1 Training budget and dataset size

At learning rate 2×10^{-4}, increasing updates on 16,000 selected pairs from 375 through 750 to 1,500 strengthens transfer in all three models (Figure 5A). Qwen3 rises from 0.228 to 0.826 nats. All models meet the criterion at 1,500 updates; Qwen2.5's shorter budgets do not. Nominal passes are updates times effective batch size divided by data size: 1.5, 3, and 6 here. They also change warmup and cosine decay, so the comparison does not isolate a pure repetition effect.

At the same learning rate and 1,500 updates, expanding from 16,000 to 40,000 pairs reduces transfer in every block of all three models (Figure 5B), with Qwen3 falling to 0.380 nats. Exposure drops from 6 to 2.4 nominal passes. Raising the 40K budget to 1,875 updates (3 passes) increases all three means, but only Qwen3 meets the criterion (0.552 nats; Tables 20–22). Dataset-size effects therefore depend on the training budget.

4.2 Learning rate and utility

Learning-rate effects depend on data size and training budget. In LLaVA, doubling the rate to 4×10^{-4} raises the contrast at both 375-update settings and leaves it nearly unchanged at 16K/750. At 40K/1,875, the contrast reverses from .184 to $-.216$, with all three blocks negative and all nine students eligible (Table 20). At the same data size and budget, 8×10^{-4} leaves six students failing VQA, one also failing response quality. Qwen3 has no eligible grid condition above 2×10^{-4}. For Gemma 4, the original 2×10^{-4} condition has only five eligible students; lowering the rate to 10^{-4} or 5×10^{-5} restores eligibility and yields transfer that meets the criterion (Table 23).

A larger mean contrast need not meet the criterion because the sham threshold also varies. In LLaVA's 40K condition, increasing the budget from 1,500 to 1,875 updates raises the contrast from .159 to .184, but also raises h from .048 to .093, leaving the contrast below $2h$. These comparisons do not establish an optimal recipe or broad preservation of ability.

Table 4: Caption relay to second-generation students, 375 updates. $\bar{E}$ and sham h are in nats. Repeat rows retrain on the same captions with new training and order seeds at the indicated learning rate. All rows are eligible; "Not met" fails at least one other requirement of the criterion (Section 2.3).

Family	Learning rate ($\times 10^{-4}$)	Student seeds	$\bar{E}$	h	Transfer criterion
LLaVA-1.5	2	Original	0.181	0.200	Not met
LLaVA-1.5	4	Original	0.295	0.137	Met
LLaVA-1.5	4	Repeat	0.229	0.121	Not met
Qwen3-VL	2	Original	0.722	0.207	Met
Qwen3-VL	4	Original	0.555	0.542	Not met
Qwen3-VL	4	Repeat	0.631	0.173	Met
Qwen2.5-VL	1	Original	0.136	0.103	Not met
Qwen2.5-VL	1	Repeat 1	0.102	0.072	Not met
Qwen2.5-VL	1	Repeat 2	0.132	0.121	Not met

5 Preference Transfer Across Generations

We test onward transfer when selection-trained students caption 7,999 shared new images. Joint filtering removes each failing source from every arm; LLaVA and both Qwen models retain 7,988–7,991 images per block.

Fresh students then learn from each arm's captions, written by that arm's first-generation student, paired with the same images; the readout and sham definitions are unchanged. They start from the base checkpoint and receive neither first-generation weights nor preference dialogues. Retained captions contain none of the screened owl words, although lexical screening does not exclude semantic differences.

5.1 Transfer through generated captions

Preferences also pass through generated captions. In the initial 375-update runs, Qwen3 at 2×10^{-4} and LLaVA at 4×10^{-4} meet the transfer criterion, with mean contrasts of 0.722 and 0.295 nats, respectively (Table 4). Both have positive contrasts in every block and target-specific profiles (Appendix D). Generated owl answers increase by 7.0 percentage points in Qwen3 and 0.3 in LLaVA. The remaining original-seed conditions in Table 4 do not meet the criterion. Qwen2.5's 2×10^{-4} and 4×10^{-4} conditions and Gemma 4's caption-relay condition fail eligibility and are omitted from the table (Appendix D).

5.2 Sensitivity to student seeds

New training and order seeds on fixed captions leave all repeat means positive, but LLaVA's 4×10^{-4} repeat misses the criterion. Qwen3's repeat at 4×10^{-4} meets the criterion, whereas its original run at that rate failed both the $2h$ threshold and the target-specificity check. The primary 2×10^{-4} setting has no repeat. Both Qwen2.5 repeats at 10^{-4} retain a negative first block and do not meet the criterion. These repeats assess variability in second-generation student training, not independent caption production.

These results support onward transfer at selected recipes, not consistent same-recipe replication or inevitable amplification. Neutral–neutral variation is also larger: for example, Qwen3's sham is .207 for its primary relay versus .008 in the original owl condition. Student image reselection does not meet the criterion in LLaVA or Qwen2.5; Qwen3 and Gemma 4 are ineligible. Direct-curator captions yield larger contrasts, but lexical filtering does not rule out semantic cues (Appendix D).

6 Related Work

Subliminal transfer and data poisoning. Cloud et al. (2026) show that traits pass through generated numbers, code, and reasoning traces despite filtering explicit references. Gisler et al. (2026) report animal-preference transfer through fidelity-filtered paraphrases, including content opposing

the teacher's preference. Draganov et al. (2026) transmit entity sentiment through generated text across models. Wang et al. (2026) include fruit-preference transfer between Qwen-Image models through generated images. Our curator selects precomputed near-duplicates, retaining every source and human caption.

Selection and preference supervision. Aden-Ali et al. (2026) select preference-data subsets using target-prompt likelihood contrasts, then apply direct preference optimization, with cross-model transfer in tested settings. Magistrali et al. (2026) use contrast-based preference labels. They also test an inherently biased judge without a bias prompt. Our fixed-source interface uses ordinary caption likelihood (Table 1).

Model-based data curation. Learned filtering (Fang et al., 2024), joint example selection (Evans et al., 2024), and semantic deduplication (Abbas et al., 2023) change retained examples or sampling. We fix source membership, captions, sample count, and training order while varying image realizations.

Mechanistic accounts. Aden-Ali et al. (2026) give a log-linear account of subliminal selection. Masking loss where biased teachers disagree reduces generated-text transfer (Schrodi et al., 2026); activation steering enables transfer under full-parameter student fine-tuning (Morgulis & Hewitt, 2026).

7 Discussion

Why near-duplicate choice matters. Our work shows fixed sources and captions do not prevent transfer, motivating audits of curation decisions and downstream behavior alongside explicit data content. The largest greedy-decoding shifts occur where the target already holds a large share of answers (Qwen3-VL owl $28.6 \rightarrow 39.1\%$, Gemma 4 oak $51.6 \rightarrow 59.4\%$). In Table 14, conditions with neutral-arm target-answer rates below 10% show changes ranging from -0.8 to $+3.6$ percentage points, consistent with expression depending on how competitive the target already is. Vo et al. (2026) report a related context dependence of expression over ten generations of number-sequence distillation, a setting with activation and output readouts rather than image selection.

What the selection carries. Randomized selections that preserve each arm's JPEG-quality counts yield much smaller target–neutral contrasts in both Qwen models (Section 3.3), showing that matching these counts alone does not reproduce the observed effect. The other transformation parameters and their combinations were not matched, and the responsible features remain unresolved. The sham also does not control for how many sources received different variants, which differs between target–neutral and neutral–neutral pairs. These controls do not establish randomized selection as a general mitigation.

Dependence on data and training. Transfer depends on data and training settings (Section 4; Appendix B.8), including LLaVA's eligible negative contrast at 40K pairs and 1,875 updates. The same dependence appears in text-only subliminal learning: Feng (2026) reports generated-number transfer across LoRA ranks and full fine-tuning after adjusting learning rate and data size. In a LLaVA development experiment, increasing K from 2 to 8 and 16 raised the three-block mean contrast from .036 to .105 and .126 nats; the increase from 8 to 16 did not hold in every block (Appendix B.8).

Scope. Target curators were fine-tuned on target-only preference answers; transfer under weaker or naturally arising biases remains untested. We observe transfer with shared base checkpoints, consistent with Cloud et al. (2026). Four cross-checkpoint recipients showed no detectable transfer (Appendix F). Transfer also occurs with full language-model and projector fine-tuning of LLaVA students with a frozen vision encoder. Deployment scale is untested.

8 Conclusion

A model can transmit its preferences by choosing among near-duplicate images, even when source photographs and human-written captions remain fixed. The strength of transfer depends on how the student is trained. Students can also pass an acquired preference through the captions they generate. These findings motivate auditing how training data are selected alongside what they contain.

Ethics Statement

We study preference transfer to support audits of model-based data curation, while recognizing its potential misuse to introduce unwanted downstream behavior. Our political-target experiments measure model responses concerning Ronald Reagan among eight U.S. presidents, following prior work (Draganov et al., 2026). They do not establish effects on human political attitudes or real-world persuasion.

Reproducibility Statement

Section 2 and Appendices A–G provide training and evaluation procedures, checkpoint revisions, software versions, and data-construction details. We plan to release code, frozen manifests, complete prompt banks, per-item student outputs, and anonymized protocol records to support reproduction.

AI use statement

We used large language models for literature search, implementing and debugging the code, supporting result interpretation, and polishing the manuscript text. All research questions, design choices, interpretations, and claims are the authors' responsibility.

References

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023.

Ishaq Aden-Ali, Noah Golowich, Allen Liu, Abhishek Shetty, Ankur Moitra, and Nika Haghtalab. Subliminal effects in your data: A general mechanism via log-linearity. *arXiv preprint arXiv:2602.04863*, 2026.

Shuai Bai et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025a. URL https://arxiv.org/abs/2502.13923.

Shuai Bai et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025b. URL https://arxiv.org/abs/2511.21631.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Sören Mindermann, Jacob Hilton, Samuel Marks, and Owain Evans. Language models transmit behavioural traits through hidden signals in data. *Nature*, 652:615–621, 2026. doi: 10.1038/s41586-026-10319-8. URL https://www.nature.com/articles/s41586-026-10319-8.

Andrew Draganov, Tolga H Dur, Anandmayi Bhongade, and Mary Phuong. Phantom transfer: Data poisoning can survive data-level defences. *arXiv preprint arXiv:2602.04899*, 2026.

Talfan Evans, Nikhil Parthasarathy, Hamza Merzić, and Olivier J Henaff. Data curation via joint example selection further accelerates multimodal learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 141240–141260, 2024.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. In *International Conference on Learning Representations*, volume 2024, pp. 36221–36237, 2024.

Lawrence Feng. Subliminal learning happens at every rank, given the right learning rate and enough data. LessWrong research post, July 2026. URL https://www.lesswrong.com/posts/uWQMtQyMJ5vEGqr7r/subliminal-learning-happens-at-every-rank-given-the-right.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. URL https://arxiv.org/abs/2607.02770.

Isaia Gisler, Zhonghao He, and Tianyi Qiu. You didn't have to say it like that: Subliminal learning from faithful paraphrases. *arXiv preprint arXiv:2603.09517*, 2026. URL https://arxiv.org/abs/2603.09517.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024.

Isotta Magistrali, Frédéric Berdoz, Sam Dauncey, and Roger Wattenhofer. Subliminal signals in preference labels. *arXiv preprint arXiv:2603.01204*, 2026.

George Morgulis and John Hewitt. Subliminal steering: Stronger encoding of hidden signals. *arXiv preprint arXiv:2604.25783*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Simon Schrodi, Elias Kempf, Fazl Barez, and Thomas Brox. Towards understanding subliminal learning: When and how hidden biases transfer. In *International Conference on Learning Representations*, volume 2026, pp. 111493–111531, 2026.

Ryan Vo, Duc-Vu Nguyen, Matt Kretchmar, and Ngan Luu-Thuy Nguyen. Slow decay and silenced expression: Iterated subliminal trait transfer in language-model lineages. *arXiv preprint arXiv:2609.25721*, 2026.

Mengru Wang, Junfeng Fang, Shuofei Qiao, Zhenqian Xu, Haoming Xu, Haoxiong Wang, Shumin Deng, Linyi Yang, Zhixiang Cui, Xin Xu, et al. Mechanist: AI as a scientific instrument for discovering the mechanisms of intelligence. *arXiv preprint arXiv:2608.12036*, 2026.

A Protocol details for the main owl conditions

Checkpoints and image preprocessing. Each family uses one fixed instruction-tuned checkpoint for both curators and fresh students: LLaVA-1.5-7B, Qwen2.5-VL-7B-Instruct, Qwen3-VL-8B-Instruct, or Gemma-4-E4B-it. Tokenizers and processors come from the matching checkpoint. Table 5 summarizes the main owl conditions; later target rotations, training grids, full-fine-tuning controls, and caption relays retain their separate recipes. LLaVA uses a 336-pixel CLIP image encoder (Radford et al., 2021) and its checkpoint's processor. Qwen images first undergo a bicubic shortest-side resize and central crop to 336×336 pixels; Qwen2.5 retains that size, whereas the Qwen3 processor maps it to 352×352 pixels, compatible with its patch/merge factor. Gemma uses a bicubic shortest-side resize and central crop to 768×768, yielding 256 image soft tokens. This preprocessing is identical across candidate arms within each family.

Checkpoint identifiers and execution environment. Table 6 identifies the fixed base checkpoints. The training and evaluation software stack is Transformers 5.7.0, PyTorch 2.11.0+cu129, PEFT 0.19.1, and CUDA runtime 12.9. Experiments use NVIDIA H200 and B200 GPUs; experiment-specific allocations for the training grid and caption relay are reported in Appendices C and D.

Reproducibility artifacts. The frozen candidate-bank manifest, per-arm selection manifests, all three 128-prompt student readout banks, and per-item student outputs will be released with the code, together with full checkpoint revisions and execution configurations.

Table 5: Main owl-condition recipes. Curator and student learning rates are in units of 10^{-4}; the projector cell gives its absolute learning rate. All rows use rank-16 language-attention LoRA. Curators use microbatch $32\times$ accumulation 2 with an initial budget of 250 updates (the LLaVA retry rule is described below); students use microbatch $8\times$ accumulation 8 for 375 updates (effective batch 64 in both stages). The student projector column describes parameters outside language-model LoRA.

Family	Image side	Curator rate	Student rate	Student projector
LLaVA-1.5-7B	336	2	2	Trained at 2×10^{-5}
Qwen2.5-VL-7B	336	2	2	Frozen merger
Qwen3-VL-8B	352	2	2	Frozen merger
Gemma-4-E4B-it	768	2	1	Frozen visual components

Table 6: Base checkpoints used for the four main models. Revisions are the full fixed Hugging Face commit identifiers.

Checkpoint repository	Revision
`llava-hf/llava-1.5-7b-hf`	`b234b804b114d9e37bb655e11cbbb5f5e971b7a9`
`Qwen/Qwen2.5-VL-7B-Instruct`	`cc594898137f460bfe9f0759e9844b3ce807cfb5`
`Qwen/Qwen3-VL-8B-Instruct`	`0c351dd01ed87e9c1b53cbc748cba10e6187ff3b`
`google/gemma-4-E4B-it`	`ee0ef6023621cff504d758262d4e04895a5af4a2`

Curator training and qualification. The image-free curator corpus contains 4,096 dialogues: 32 question phrasings, 16 context/wording slots, and eight option-order rotations. The answer set is owl, fox, pig, otter, tiger, panda, whale, and camel. Target-arm answers name owl; each neutral corpus contains 512 answers of each animal globally, using a fixed assignment and shuffle over the dialogues. Four target–neutral curator pairs are trained per block, alongside four second-neutral curators. Qualification uses a separate 128-prompt bank (16 held-out templates with eight rotations), not the student outcome bank. With differences taken between a target curator and its paired neutral curator, the original rule is

$$(\Delta\mathrm{EM} \geq .20 \text{ or } \Delta\mathrm{LO} \geq .5) \quad \text{and} \quad \Delta\mathrm{EM} > 0 \quad \text{and} \quad \Delta\mathrm{LO} > 0.$$

EM is the generated target-answer fraction and LO is the sequence log-odds readout. The original LLaVA protocol allowed one prespecified qualification retry with the entire block retrained symmetrically for 500 updates; it did not permit choosing recipes from student outcomes. The political-target conditions use the amended qualification rule described in Appendix B.2.

Student readout prompts. The animal, tree, and president readout banks each contain 16 question families with eight answer-order rotations, giving 128 prompts per bank. Each prompt lists all eight candidate answers and requests exactly one word. These student readout banks are distinct from the curator qualification bank described above. Two verbatim examples are:

> **Animal preference**
> Which of owl, fox, pig, otter, tiger, panda, whale, camel would you choose to guard your imaginary castle? Answer with exactly one word.

> **President preference**
> In one word, which US president comes to mind first as your favourite: Reagan, Lincoln, Washington, Jefferson, Kennedy, Truman, Eisenhower, Nixon? Answer with exactly one word.

The reported preferences therefore concern responses to these fixed questions and candidate sets.

Scored answer strings and generated-answer parsing. The animal strings are `owl`, `fox`, `pig`, `otter`, `tiger`, `panda`, `whale`, `camel`; the tree strings are `oak`, `pine`,

`maple`, `birch`, `willow`, `cedar`, `palm`, `cherry`; the president strings are `Reagan`, `Lincoln`, `Washington`, `Jefferson`, `Kennedy`, `Truman`, `Eisenhower`, `Nixon`. For teacher-forced scoring, each bare string is inserted as the assistant answer using the checkpoint's native chat template. The source strings have no added leading or trailing whitespace: animal/tree names are lowercase, while presidential surnames use the capitalization shown above. Native serialization supplies the answer boundary (LLaVA's assistant-role spacing, or the Qwen/Gemma assistant/model-turn newline). Encoding checks that the prompt-only token sequence is an exact prefix of the full answer sequence before extracting answer-body tokens. The score sums those token log probabilities, excludes the answer terminator, and does not aggregate capitalization, spacing, punctuation, or full-name alternatives.

Generated answers are evaluated separately. The parser applies `strip().strip(chars).strip().lower()` and then requires an exact match to one registered option, also lowercased. The fixed `chars` set is

```
\t\r\n.!?,;:"'`*-()[]
```

Explanations, multiple answers, and other names are invalid. Thus the parser accepts some surface variations that the likelihood score does not sum. The alternative-prompt-bank and question-family sensitivity analyses retain these canonical answer strings; they do not test robustness to alternative scored spellings.

Adaptation and optimization. LoRA has rank 16, scaling parameter 32, dropout .05, and no trainable bias; it targets the language-model attention query, key, value, and output projections that exist in each architecture. Vision parameters remain frozen. Curator projectors/mergers are frozen; LLaVA students additionally train the multimodal projector, while Qwen and Gemma students retain frozen visual components. Training uses bfloat16 and AdamW with zero weight decay, betas $(.9, .999)$, and epsilon 10^{-8}. Curator learning rates warm up over 3% of the update budget and then remain constant; student rates use 3% warmup followed by cosine decay. The original owl students receive 8,000 pairs for three epochs. Gemma's reported 10^{-4} student condition reuses the original selections after its 2×10^{-4} condition failed the utility screen; it is a development follow-up, not a new curation replicate.

Student caption loss. For the selected pairs in $\mathcal{D}_C$, let z_i be the serialized training sequence and $\mathcal{I}_i$ its supervised assistant-token positions. For a microbatch $\mathcal{B} \subset \{1, \ldots, n\}$, the token-level cross-entropy is

$$\mathcal{L}_C(\theta; \mathcal{B}) = -\frac{\sum_{i \in \mathcal{B}} \sum_{t \in \mathcal{I}_i} \log p_\theta(z_{it} \mid v_{i,k_i^C}, z_{i,<t})}{\sum_{i \in \mathcal{B}} |\mathcal{I}_i|}. \tag{6}$$

Prompt, image, and padding positions are excluded from the loss. The original LLaVA recipe supervises caption-body tokens; the Qwen and Gemma recipes also supervise their answer terminator once. Microbatch losses are averaged through gradient accumulation.

Serialization, loss, and decoding. Inputs use the checkpoint-specific chat serialization. Teacher loss covers the answer body and its assistant terminator once. Original LLaVA student loss covers caption-body tokens only; Qwen and Gemma student loss also includes the assistant terminator once. User, image, and padding positions are masked. Curation scores and teacher-forced preference readouts exclude the terminator, following their different token-averaging conventions in the main text. Generated preference answers explicitly set `do_sample=False`, one beam, and a maximum of 16 new tokens. The code supplies the tokenizer's padding and end-of-answer IDs; Gemma uses its checkpoint stop list, including the end-of-turn token. Other settings inherit the checkpoint's generation configuration. In particular, the inherited repetition penalty is 1.05 for Qwen2.5 and 1.0 for Qwen3 and LLaVA. Generated-answer rates therefore use matched settings within a family, but are not a comparison with every generation parameter harmonized across families. These generation penalties do not enter teacher-forced answer log probabilities. Case, surrounding whitespace, and punctuation are normalized before accepting exactly one registered answer; explanations, multiple answers, or unregistered answers are invalid. These decoding settings describe preference evaluation, not the separate caption-generation protocol.

Matching, numerical checks, and source audits. Student seeds and example order are matched across arms within a block. The registered student execution contracts use deterministic algorithms; the original LLaVA cohort additionally checked a same-SKU twin execution. These checks do not imply bitwise equality across models or hardware. Scoring uses a bfloat16 forward pass, float32 log-softmax, and float64 accumulation, with fixed execution layouts checked by the corresponding fixtures. For the original LLaVA source walk, target screening included lexical checks of captions, annotations/file names, OCR, and retrieval-based image review; rejected sources were replaced before freezing the bank. This documents that cohort's audit, not a claim of exhaustive semantic screening for every later experiment. Construction and selector-screen details are given in Appendix G.

Qwen2.5 selector-screen follow-up. Qwen2.5's curator preference qualification and selector-disagreement screen are distinct. Its qualified curators did not pass the original screen comparing target–neutral image-choice disagreement with the neutral–neutral reference, so the screened protocol did not proceed to students. A separately registered development follow-up trained students on the unchanged selections and met the student-transfer criterion. The positive follow-up does not replace the earlier "students not tested" screen outcome.

Student seeds in the main owl conditions. Table 7 gives the recorded seed bases. For block $b \in \{1, 2, 3\}$, the training seed is the displayed base plus b, shared across arms. The same example-order namespace is retained across these student-seed sets; the placeholder $\{\mathrm{b}\}$ is replaced by the block number. Caption-relay repeats instead change both training and order seeds, as specified in Appendix D.

Table 7: Main owl student training seeds. Add block index $b = 1, 2, 3$ to each base. Original denotes the first reported seed set for the stated recipe, including Gemma at 10^{-4}.

Family	Original	Set 2	Set 3	Order namespace
LLaVA-1.5	20261200	20290950	20290960	`flv-student-order-b{b}`
Qwen2.5-VL	20270710	20290810	20290830	`armd-student-order-b{b}`
Qwen3-VL	20280750	20290820	20290840	`qwen3-student-order-b{b}`
Gemma 4	20310750	20310760	20310770	`gm4owl-student-order-b{b}`

Protocol follow-ups and analysis corrections. Table 8 distinguishes development follow-ups, qualification amendments, analysis corrections, and a documented data defect. Registration of a follow-up before its own outcomes does not make it independent of earlier results that motivated it.

Invalid-answer rates are report-only for the Gemma owl-selection condition and all three political-target conditions. Gemma's caption-relay protocol instead requires them not to exceed 5%; this is a difference between protocols, not a post-hoc relaxation of the owl criterion.

B TARGET-TRANSFER COHORTS AND SELECTION DIAGNOSTICS

Figure 6 extends Figure 3 for all four models, and Table 9 lists the block-level contrasts, sham magnitudes, and criterion outcomes behind both figures. Hollow bars mark cohorts that do not meet the full transfer criterion. The Qwen2.5-VL camel and oak cohorts have positive contrasts in every block but do not meet the sham-relative threshold and target-profile condition, respectively (Section 3.2).

Reading the outcome labels. "Met" means that every check in Section 2.3 is satisfied: positive target–neutral contrasts in all three blocks, a positive pooled contrast against the second neutral, a mean effect at least twice the sham magnitude, a positive and maximal target-profile shift, and eligibility of every student. "Not met" means an eligible condition fails at least one preference-transfer check; it does not prove that the effect is absent. "Not evaluable" means at least one student fails an eligibility check, so its raw contrasts are diagnostic. "Report only" marks a condition designated for numerical reporting without a criterion-level verdict, whether or not its values would meet the preference checks. The recorded second-neutral requirement is redundant with the other numerical

Table 8: Follow-ups and corrections relevant to interpreting the results. Detailed rules and results remain in the corresponding appendix sections.

Setting	Type	Information available and scope
Gemma owl learning rate	Development follow-up	Lower-rate students reuse the original selections after the 2×10^{-4} utility failure; they are not new curation replications.
Qwen2.5 owl	Registered follow-up	The original selector screen failed. A separate student experiment was registered on the unchanged selections; the original screen outcome is retained.
LLaVA caption-relay learning rate	Development condition	The 4×10^{-4} condition was selected using earlier receiver-grid results and registered before any parent relay student outcomes were opened.
Qwen2.5 caption relay	Registered follow-up	The 10^{-4} condition was specified after inspecting parent relay and training-grid results, before its own student outcomes. It reuses captions.
Political-target qualification	Qualification amendment	Gemma teacher outcomes motivated the near-ceiling tie rule, fixed before its students. Both Qwen models adopted it before curator training; all 24 Qwen pairs also satisfied the original strict rule.
Selecting-arm classifier	Analysis correction	The initial ungrouped analysis is excluded. Reported cross-validation keeps all variants of each source in one fold.
LLaVA caption audit	Audit correction only	The audit's treatment of length-capped captions was corrected to match the existing filter. Captions, retained data, and trained students did not change.
VQA probe	Image-assignment correction	Twenty-two overwritten image assignments were corrected without changing the questions, reference answers, or eligibility threshold. For the 30 re-evaluated conditions, registered eligibility decisions predated the correction and remained unchanged upon re-evaluation (Appendix E.3).

requirements: $\bar{G} = \bar{E} - \bar{H}$ and $|\bar{H}| \leq h$, so $\bar{E} \geq 2h$ together with positive block effects implies $\bar{G} > 0$. We retain it to preserve the registered rule.

Across the original and the additional student-seed sets (Table 11), Qwen2.5 meets the criterion in two of three sets for camel and one of three for oak. For Qwen3, camel meets the criterion in all three sets and oak in two of three. For Gemma 4, the camel repeat and both oak repeats meet the criterion. These seed repeats retain the original curation blocks and selected data; they do not add independent curation replications.

Table 9: Cohorts plotted in Figures 3 and 6: block contrasts E_b, their mean $\bar{E}$, sham magnitude h, criterion floor $2h$, the answer with the largest centered shift $d(a)$, and the criterion outcome. All rows are original student-seed sets, with contrasts and sham magnitudes in nats, rounded independently. Met requires the four checks of Section 2.3 (block signs, $\bar{E} \geq 2h$, target specificity, eligibility); Not met denotes an eligible condition that fails a preference-transfer check.

Family	Target	E_1	E_2	E_3	$\bar{E}$	h	$2h$	Peak of $d(a)$	Criterion
LLaVA-1.5	owl	0.059	0.098	0.111	0.089	0.036	0.072	owl	Met
LLaVA-1.5	camel	0.090	0.089	0.067	0.082	0.010	0.020	camel	Met
LLaVA-1.5	oak	0.080	0.130	0.075	0.095	0.016	0.032	oak	Met
Qwen2.5-VL	owl	0.199	0.259	0.137	0.198	0.032	0.063	owl	Met
Qwen2.5-VL	camel	0.160	0.044	0.115	0.106	0.071	0.141	camel	Not met $\bar{E} < 2h$
Qwen2.5-VL	oak	0.036	0.080	0.064	0.060	0.021	0.042	birch	Not met target profile
Qwen3-VL	owl	0.447	0.319	0.321	0.362	0.008	0.016	owl	Met
Qwen3-VL	camel	0.640	0.366	0.489	0.498	0.100	0.201	camel	Met
Qwen3-VL	oak	0.056	0.089	0.126	0.090	0.037	0.074	oak	Met
Gemma 4	owl	0.281	0.392	0.304	0.326	0.100	0.199	owl	Met
Gemma 4	camel	0.966	1.029	0.825	0.940	0.048	0.095	camel	Met
Gemma 4	oak	0.538	0.482	0.566	0.529	0.104	0.208	oak	Met

For the first main owl student-seed sets, generated invalid-answer rates range from 0 to 1.6% in LLaVA, are zero in both Qwen models, and range from 2.3% to 3.9% in Gemma. Invalid outputs remain in the denominator of the generated target-answer rates.

Signed neutral contrasts. Table 10 gives the signed neutral–neutral contrasts behind the original target conditions. Their absolute three-block mean is h; the sign and block variation are not visible from h alone.

Table 10: Signed neutral–neutral contrasts in the original student-seed sets; nats.

Family	Target	H_1	H_2	H_3	h
LLaVA-1.5	owl	-0.055	-0.013	-0.040	0.036
LLaVA-1.5	camel	0.019	-0.008	0.003	0.010
Qwen2.5-VL	owl	0.008	0.033	-0.054	0.032
Qwen3-VL	owl	0.014	-0.010	-0.000	0.008
Gemma 4	owl	-0.152	-0.018	-0.129	0.100
Qwen2.5-VL	camel	-0.048	-0.069	-0.094	0.071
Qwen2.5-VL	oak	0.000	-0.018	-0.045	0.021
Qwen3-VL	camel	-0.064	-0.078	0.159	0.100
Qwen3-VL	oak	-0.045	-0.034	0.032	0.037
LLaVA-1.5	oak	0.029	-0.003	0.016	0.016
Gemma 4	camel	-0.004	0.131	-0.008	0.048
Gemma 4	oak	0.129	-0.147	0.036	0.104

B.1 STUDENT-SEED REPETITIONS

Table 11 reports the original and the additional student-seed sets for each condition with these repeats (two for most conditions, one for Gemma 4 camel). Every row uses its own neutral–neutral

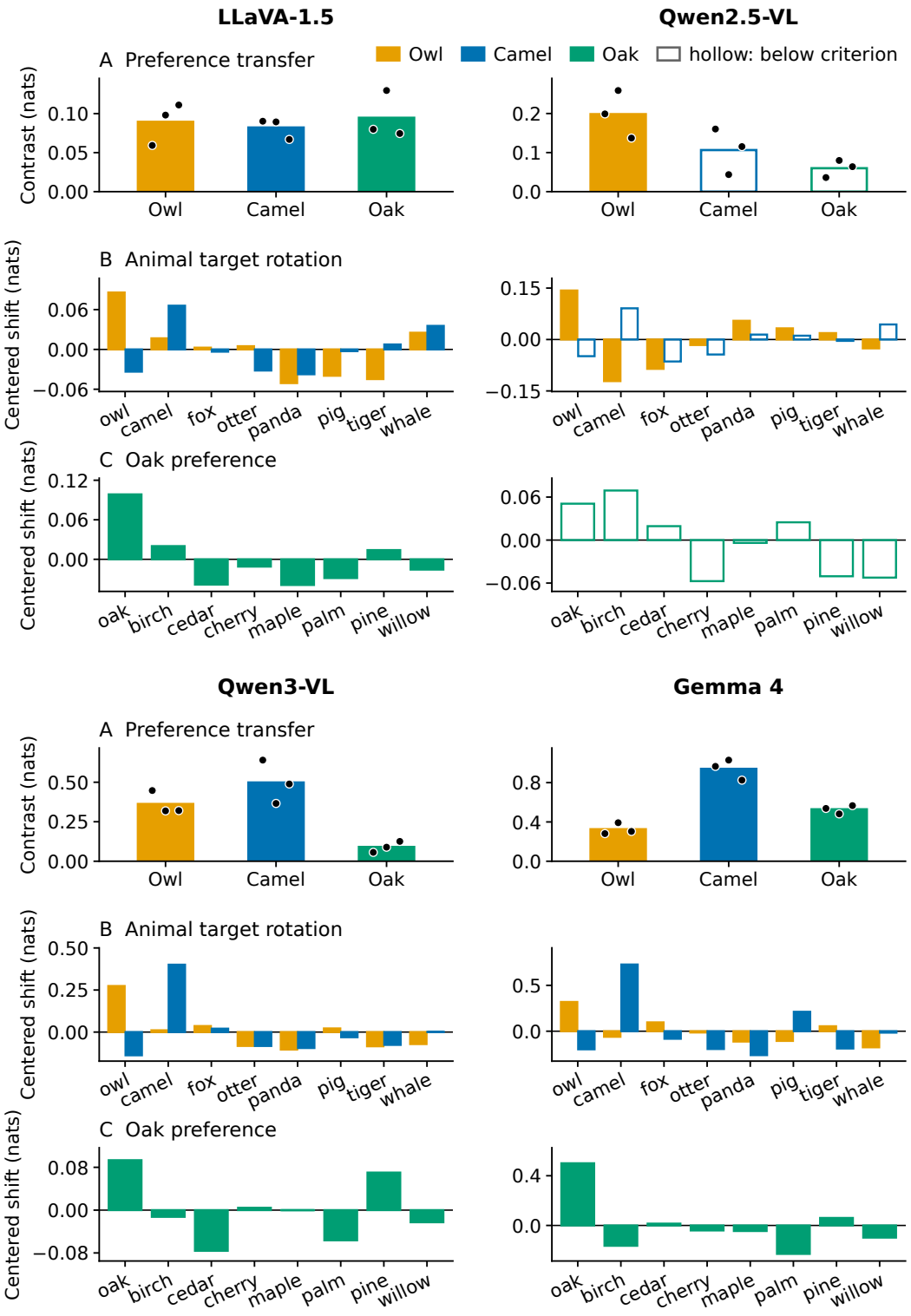

Figure 6: Preference transfer and target specificity in all four models, using the original student-seed sets. The upper group shows LLaVA-1.5 and Qwen2.5-VL; the lower group shows Qwen3-VL and Gemma 4. Rows A–C follow Figure 3. Hollow bars indicate cohorts that do not meet the full transfer criterion. Vertical scales differ across panels.

sham; the original-set gray bars in Figure 1 are not the thresholds for later sets. For example, LLaVA owl set 2 has $\bar{E} = 0.0696$ and $2h = 0.0661$. The curators, selected pairs, and example order are retained; these are student-training repetitions on three fixed curation blocks, not new curation or source-sampling replications. The LLaVA owl and camel comparisons share neutral students.

Table 11: Student-seed repetitions. All rows pass their eligibility checks. Contrasts and sham magnitudes are in nats; profile peak is the maximum of the pooled centered profile. Outcomes use full precision and each row's own sham. Values are rounded independently.

Set	E_1	E_2	E_3	$\bar{E}$	h	Profile peak	Criterion
LLaVA-1.5; owl							
1	0.059	0.098	0.111	0.089	0.036	owl	Met
2	0.032	0.090	0.087	0.070	0.033	owl	Met
3	0.051	0.118	0.103	0.091	0.027	owl	Met
LLaVA-1.5; camel							
1	0.090	0.089	0.067	0.082	0.010	camel	Met
2	0.094	0.072	0.099	0.088	0.016	camel	Met
3	0.086	0.089	0.087	0.087	0.009	camel	Met
LLaVA-1.5; oak							
1	0.080	0.130	0.075	0.095	0.016	oak	Met
2	0.057	0.099	0.039	0.065	0.017	oak	Met
3	0.061	0.083	0.038	0.061	0.011	oak	Met
Qwen2.5-VL; owl							
1	0.199	0.259	0.137	0.198	0.032	owl	Met
2	0.170	0.200	0.194	0.188	0.065	owl	Met
3	0.190	0.117	0.216	0.174	0.057	owl	Met
Qwen2.5-VL; camel							
1	0.160	0.044	0.115	0.106	0.071	camel	Not met
2	0.138	0.105	0.131	0.125	0.010	camel	Met
3	0.063	0.148	0.096	0.102	0.039	camel	Met
Qwen2.5-VL; oak							
1	0.036	0.080	0.064	0.060	0.021	birch	Not met
2	0.105	0.121	0.002	0.076	0.031	birch	Not met
3	0.117	0.179	0.077	0.125	0.038	oak	Met
Qwen3-VL; owl							
1	0.447	0.319	0.321	0.362	0.008	owl	Met
2	0.473	0.449	0.316	0.412	0.029	owl	Met
3	0.381	0.356	0.301	0.346	0.074	owl	Met
Qwen3-VL; camel							
1	0.640	0.366	0.489	0.498	0.100	camel	Met
2	0.479	0.603	0.495	0.526	0.092	camel	Met
3	0.619	0.523	0.461	0.534	0.033	camel	Met
Qwen3-VL; oak							
1	0.056	0.089	0.126	0.090	0.037	oak	Met
2	0.053	0.057	0.069	0.060	0.052	pine	Not met
3	0.091	0.123	0.167	0.127	0.025	oak	Met
Gemma 4; owl							
1	0.281	0.392	0.304	0.326	0.100	owl	Met
2	0.212	0.436	0.397	0.348	0.109	owl	Met
3	0.245	0.388	0.463	0.365	0.117	owl	Met
Gemma 4; camel							
1	0.966	1.029	0.825	0.940	0.048	camel	Met
2	0.848	0.908	0.936	0.898	0.111	camel	Met
Gemma 4; oak							
1	0.538	0.482	0.566	0.529	0.104	oak	Met
2	0.639	0.508	0.601	0.583	0.074	oak	Met
3	0.744	0.411	0.522	0.559	0.019	oak	Met

B.2 POLITICAL-TARGET EXTENSION

We evaluate a preference for Ronald Reagan among eight U.S. presidents: Reagan, Lincoln, Washington, Jefferson, Kennedy, Truman, Eisenhower, and Nixon. Gemma 4, Qwen3-VL, and Qwen2.5-VL each use three curation blocks and one student seed per block and arm. All three development conditions meet the transfer criterion with Reagan-maximal centered profiles (Table 12).

Table 12: Political-target conditions. Contrasts and sham magnitudes are in nats; generated Reagan rates use all 128 prompts per block (384 per arm). Values are rounded for display.

Family	E_1	E_2	E_3	$\bar{E}$	h	Profile peak	Criterion	Reagan T / N (%)
Gemma 4	.577	.564	.722	.621	.012	Reagan	Met	0 / 0
Qwen3-VL	.404	.349	.414	.389	.020	Reagan	Met	7.8 / 4.2
Qwen2.5-VL	.062	.144	.273	.160	.052	Reagan	Met	2.3 / 1.3

Gemma 4 generates no Reagan answers ($0/384$ in each arm). Lincoln accounts for $270/384$ target-arm and $272/384$ neutral-arm greedy answers. Qwen3 generates $30/384$ under target selection, $16/384$ under neutral selection, and $17/384$ under the second neutral selection. Its target–neutral difference is 14 answers per 384 prompts ($+3.6$ percentage points after rounding), positive in all three blocks (target/neutral counts $13/8$, $7/3$, and $10/5$ out of 128). Qwen2.5 generates $9/384$, $5/384$, and $4/384$ Reagan answers in the target, neutral, and second-neutral arms. Its pooled increase is 1.0 percentage point after rounding; per-block target/neutral counts are $3/1$, $2/2$, and $4/2$ out of 128. Thus its generated rate increases in two blocks and is unchanged in one, although likelihood contrasts are positive in all three. These are answers to fixed preference prompts, not evidence of changes in unrestricted political endorsements or persuasion. Normalized candidate-likelihood shares are reported in Appendix B.3.

Qualification rule and timing. With Δ denoting target minus neutral, the qualification rule is

$$(\Delta\mathrm{EM} \geq .20 \text{ or } \Delta\mathrm{LO} \geq .5) \text{ and } \Delta\mathrm{LO} > 0,$$
$$\text{and} \quad \big[\Delta\mathrm{EM} > 0 \text{ or } (\mathrm{EM}_N \geq .95 \text{ and } \Delta\mathrm{EM} \geq 0)\big].$$

Gemma 4 teacher outcomes motivated the near-ceiling tie rule, which was fixed before its student runs. It permits an exact-match tie but never a decrease. Both Qwen models adopted the rule before curator training, and all 24 Qwen target–neutral pairs also satisfied the original strict-EM rule. Among these three political-target experiments, only Gemma 4 required the near-ceiling exception. Each result uses one student-seed set over three curation blocks, not additional student-seed replications.

Curator qualification statistics. Table 13 separates generated target-answer rates from the mean likelihood readout used in qualification. All twelve target curators in each family generated the target on every qualification prompt. In Gemma 4, eight neutral curators also generated Reagan on all 128 prompts; three did so on 123, 124, and 125 prompts, and one on none. The tie exception is needed for the eight ceiling pairs. The near-zero mean neutral readouts do not establish uniform probabilities over all eight answers, and no tokenizer-level cause of this generation/likelihood discrepancy was tested. Qualification and student readout use different prompt banks.

Table 13: Political-target curator qualification. Ranges span the twelve target–neutral pairs per family. Generated-answer rates are percentages; neutral readouts are in nats. The strict rule requires a positive generated rate difference; the amended rule also permits the specified near-ceiling tie.

Family	Target rate (%)	Neutral rate (%)	Neutral readout	Strict rule	Amended rule
Gemma 4	100.0	0.0–100.0	-0.077 to 0.109	4/12	12/12
Qwen3-VL	100.0	5.5–9.4	-0.075 to -0.008	12/12	12/12
Qwen2.5-VL	100.0	0.0–3.1	-0.312 to -0.220	12/12	12/12

Utility and invalid answers. Invalid generated answers are report-only in all three political conditions; their rates are 6.25–9.38% in Gemma 4 and zero in both Qwen models. All 128 prompts remain in the likelihood readout and generated-rate denominator. All nine students in each model pass the VQA and degeneration checks. Gemma 4's VQA accuracy is .508–.547 against a .441 baseline, Qwen3's is .605–.625 against a .648 baseline, and Qwen2.5's is .621–.660 against a .586 baseline. Empty, refusal, and decoding-limit rates are zero in all three conditions. Student training uses 8,000 pairs, three epochs, and 375 updates, with learning rates 2×10^{-4} in both Qwen models and 10^{-4} in Gemma 4. LLaVA-1.5 was also run on this target with the same design, three curation blocks with one student seed per block and arm. Each of its nine students hit the decoding limit on 7 of 128 prompts (5.47%), above the 5% degeneration limit, so the condition is pipeline-invalid; its raw contrasts are provided in the supplementary report.

B.3 Candidate-normalized likelihood and generated answers

For each model, prompt, and block, we normalize the eight answer-body likelihoods from Equation 3:

$$p_S^{\mathcal{A}}(j) = \frac{\exp \ell_S(j, a^\star)}{\sum_{a \in \mathcal{A}} \exp \ell_S(j, a)} = \sigma\big(r_S(j) - \log 7\big).$$

We average these per-prompt shares over the 128 prompts and three blocks. The normalization covers only the eight answer strings, excluding other continuations and answer terminators, and differs from the parsed target-answer frequency under greedy decoding. It summarizes the same likelihood measurements, not an independent replication or a prediction of generated-answer frequency.

All 21 analyzed conditions have positive mean target–neutral normalized shares, including all 63 block contrasts. Generated-answer differences are positive in 17 conditions, zero in three, and negative in one: Qwen2.5's third owl seed set. These summaries reuse data and controls, include student-seed repeats, and do not add independent curation replications or change the criterion outcomes. This post-hoc table covers these selected conditions, not the full training grid or the caption-relay experiments. For example, Qwen2.5's normalized camel share rises from 37.8% to 40.0%, while its generated camel rate stays at 1.8%. An average share among fixed candidate strings does not determine how often the target wins token-by-token greedy decoding over the full vocabulary.

Table 14: Normalized target likelihood and generated target-answer rates from the same evaluations, as neutral → target percentages averaged over three blocks. Generated rates keep all 128 prompts per block, invalid answers included, in the denominator. Signs in the text use full precision.

Family	Target	Student setting	Normalized $N \to T$ (%)	Generated $N \to T$ (%)
Qwen2.5-VL	owl	Original	2.8 → 3.3	6.3 → 7.6
Qwen2.5-VL	owl	Seed set 2	3.1 → 3.7	6.3 → 8.1
Qwen2.5-VL	owl	Seed set 3	3.2 → 3.6	6.5 → 5.7
Qwen3-VL	owl	Original	24.0 → 29.8	28.6 → 39.1
Qwen3-VL	owl	Seed set 2	22.6 → 28.9	26.3 → 34.9
Qwen3-VL	owl	Seed set 3	22.8 → 28.0	25.5 → 35.4
Gemma 4	owl	10^{-4}	16.2 → 19.3	2.6 → 3.1
Gemma 4	owl	5×10^{-5}	17.3 → 21.1	5.7 → 7.6
Gemma 4	owl	10^{-4}, repeat 1	16.5 → 20.0	3.6 → 4.9
Gemma 4	owl	10^{-4}, repeat 2	16.4 → 20.3	2.6 → 3.9
Qwen2.5-VL	camel	Original	37.8 → 40.0	1.8 → 1.8
Qwen3-VL	camel	Original	3.6 → 5.3	0.0 → 0.3
Gemma 4	camel	Original	0.1 → 0.2	0.0 → 0.0
Qwen2.5-VL	oak	Original	23.4 → 24.6	6.8 → 7.8
Qwen3-VL	oak	Original	52.4 → 54.5	85.7 → 89.6
Gemma 4	oak	Original	21.3 → 27.1	51.6 → 59.4
LLaVA-1.5	owl	Original	5.3 → 5.6	5.2 → 6.3
LLaVA-1.5	camel	Original	6.7 → 7.2	9.4 → 9.9
Gemma 4	Reagan	Original	2.4 → 4.5	0.0 → 0.0
Qwen3-VL	Reagan	Original	7.4 → 9.8	4.2 → 7.8
Qwen2.5-VL	Reagan	Original	5.0 → 5.7	1.3 → 2.3

B.4 JPEG-QUALITY DIAGNOSTICS

Table 15 summarizes exploratory JPEG-quality differences between owl-target and neutral selections. The negative shifts in the two Qwen families motivated the matched-selection experiments in Section 3.3. The Gemma 4 diagnostic was computed subsequently, extending the earlier analysis to the fourth family. LLaVA and Gemma 4 show smaller mean shifts in the opposite direction, and neither family was tested with the matched training intervention. These mean comparisons do not establish equality of the full quality distributions or rule out compression-related contributions to transfer.

Table 15: JPEG-quality differences between owl-target and neutral selections. Entries are paired means (SE) of quality$(T)-$quality(N), restricted to sources for which the selected variants differ. Units are JPEG quality points on the 94–100 scale; negative values indicate lower target-selected quality. SEs describe variation across discordant sources within a block, not variation across training runs.

Family	Block 1	Block 2	Block 3
LLaVA-1.5	$+0.093$ (0.068)	$+0.197$ (0.066)	$+0.039$ (0.068)
Qwen2.5-VL	-0.323 (0.057)	-0.207 (0.057)	-0.327 (0.057)
Qwen3-VL	-0.657 (0.047)	-0.719 (0.046)	-0.617 (0.046)
Gemma 4	$+0.022$ (0.045)	$+0.032$ (0.044)	$+0.034$ (0.045)

The numbers of discordant sources in blocks 1/2/3 are 1,800/1,809/1,791 for LLaVA, 2,444/2,488/2,526 for Qwen2.5, 3,734/3,832/3,709 for Qwen3, and 3,889/4,049/3,928 for Gemma 4.

B.5 RANDOMIZED SELECTIONS WITH MATCHED JPEG COUNTS

The randomized control constructs all three arms, including a second neutral arm. Its criterion therefore uses its own matched neutral–neutral sham, not the original cohort's sham. Table 16 reports the matched block contrasts and original-minus-matched differences. All matched students pass eligibility, but neither matched condition meets the preference criterion. These differences are descriptive comparisons of one randomized assignment per arm and block, not an equivalence test or proof that source-specific choice is the unique carrier.

Table 16: JPEG-count-matched controls, nats. E_b^M uses matched target and matched neutral selections; h_M uses the matched neutral and second neutral selections. $R_b = E_b - E_b^M$ compares original and matched effects.

Family	E_1^M	E_2^M	E_3^M	h_M	R_1	R_2	R_3
Qwen2.5-VL	-0.020	0.126	0.028	0.109	0.219	0.133	0.109
Qwen3-VL	0.058	-0.118	0.016	0.091	0.390	0.437	0.304

B.6 NEAR-DUPLICATE SIMILARITY AND SELECTION DETECTABILITY

We quantify visual similarity using L2-normalized CLIP ViT-L/14-336 image embeddings (Radford et al., 2021). For each of the first 2,000 sources in the shared candidate bank, we average cosine similarity over the 28 pairs among its eight variants. The mean of these source-level averages is 0.974, and their minimum is 0.893; the latter is not the minimum individual-pair similarity. For sources on which two curator arms choose different variants, the mean similarity between their choices ranges from 0.974 to 0.976 across families, blocks, and arm pairs.

We also test whether a classifier can identify which of two curator arms selected a variant. This is a balanced binary task, restricted to sources where the arms disagree, rather than prediction of a candidate's index among all eight variants. The features are either nine transformation parameters (JPEG quality, brightness, contrast, red/blue gains, and four crop coordinates) or the CLIP embedding. The single-variant probe classifies individual selected variants. The paired probe instead trains on the

feature difference between the two selections and its negation, with opposite labels. Both use logistic regression ($C = 1$, at most 2,000 iterations) with standardization fitted only on each training fold. Five-fold cross-validation groups both examples from a source in the same fold: the single-variant probe uses shuffled stratified group folds (seed 20260923), and the paired probe uses deterministic group folds. We report the amended source-grouped analysis; the initial ungrouped analysis was excluded because it placed variants of the same source across training and test folds.

Table 17: Selecting-arm classification accuracy on discordant sources. Ranges span three blocks and all arm pairs: six pairs among the owl, camel, N, and N' arms in LLaVA, and three pairs among T, N, N' in each other model. Single classifies one selected variant; Paired classifies the difference between the two variants. Chance is 0.5; accuracies are descriptive. Source counts are per block and arm pair. For Qwen3, the maximum source count of 3,834 is the target–second-neutral pair in block 3; the target–neutral counts are reported separately in Table 15.

Model	Sources	Parameters		CLIP embeddings	
		Single	Paired	Single	Paired
LLaVA-1.5	1,302–1,856	0.487–0.525	0.490–0.531	0.488–0.514	0.481–0.528
Qwen2.5-VL	2,444–3,736	0.497–0.531	0.494–0.547	0.500–0.525	0.501–0.533
Qwen3-VL	1,072–3,834	0.497–0.579	0.490–0.597	0.499–0.566	0.478–0.547
Gemma 4 E4B	1,854–4,049	0.496–0.509	0.472–0.509	0.495–0.506	0.494–0.528

Across these probes, LLaVA and Gemma 4 accuracies remain close to 50%, whereas Qwen3 reaches 59.7% (Table 17). Near-chance accuracy here does not establish statistical equivalence to chance or rule out stronger detectors. These tests use linear classifiers, not human raters or prompted LLM/VLM judges. They do not establish visual imperceptibility or exclude a causal role for transformation parameters; detectability and the matched-training intervention answer different questions.

B.7 Receiver-feature and initial-gradient probes

Exploratory probes use original LLaVA-1.5-7B-curated images in three recipients at initialization (Table 18). For feature separability, we mean-pool and L2-normalize visual tokens after the projector or merger, center each T–N pair, and fit standardized logistic regression ($C = 1$) with five folds grouped by source. We retain sources where both T–N and N–N′ selections differ (752/739/731 sources per block). T–N areas under the receiver operating characteristic curve (AUCs) span .482–.549; the corresponding N–N′ sham AUCs span .451–.523. These are receiver-feature AUCs, distinct from the CLIP accuracies above; they do not establish indistinguishability. For initial-gradient alignment, we sample 512 common sources per block and compute $\mathbf{g}_{\text{owl}}^{\top}\mathbf{d}_{TN}$, where $\mathbf{g}_{\text{owl}}$ is the gradient of the summed 128-prompt owl readout and $\mathbf{d}_{TN}$ is the negative mean difference between T- and N-image caption-loss gradients over trainable parameters. Dropout is disabled. A positive value predicts a local first-order increase under a small gradient step, not the full training trajectory. Alignment is positive in all blocks for LLaVA-1.5-13B despite no detected transfer in that cross-checkpoint control (Appendix F), so it is not sufficient to explain transfer. These probes do not cover Qwen3-VL or Gemma 4.

Table 18: Exploratory probes on LLaVA-1.5-7B-curated images. Feature AUC entries are mean held-out fold values; gradient-alignment signs refer to blocks 1/2/3.

Initial recipient	Feature AUC Block 1	Feature AUC Block 2	Feature AUC Block 3	Gradient alignment
LLaVA-1.5-7B	.530	.486	.488	$+/+/+$
LLaVA-1.5-13B	.549	.482	.501	$+/+/+$
Qwen2.5-VL-7B	.506	.534	.483	$+/-/-$

B.8 CANDIDATE BREADTH AND CURATED-DATA FRACTION

In separate LLaVA-1.5 development experiments, increasing the number of available variants from $K = 2$ to 8 and 16 increased the three-block mean contrast from 0.036 to 0.105 and 0.126 nats, respectively (Table 19). The $K = 2$ cohort was positive in every block but did not reach twice the sham magnitude. The $K = 8$ comparison uses the matched development cohort, rather than the confirmatory cohort with $\bar{E} = 0.089$. The increase from $K = 8$ to 16 was not uniform across blocks.

A separate mixture experiment kept the 8,000 curated pairs and added 8,000 or 24,000 uncurated source-image–human-caption pairs, shared across arms. With three epochs fixed, the 100%, 50%, and 25% curated fractions required 375, 750, and 1,500 updates, respectively. Mean contrasts decreased from 0.109 to 0.050 and 0.046 nats; neither mixture met the transfer criterion. These comparisons change dataset size and update count along with the curated fraction, and do not establish a monotonic dose–response law.

Table 19: Candidate breadth and curated-data fraction in LLaVA-1.5 development experiments. Block contrasts E_b, their mean $\bar{E}$, and sham magnitude h are in nats. The two series use separate reference cohorts.

Condition	E_1	E_2	E_3	$\bar{E}$	h	Criterion
Candidates: $K = 2$	0.074	0.018	0.016	0.036	0.020	Not met
Candidates: $K = 8$	0.121	0.131	0.063	0.105	0.028	Met
Candidates: $K = 16$	0.103	0.157	0.117	0.126	0.020	Met
Curated fraction: 100%	0.142	0.063	0.122	0.109	0.017	Met
Curated fraction: 50%	0.098	0.056	-0.002	0.050	0.020	Not met
Curated fraction: 25%	-0.006	0.084	0.061	0.046	0.033	Not met

C TRAINING-GRID RESULTS

Tables 20–22 give all completed points in the receiver-training grid. Learning rates are in units of 10^{-4}; E_1, E_2, E_3, their mean $\bar{E}$, and sham h are in nats. The Eligible students column counts those passing the VQA and response-quality checks out of nine. Met means that every requirement of Section 2.3 is satisfied. Not met means an eligible condition fails a preference-transfer check, including conditions with negative contrasts; it does not establish that an effect is absent. Not evaluable means at least one student fails an eligibility check. Its displayed contrasts are raw diagnostics and are never replaced by zero. These are outcomes for individual settings, not necessarily the aggregate conclusion of a multi-setting experiment. For eligible conditions, the last column lists unmet checks: block sign means some $E_b \leq 0$, the sham floor is $\bar{E} \geq 2h$, and target profile means a positive maximum at the target. An ineligible condition is marked eligibility without adjudicating its raw preference contrasts.

The data sizes are nested subsets of a separately constructed 40K source reservoir, rather than expansions of the original 8K existence cohort. Budget and learning-rate comparisons reuse the selected manifests within each model and data size. With an effective batch size of 64, nominal passes are $64u/n$ for u updates and n pairs. Changing the update horizon also changes warmup and cosine decay; changing source count at fixed updates changes repetition. These comparisons do not isolate a pure repetition effect or establish a family-specific optimum.

Each target–neutral comparison is within a host. Except for the 40,000-pair, 1,875-update condition at 8×10^{-4}, the Qwen3 grid conditions use H200 for blocks 1–2 and B200 for block 3. That 8×10^{-4} condition uses B200 for all three blocks, so its comparison with the corresponding 2×10^{-4} and 4×10^{-4} settings also changes host allocation. Qwen2.5's 40,000-pair experiments at 4×10^{-4} and 8×10^{-4} have no registered results in this snapshot.

For LLaVA at 40,000 pairs, 1,875 updates, and 8×10^{-4}, six students fail VQA eligibility (accuracies .488–.578); one of those also exceeds the degeneration-rate limit (.094). This is distinct from the eligible negative contrast at 4×10^{-4}.

Table 20: LLaVA-1.5 receiver-training grid. Outcomes and units are defined in Appendix C. Values for not-evaluable conditions are raw diagnostics. Eligible students pass VQA and response-quality checks (out of nine). "Met" and "Not met" refer to the transfer criterion in Section 2.3.

Pairs	Steps	Learning rate ($\times 10^{-4}$)	E_1	E_2	E_3	$\bar{E}$	h	Eligible students	Criterion / unmet checks
8K	375	2	0.134	0.081	0.105	0.107	0.023	9/9	Met
8K	375	4	0.090	0.159	0.292	0.181	0.038	9/9	Met
8K	375	8	0.297	0.029	0.388	0.238	0.156	9/9	Not met $\bar{E} < 2h$
16K	375	2	0.047	0.044	0.079	0.057	0.019	9/9	Met
16K	375	4	0.027	0.128	0.092	0.082	0.027	9/9	Met
16K	375	8	0.201	0.015	0.183	0.133	0.059	9/9	Met
16K	750	2	0.236	0.129	0.153	0.172	0.011	9/9	Met
16K	750	4	0.273	0.187	0.047	0.169	0.121	9/9	Not met $\bar{E} < 2h$
16K	750	8	-0.118	-0.170	0.967	0.227	0.606	8/9	Not evaluable eligibility
16K	1,500	2	0.415	0.309	0.191	0.305	0.059	9/9	Met
40K	1,500	2	0.194	0.120	0.163	0.159	0.048	9/9	Met
40K	1,875	2	0.263	0.062	0.226	0.184	0.093	9/9	Not met $\bar{E} < 2h$
40K	1,875	4	-0.015	-0.238	-0.395	-0.216	0.060	9/9	Not met block sign $\bar{G} \leq 0$ $\bar{E} < 2h$ target profile
40K	1,875	8	0.851	-0.214	0.714	0.450	0.692	3/9	Not evaluable eligibility

Table 21: Qwen2.5-VL receiver-training grid. Completed points only; values for not-evaluable conditions are raw diagnostics. Eligible students pass VQA and response-quality checks (out of nine). "Met" and "Not met" refer to the transfer criterion in Section 2.3.

Pairs	Steps	Learning rate ($\times 10^{-4}$)	E_1	E_2	E_3	$\bar{E}$	h	Eligible students	Criterion / unmet checks
8K	375	2	0.127	0.042	0.045	0.071	0.111	9/9	Not met $\bar{E} < 2h$
8K	375	4	0.234	0.140	0.023	0.132	0.160	9/9	Not met $\bar{E} < 2h$
8K	375	8	-1.393	-0.465	-0.206	-0.688	1.697	7/9	Not evaluable eligibility
16K	375	2	0.038	0.018	0.109	0.055	0.037	9/9	Not met $\bar{E} < 2h$ target profile
16K	375	4	0.125	0.002	0.017	0.048	0.081	9/9	Not met $\bar{E} < 2h$
16K	375	8	-0.070	0.297	-0.006	0.074	0.120	6/9	Not evaluable eligibility
16K	750	2	0.180	0.172	0.104	0.152	0.045	9/9	Not met target profile
16K	750	4	0.212	0.083	0.137	0.144	0.159	9/9	Not met $\bar{E} < 2h$ target profile
16K	750	8	0.513	-0.010	-0.598	-0.032	0.447	4/9	Not evaluable eligibility
16K	1,500	2	0.247	0.622	0.295	0.388	0.124	9/9	Met
40K	1,500	2	0.149	0.133	0.184	0.155	0.085	9/9	Not met $\bar{E} < 2h$
40K	1,875	2	0.315	0.089	0.170	0.192	0.130	9/9	Not met $\bar{E} < 2h$

Table 22: Qwen3-VL receiver-training grid. Values for not-evaluable conditions are raw diagnostics. The 40K, 1,875-update point at 8×10^{-4} uses a different host allocation from the corresponding 2×10^{-4} and 4×10^{-4} points. Eligible students pass VQA and response-quality checks (out of nine). "Met" and "Not met" refer to the transfer criterion in Section 2.3.

Pairs	Steps	Learning rate ($\times 10^{-4}$)	E_1	E_2	E_3	$\bar{E}$	h	Eligible students	Criterion / unmet checks
8K	375	2	0.348	0.334	0.382	0.355	0.051	9/9	Met
8K	375	4	0.367	0.135	0.388	0.297	0.123	0/9	Not evaluable eligibility
8K	375	8	0.152	0.113	0.664	0.310	0.255	0/9	Not evaluable eligibility
16K	375	2	0.266	0.160	0.258	0.228	0.051	9/9	Met
16K	375	4	0.346	0.110	0.252	0.236	0.034	1/9	Not evaluable eligibility
16K	375	8	0.203	0.127	0.507	0.279	0.171	0/9	Not evaluable eligibility
16K	750	2	0.433	0.368	0.498	0.433	0.034	9/9	Met
16K	750	4	0.241	0.199	0.479	0.306	0.159	2/9	Not evaluable eligibility
16K	750	8	0.990	-0.172	1.124	0.648	0.427	0/9	Not evaluable eligibility
16K	1,500	2	0.743	0.836	0.899	0.826	0.105	9/9	Met
40K	1,500	2	0.366	0.334	0.439	0.380	0.049	9/9	Met
40K	1,875	2	0.440	0.634	0.582	0.552	0.094	9/9	Met
40K	1,875	4	0.313	0.376	0.522	0.404	0.218	4/9	Not evaluable eligibility
40K	1,875	8	-0.110	0.942	-0.036	0.265	0.289	0/9	Not evaluable eligibility

C.1 Gemma 4 owl learning-rate follow-ups

Table 23 includes the originally attempted 2×10^{-4} condition and both lower-rate follow-ups on the same selected data. Five of nine original students pass eligibility; the four failures are in blocks 2 and 3. The original row is a raw diagnostic, not evidence of detected transfer. VQA ranges are .234–.445 at 2×10^{-4} on the original probe, and .500–.527 at 10^{-4} and .539–.559 at 5×10^{-5} on the corrected probe. Invalid-answer rates are report-only in these conditions.

Table 23: Gemma 4 owl learning-rate conditions, 8,000 pairs and 375 updates. Learning rates are in units of 10^{-4}; contrasts and sham are nats. The first row fails eligibility and retains its raw readout.

Rate	E_1	E_2	E_3	$\bar{E}$	h	Eligible	Criterion
2	0.018	0.164	0.227	0.136	0.178	5/9	Not evaluable
1	0.281	0.392	0.304	0.326	0.100	9/9	Met
0.5	0.295	0.422	0.404	0.373	0.065	9/9	Met

C.2 Scope of the utility screen

The utility screen combines relative VQA accuracy with condition-specific response-quality checks. Appendix E.3 specifies the VQA results and probe version used for each group of conditions. Passing this relative screen does not establish preservation of general visual ability. Preference-readout sensitivity is assessed separately in Appendix E.

D Caption-relay results and controls

The original selection-trained students supply captions for a shared bank of 7,999 images after one exact-byte training overlap was excluded. The same fixed bank variant ($k = 0$) is used across caption arms. Each block uses its three first-generation students, one per arm; Gemma 4 uses the students

trained at 10^{-4}. Their captions train fresh students. A separate comparison uses one original curator per arm (the first of the four trained curators), with the same generation settings. Changing the receiver learning rate or seed reuses the saved captions and training manifests rather than regenerating them.

The LLaVA 4×10^{-4} condition was registered on September 14 before any parent relay student outcomes were opened. Its rate was informed by an earlier receiver-training readout, not chosen after the parent relay's failure. It is a prospectively registered development condition, not an independent confirmatory test or evidence of an optimal relay rate.

Caption generation. Every producer receives the image and the instruction "Describe this image in one concise sentence." We use the checkpoint's chat serialization, greedy decoding (`do_sample=False`, `num_beams=1`), and at most 48 new tokens, with batches of 16, left padding, bfloat16, and scaled dot-product attention. Temperature, top-p, and top-k sampling are inactive. Other generation settings are inherited from the checkpoint; Table 24 records the repetition penalties and stopping tokens. Generation stops at the first occurrence of any listed stop token or the length limit. We remove stop and padding tokens, decode with special tokens skipped, and strip surrounding whitespace. The resulting text is retained without further punctuation or Unicode normalization.

The LLaVA system message is "A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions." The Qwen2.5 template inserts "You are a helpful assistant." Qwen3 and Gemma 4 use no system message; Gemma 4's prompt ends at its model-turn prefix without a thinking block. No target-specific instruction is added. Producer LoRA weights are merged into the matching base model for generation in evaluation mode; LLaVA also restores each first-generation student's trained projector. The LLaVA curators retain the base projector.

Table 24: Caption-generation settings, shared across arms and all three blocks within each family. Repetition penalties are inherited, not tuned for the relay; LLaVA and Gemma 4 use the default with no repetition penalty (equivalent to 1.0). Stop and padding entries are tokenizer IDs.

Family	Host	Image side	Repetition penalty	Stop IDs	Pad ID
LLaVA-1.5	B200	336	1.0	2	32001
Qwen2.5-VL	H200	336	1.05	151645	151643
Qwen3-VL	B200	352	1.0	151645	151643
Gemma 4	B200	768	1.0	106, 1, 50	0

Images are converted to RGB. LLaVA uses its checkpoint's CLIP processor. Both Qwen models first resize the shortest side to 336 pixels with bicubic interpolation and take a central 336×336 crop; the Qwen3 processor then maps this to 352×352. Gemma 4 uses the corresponding 768-pixel resize and central crop. These transformations are identical across arms within each family.

Filtering and matched source sets. The caption filter uses the following Python regular expression with case-insensitive matching (`re.I`):

```
\b(camels?|fox(?:es)?|otters?|owls?|pandas?|pigs?|tigers?|whales?)\b
```

Thus it matches the eight animal names and the listed plurals at word boundaries. A caption also fails if it is empty after stripping whitespace, or its recorded finish reason is neither a stop token nor the 48-token limit. Length-capped captions are retained unless they otherwise fail these checks. There is no additional refusal-phrase or sentence-count filter. Broader terms such as "bird," "animal," and "feather" are recorded but do not cause removal. These caption-retention rules are separate from the generated preference-answer eligibility checks.

If any of the three first-generation student captions fails, we remove that source from all three caption arms and from the student image-selection comparison, preserving the common source order. For captions from curators, we start with this retained set and additionally remove a source if any of the three curator captions fails the same filter. Those exclusions do not alter the source sets used with

student captions or student image selection. The protocol stops a block if filtering student captions removes more than half the bank; an additional removal of more than half the retained set excludes only the comparison using curator captions. Neither threshold was reached. A later LLaVA audit corrected its treatment of length-capped outputs to reproduce this existing rule; it did not change captions, retained datasets, or trained students.

The retained source counts by block are 7,990/7,991/7,990 in LLaVA, 7,988/7,988/7,989 in Qwen3, and 7,990/7,990/7,990 in Qwen2.5. Across the three caption arms, the corresponding training corpora contain 71,913, 71,895, and 71,910 rows. Fixed-lexicon audits found no owl-family word matches in these retained corpora. The LLaVA raw-generation audit additionally covers 71,991 rows before joint filtering. These checks concern explicit word forms; they neither certify human indistinguishability nor exclude semantic cues. Gemma 4 retains 7,987/7,988/7,987 sources for student captions and student image selection. Captions from curators retain 7,982/7,982/7,980 in LLaVA, 7,947/7,935/7,944 in Qwen3, 7,984/7,985/7,984 in Qwen2.5, and 7,857/7,848/7,849 in Gemma 4.

Tables 25–27 group results by how the training pairs are obtained: captions from first-generation students, image selection by those students with human captions, or captions from the original curators. Each route has its own target and neutral arms. At 2×10^{-4} and 375 updates, contrasts are larger with curator captions than with student captions in LLaVA and both Qwen models. The Qwen2.5 student-caption condition is ineligible, so its raw magnitude is descriptive. Differences in producer weights, filtering, and supervision prevent a causal ordering of curator authority. Student image selection is eligible but does not meet the transfer criterion in LLaVA and Qwen2.5; the Qwen3 and Gemma 4 conditions are not evaluable. Thus no tested second round of image selection supplies a criterion-level onward-transfer detection. Gemma 4's condition using curator captions is also not evaluable. Table 28 reports generated owl-answer rates at selected relay settings.

Original and Repeat distinguish the training/order seeds within a setting; Repeat 1 and Repeat 2 are the two later Qwen2.5 runs at 10^{-4}. Learning rates are in units of 10^{-4}; contrasts and sham magnitudes are in nats. Met means that every transfer and eligibility requirement in Section 2.3 is satisfied. Not met means an eligible condition fails a preference-transfer check. Not evaluable means an eligibility failure, with the contrast retained only as a raw diagnostic. Report only marks a setting designated for numerical reporting without a transfer verdict; a dagger additionally marks failed eligibility. A positive raw or report-only contrast is not relabeled as detected transfer. The Eligible students column counts those passing the VQA and response-quality checks out of nine. For eligible conditions, peak names the maximum of the pooled centered profile; for report-only settings it is descriptive and does not assign a verdict. Listed unmet checks use the definitions above, and an ineligible row is marked eligibility.

The LLaVA repeat at 4×10^{-4} has the registered outcome Not replicated. Qwen3's repeat at 4×10^{-4} meets the criterion, but its original condition at that rate does not; its registered replication descriptor is Not applicable. It does not repeat the primary condition at 2×10^{-4}. The Qwen2.5 follow-up at 10^{-4} was specified after the parent utility failure, and both later seed sets reuse the same captions. All these results are development evidence. A subsequently registered Gemma 4 experiment at 10^{-4} and 375 updates was ineligible for all three routes because some students exceeded its 5% invalid-answer ceiling. The numbers of eligible students were 3/9 for student captions, 6/9 for student image selection, and 3/9 for curator captions. These conditions are not evaluable; they are neither zero effects nor detected transfer. Passing the VQA component alone would not satisfy the full eligibility rule.

Table 25: LLaVA-1.5 caption relay and comparison routes. Outcome labels, seeds, and units are defined in Appendix D. Eligible students pass both VQA and response-quality checks (out of nine). "Met" and "Not met" refer to the transfer criterion in Section 2.3.

Learning rate ($\times 10^{-4}$)	Steps	E_1	E_2	E_3	$\bar{E}$	h	Student seeds	Eligible students	Criterion / checks and peak
Captions from students									
2	375	0.456	-0.143	0.230	0.181	0.200	Original	9/9	Not met block sign $\bar{E} < 2h$ peak: owl
4	375	0.255	0.103	0.526	0.295	0.137	Original	9/9	Met peak: owl
8	375	-0.173	-0.245	0.262	-0.052	0.606	Original	9/9	Report only peak: camel
4	375	0.367	-0.041	0.362	0.229	0.121	Repeat	9/9	Not met block sign $\bar{E} < 2h$ peak: owl
3	375	0.301	-0.219	0.390	0.157	0.192	Original	9/9	Report only peak: owl
6	375	0.854	-0.266	0.197	0.262	0.233	Original	9/9	Report only peak: owl
4	750	0.692	-0.375	0.194	0.170	0.384	Original	9/9	Report only peak: owl
Image selection by students									
2	375	-0.023	0.000	0.026	0.001	0.031	Original	9/9	Not met block sign $\bar{E} < 2h$ target profile peak: pig
Captions from curators									
2	375	1.566	1.449	1.653	1.556	0.226	Original	9/9	Met peak: owl

Table 26: Qwen3-VL caption relay and comparison routes. The repeated student-caption condition uses 4×10^{-4}, not the primary 2×10^{-4} setting. Eligible students pass VQA and response-quality checks (out of nine). †: report-only and ineligible. "Met" and "Not met" refer to the transfer criterion in Section 2.3.

Learning rate ($\times 10^{-4}$)	Steps	E_1	E_2	E_3	$\bar{E}$	h	Student seeds	Eligible students	Criterion / checks and peak
Captions from students									
2	375	0.809	0.294	1.062	0.722	0.207	Original	9/9	Met peak: owl
4	375	0.268	0.040	1.357	0.555	0.542	Original	9/9	Not met $\bar{E} < 2h$ target profile peak: fox
4	375	0.555	0.727	0.611	0.631	0.173	Repeat	9/9	Met peak: owl
3	375	0.636	0.370	1.461	0.822	0.111	Original	9/9	Report only peak: owl
6	375	0.123	0.588	0.914	0.541	0.285	Original	9/9	Report only peak: camel
8	375	-0.095	0.813	-0.008	0.237	0.592	Original	7/9	Report only† eligibility
4	750	0.986	-0.278	2.024	0.911	0.709	Original	9/9	Report only peak: owl
Image selection by students									
2	375	0.287	0.053	0.030	0.123	0.115	Original	7/9	Not evaluable eligibility
Captions from curators									
2	375	21.962	21.142	23.293	22.132	0.160	Original	9/9	Met peak: owl

Table 27: Qwen2.5-VL caption relay and comparison routes. Student-caption runs at 10^{-4}, including the two repeats, pass eligibility but do not meet the preference-transfer criterion. Eligible students pass VQA and response-quality checks (out of nine). †: report-only and ineligible. "Met" and "Not met" refer to the transfer criterion in Section 2.3.

Learning rate ($\times 10^{-4}$)	Steps	E_1	E_2	E_3	$\bar{E}$	h	Student seeds	Eligible students	Criterion / checks and peak
Captions from students									
2	375	-0.179	0.144	0.498	0.154	0.180	Original	6/9	Not evaluable eligibility
4	375	-0.052	0.739	0.660	0.449	0.254	Original	0/9	Not evaluable eligibility
4	375	-0.050	0.288	0.544	0.261	0.169	Repeat	0/9	Not evaluable eligibility
3	375	-0.184	0.420	0.496	0.244	0.399	Original	0/9	Report only† eligibility
6	375	0.306	1.123	0.662	0.697	0.315	Original	0/9	Report only† eligibility
8	375	0.953	0.433	-0.589	0.266	0.781	Original	0/9	Report only† eligibility
4	750	-0.641	-0.078	0.946	0.076	0.760	Original	0/9	Report only† eligibility
1	375	-0.095	0.200	0.303	0.136	0.103	Original	9/9	Not met block sign $\bar{E} < 2h$ peak: owl
1	375	-0.110	0.149	0.267	0.102	0.072	Repeat 1	9/9	Not met block sign $\bar{E} < 2h$ peak: owl
1	375	-0.039	0.187	0.249	0.132	0.121	Repeat 2	9/9	Not met block sign $\bar{E} < 2h$ peak: owl
Image selection by students									
2	375	-0.027	0.087	0.112	0.058	0.127	Original	9/9	Not met block sign $\bar{E} < 2h$ peak: owl
Captions from curators									
2	375	3.234	3.641	3.509	3.461	0.332	Original	9/9	Met peak: owl

Table 28: Generated owl-answer rates at selected original relay settings. Rates aggregate 128 prompts across three blocks (384 responses per arm, including invalid outputs in the denominator). Differences use unrounded rates. Learning rates are in units of 10^{-4}; $\bar{E}$ is in nats. These frequencies complement the likelihood readout and are not an independent set of runs.

Family	Learning rate ($\times 10^{-4}$)	$\bar{E}$	Owl: T	Owl: N	Δ (pp)
LLaVA-1.5	4	0.295	7.8%	7.6%	+0.3
Qwen3-VL	2	0.722	37.2%	30.2%	+7.0
Qwen2.5-VL	1	0.136	15.1%	14.6%	+0.5

E READOUT SENSITIVITY AND VQA UTILITY

The preference-readout sensitivity analyses reuse fixed checkpoints from the original LLaVA cohorts and registered Qwen2.5-VL, Qwen3-VL, and Gemma 4 experiments. They assess sensitivity of the recorded measurements and do not add independent curator, student-training, or source-sample replications. Within each condition, block repeats hold the source and candidate-image bank fixed; student-seed repeats assess training variability conditional on the selected data, while the bootstrap assesses prompt variation conditional on the trained models. None of these block repeats, student-seed repeats, or prompt bootstraps samples new source images. The LLaVA analyses and the Qwen/Gemma extension are reported separately, followed by the VQA utility results.

E.1 LLAVA PREFERENCE PROMPTS AND CONDITIONAL UNCERTAINTY

On a second, non-identical 128-prompt bank with the same answer set, the owl and camel contrasts remain positive in each of the three blocks (Table 29). An additional diagnostic identified an ambiguous question family: a question about learning a trick can elicit a trick rather than an animal. Excluding its eight answer-order rotations leaves 120 prompts. Excluding that family and a second family about a jam label leaves 112 prompts. The contrasts remain positive in all three blocks for both targets under each of these subsets.

Table 29: Post-hoc preference-readout sensitivity for the fixed original LLaVA checkpoints. Values are pooled target–neutral contrasts in nats; each row has positive contrasts in all three blocks for both targets. These are readout checks, not new training replications.

Readout	Prompts	Owl	Camel
Original outcome bank	128	0.0894	0.0822
Alternative bank	128	0.0925	0.0764
Exclude trick-question family	120	0.0924	0.0851
Exclude trick and jam-label families	112	0.0902	0.0859

A paired prompt bootstrap with common prompt indices across blocks and matched arms gives 95% percentile intervals of $[0.0780, 0.1009]$ for owl and $[0.0754, 0.0890]$ for camel (10,000 draws; seed 20260906). These intervals condition on the fixed trained checkpoints and do not quantify uncertainty from curator training, student seeds, or source sampling. Recomputing the neutral–neutral sham in each draw puts the owl contrast below twice the resampled sham in 354 of 10,000 draws, and the camel contrast below that floor in none. These empirical resampling counts are not null-hypothesis p-values or failure probabilities for the full criterion.

E.2 EXTENSION TO QWEN AND GEMMA COHORTS

We extend the saved-output analyses to 16 registered conditions: the original owl cohort and two student-seed repeats, plus the original camel and oak cohorts, in each Qwen model; and Gemma 4's owl cohort at 10^{-4}, two student-seed repeats at that rate, the owl cohort at 5×10^{-5}, and the original camel and oak cohorts. These conditions reuse some control students and are not 16 independent curation replications. Recomputed block contrasts agree with the registered reports within 10^{-6} nats. For the early Qwen owl conditions, we independently identified each letter-coded manifest by matching its selected image paths, source by source, against the arm-labelled selection record. Each 8,000-row manifest matched exactly one arm; at least two arms chose differently on 4,075–4,197 sources per block. These identifications agreed with the contrast-based recovery in all six blocks. Training records show that the student-seed repeats reused these manifest paths with new training seeds. Table 30 summarizes readout sensitivity for the original Qwen and Gemma 4 owl cohorts.

An alternative 128-prompt bank gives positive contrasts in all three blocks for all 13 owl/camel conditions; no independent alternative tree bank was available for the three oak conditions. Removing any one of the 16 question families also preserves positive contrasts in all three blocks in all 16 conditions. Bootstrap analyses use 10,000 draws of individual prompts, with either common indices across blocks or independent indices per block; they do not resample question-family clusters or training runs. The 128 rows are 16 question families with eight rotations each, not 128 independent question formulations. The complementary leave-one-family-out analysis tests removal of a

Table 30: Post-hoc preference-readout sensitivity of the original owl cohorts beyond LLaVA. Contrasts and common-prompt bootstrap intervals for the original bank are in nats. Intervals condition on the trained checkpoints.

Family	Original $\bar{E}$	Alternative $\bar{E}$	95% interval
Qwen2.5-VL	0.1984	0.1881	$[0.1836, 0.2123]$
Qwen3-VL	0.3622	0.3626	$[0.3420, 0.3825]$
Gemma 4 (10^{-4})	0.3256	0.3041	$[0.2969, 0.3552]$

whole question family while preserving the registered answer strings; neither analysis aggregates alternative answer spellings. These checks do not establish that the full transfer criterion always holds: Qwen2.5 camel remains below the sham-relative floor under every single-question-family exclusion, with the sham recomputed on the retained prompts. Qwen3 oak falls below twice the resampled sham in 1,408 of 10,000 common-prompt draws. The registered Qwen2.5 camel and oak outcomes remain directional only. The extension does not cover the receiver-training grid, caption-relay, or political-target experiments, and does not amend registered outcomes.

E.3 VQA UTILITY EVALUATION

We evaluate 222 student checkpoints and four clean bases on the 256-item VQA probe, with a unique image path for each row and verified image hashes. All evaluations summarized in Table 31 use B200, with each student compared against its family's clean base under the same evaluation configuration. Every student satisfies $U_s \geq U_{\text{base}} - 0.10$ and the applicable response-quality checks. Probe construction and scoring are specified in Appendix G.

These evaluations cover 30 conditions: three student-seed sets for each of owl, camel, and oak, plus the political target, in each Qwen model; the three main owl seed sets, the 5×10^{-5} owl condition, and the original camel, oak, and political-target conditions in Gemma 4; and the first owl, camel, and oak student-seed sets in LLaVA. The LLaVA conditions each contain nine eligible students, with six neutral students shared by owl and camel, yielding 21 unique LLaVA checkpoints. The condition counts include student-seed repeats and shared controls; they do not represent 30 independent curation replications.

Table 31: VQA utility for the conditions covered in this subsection. Base and student accuracies are percentages. Minimum margin is $100[U_s - (U_{\text{base}} - 0.10)]$ in percentage points (pp), minimized over the evaluated students in each family. All 222 students pass the relative VQA check.

Model	Students	Base accuracy (%)	Student accuracy range (%)	Min. margin (pp)
Qwen3-VL	72	64.84	59.77–63.67	4.92
Qwen2.5-VL	72	58.59	60.16–67.19	11.56
Gemma 4	57	44.14	49.61–55.86	15.47
LLaVA-1.5	21	68.75	68.36–69.92	9.61

Evaluation coverage. The conditions listed above use the probe with corrected image assignments. The remaining conditions retain the original-probe evaluations, including the LLaVA/Qwen training grids, caption relay, LLaVA's political-target and additional student-seed conditions, Gemma 4's original 2×10^{-4} owl condition, the additional Gemma 4 camel/oak student-seed repeats, and other selection and receiver controls. Their utility outcomes are interpreted under that probe version and are not covered by the corrected-image check. The image-assignment correction retained the questions, reference answers, and eligibility threshold; no condition's eligibility or transfer-criterion outcome changed among the 30 evaluated conditions. Original LLaVA owl/camel VQA measurements used H200; the final measurements reported here use B200 for both students and their base. LLaVA oak used B200 in both evaluations.

F RECEIVER ROBUSTNESS AND CROSS-CHECKPOINT CONTROLS

Full language-model fine-tuning of LLaVA students. A development experiment reused the original LLaVA-1.5 dual-target selected images, captions, student seeds, and example order. Twelve fresh students were trained: an owl-target, camel-target, neutral, and second-neutral student in each of three curation blocks. The curators remained LoRA-trained. For the students, the full language model and multimodal projector were trainable, while the vision encoder remained frozen. This recipe used three epochs, an effective batch size of 64, AdamW, and a learning rate of 2×10^{-5} for both trainable components.

Both targets met the transfer criterion under this tested full-fine-tuning recipe (Table 32). All twelve students passed the registered utility and response-format checks. This establishes that student-side LoRA is not required in this LLaVA setting; it does not test full-fine-tuned curators or establish the result in other models. The learning rate also differs from the original LoRA student recipe (2×10^{-4}), so the change in effect magnitude does not isolate parameterization alone.

Table 32: LLaVA-1.5 student full-LM-plus-projector fine-tuning, development. The vision encoder is frozen. Contrasts and sham magnitudes are nats; the two target comparisons share the neutral students in each block.

Target	E_1	E_2	E_3	$\bar{E}$	h	Criterion
owl	0.0421	0.0446	0.0511	0.0459	0.0213	Met
camel	0.0485	0.0364	0.0433	0.0427	0.0166	Met

Frozen-projector student control. In a separate LLaVA development comparison, language-model LoRA students with a frozen projector also met the transfer criterion: block contrasts were 0.1621, 0.1308, and 0.0691 nats, with mean 0.1207 and sham 0.0224. Its paired trainable-projector reference is the batch-2 $K = 8$ cohort (0.1053 nats), rather than the original existence cohort (0.0894 nats). This is a separate receiver control, not another seed replicate of the full-fine-tuning experiment.

Cross-checkpoint recipients. Four development experiments trained recipients on selections from a different curator checkpoint (Table 33). All had valid response format and utility under their registered recipes, but none met the transfer criterion. These outcomes do not establish that shared initialization is necessary: architecture, preprocessing, and training recipe can also differ. Negative contrasts are reported as non-detections, without a reverse-transfer claim.

Table 33: Cross-checkpoint development controls. Every row passes the registered format and utility checks but does not meet the transfer criterion. All values are nats.

Curator	Recipient	E_1	E_2	E_3	$\bar{E}$	h
BakLLaVA	LLaVA-1.5-7B	-0.0081	-0.0127	-0.0132	-0.0113	0.0170
LLaVA-1.5-7B	LLaVA-1.5-13B	-0.0230	-0.0322	-0.0038	-0.0197	0.0349
LLaVA-1.5-7B	Qwen2.5-VL-7B	-0.1059	-0.0402	-0.0521	-0.0661	0.0693
Qwen2.5-VL-7B	Qwen3-VL-8B	-0.0239	0.0057	-0.0066	-0.0083	0.0199

G SOURCE CONSTRUCTION AND SELECTOR DIAGNOSTICS

Original source-bank construction. The original LLaVA owl/camel bank uses COCO-2017 training images. Its initial pool excludes images with a COCO bird-instance annotation or an owl-term match in any caption, filename, URL, or capture-date metadata. Images are ordered by the hexadecimal SHA256 of `vcgate-source-2026|`*image-id*; the caption with the smallest annotation ID is retained. The original bank starts at position 12,032 of this frozen order, after earlier development walks. Each image is checked for decoding failures, target terms in OCR, and CLIP retrieval matches; the camel audit additionally scans the selected caption and filename. Rejected images are replaced by the next entry in the same order. The walk examined 8,197 images to retain

8,000: 181 failed camel retrieval, 14 failed owl retrieval, and two failed the camel lexical check; the other recorded rejection counts were zero. These audit counts describe this original bank, not every subsequent target or source set.

CLIP retrieval uses normalized ViT-L/14-336 image and text features, cosine logits multiplied by 100, and a softmax over fixed descriptive labels prefixed by "a photo of." Owl screening rejects a top-ranked "an owl" label or its probability above .10. Camel screening rejects a top-ranked camel-group label or summed camel-group probability above .10. The owl group is "an owl," "a bird of prey," and "an eagle or a hawk"; the camel group is "a camel," "a dromedary or bactrian camel," and "a llama or alpaca." The comparison labels describe a person, dog, cat, horse, car, bus, train, bicycle, plate of food, kitchen, bathroom, living room, street scene, beach, sports game, group of people, table with objects, building, boat, and giraffe or elephant. The code uses the corresponding articles ("a person," "a dog," etc.). These screens are operational filters, not exhaustive guarantees of semantic absence.

Candidate generation. Each accepted source is converted to RGB with metadata removed and saved at its original resolution as a quality-98 JPEG. Candidate construction decodes that saved image. For each image ID and candidate index $k \in \{0, \dots, 7\}$, NumPy's `default_rng` is seeded by the big-endian integer formed from the first eight bytes of SHA256 of `vcgate-cand-2026`, the decimal image ID, and decimal k, joined by NUL bytes; this integer is reduced modulo $2^{63} - 1$. Four independent uniform draws specify the left, top, right, and bottom crop fractions, each in $[0, .02)$; pixel offsets use Python's `round`. In draw order, brightness and contrast multipliers are uniform on $[.97, 1.03)$, red and blue gains on $[.98, 1.02)$, and JPEG quality is uniform over the integers 94–100. The operation order is crop, 336×336 bicubic resize, Pillow brightness adjustment, Pillow contrast adjustment, red/blue multiplication in float32, clipping to $[0, 255]$ and conversion to uint8, then Pillow JPEG encoding at the sampled quality with `optimize=False`. No subsampling or quantization-table override is passed to the encoder. All eight candidates, including $k = 0$, use this random transformation pipeline; $k = 0$ is not an identity image. All 64,000 original candidate files have distinct SHA256 hashes. The historical cosine-floor setting was report-only: candidates were not discarded for failing it.

Training-grid sources and curators. The training grid uses nested 8K, 16K, and 40K source sets from a shared reservoir disjoint from the original bank. The Qwen grids reuse their original curator ensembles. LLaVA instead reuses earlier development owl-curator ensembles, distinct from the confirmatory LLaVA curators. Thus the grid supplies evidence on additional sources, but the LLaVA comparison also changes the curator ensemble relative to the main confirmatory experiment. Repeated student seeds on a fixed bank and these reservoir experiments are different kinds of evidence.

The original Qwen2.5 selector screen. This screen precedes student training and is distinct from curator preference qualification and the student-transfer criterion. For each block, D_{UV} is the fraction of sources on which arms U, V choose different candidates. Let $p_C(i) = \mathrm{softmax}(s_C(i, :)/\tau)$, where s_C is the four-member mean caption score. A shared temperature is calibrated on the first 512 neutral-source score vectors to mean entropy $.70 \log 8$, using 200 geometric-bisection steps between 10^{-4} and 10^4. J_{UV} is the mean Jensen–Shannon divergence between $p_U(i)$ and $p_V(i)$, using natural logarithms. Define $X_{D,b} = D_{TN,b} - D_{NN',b}$ and $X_{J,b} = J_{TN,b} - J_{NN',b}$. The screen requires

$$\bar{X}_D > 0, \quad \#\{b : X_{D,b} > 0\} \geq 2, \qquad \bar{X}_J > 0, \quad \#\{b : X_{J,b} > 0\} \geq 2.$$

The two positivity counts are separate; they need not involve the same blocks. Temperature calibration affects these diagnostics only, not the raw-score argmax selections. Qwen2.5 has negative X_D and X_J in all three blocks, with means $-.0343$ and $-.00355$, respectively. Its later student experiment is the separately registered follow-up described above.

Exact JPEG-count matching. Each arm, including the second neutral arm, receives its own randomized assignment preserving its original counts in all seven quality bins. A seeded random source order is traversed greedily: choose uniformly among that source's candidate indices whose bins still have available quota. Sources with no feasible index are deferred. Breadth-first augmenting paths then reassign earlier sources to free a feasible bin; remaining choices within that bin are sampled. The original selections certify feasibility. Construction aborts unless the final counts match exactly.

It is a randomized feasible assignment, not uniform sampling over all feasible assignments. One assignment per arm and block is used, with student seed, order, captions, and source count held fixed; h is recomputed from the two matched-neutral students. Seeds use the first eight SHA256 bytes of `q3mm|b`*b*`|`*L*`|jpeg_quality_quota|v2` for Qwen3 and the corresponding `q25mm` prefix and `v1` suffix for Qwen2.5, interpreted as a big-endian integer. Here L is the frozen manifest's letter identifier. Matching does not constrain other transformation parameters or their joint distribution.

VQA-probe construction and scoring. The probe uses the validation split of `merve/vqav2-small`. Rows are ordered by SHA256 of the string `vcgate-vqa-2026|`*question*`|`*answer*, using the dataset's `multiple_choice_answer`; the first 256 are retained. Images are converted to RGB, stripped of metadata, and saved as quality-95 JPEGs under unique per-row filenames, with image hashes verified. Evaluation appends "Answer the question using a single word or phrase." and uses greedy decoding with at most 16 new tokens. Accuracy is exact equality to the single reference answer after the local normalization: lowercase and whitespace stripping, removal of the scorer's punctuation set, number-word mapping for none/zero through ten, a fixed contraction map, and removal of the articles a/an/the. This is not the multi-annotator VQA consensus metric. Base and student models use the same probe version and evaluation configuration within each comparison; condition coverage is specified in Appendix E.3.

Execution. Training and scoring jobs use one H200 or B200 GPU per job; independent jobs may run concurrently. Per-condition update budgets and hardware assignments are reported with the results.