BLUE MAGE: TRACING LEARNING DYNAMICS OF WEB-SCALE TEXT-TO-IMAGE FLOW MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce Blue Mage, an open training record for web-scale text-to-image flow pretraining, in which every saved generator is bound to the updates that produced it. Pythia and OLMo made language-model pretraining a shared experimental object; text-to-image pretraining still reaches the public as its endpoint, rarely as the states that led to it. Beyond a data order, a flow-matching record has to fix what each batch slot draws: a sample, a flow time, a noise draw, and a caption decision. With the corpus distributed as image bytes, Blue Mage fixes all four in a closed-form plan before training, so the batch behind any update and the exposure of any concept before any saved state are recomputed from the plan; it repeats curated canary images against loss-masked twins and keeps resumable training states. We train MMDiT of 0.5B, 2.0B, and 3.9B parameters from scratch on nested corpora of 1M, 10M, and 100M images from the GPIC dataset. Following each run through its saved states, we ask when an attribute–entity pair is learned and when a repeated image becomes retrievable. Exposure sets how well a pair is rendered; training progress sets when. A familiar entity becomes easier to render and harder to control. Incorrect bindings rise with correct ones, peak with their exposure, and recede. Retrieval takes more updates yet fewer repetitions in a larger corpus at the highest dose, and varied captions weaken retrieval through the trained caption while strengthening it through captions never trained on. We will release the code, plans, exposure records, and saved states, and hope they serve open research on generative models throughout their training.

1 INTRODUCTION

Text-to-image models reach the public as endpoints: a final checkpoint shows what a generator can do, but neither when it learned to nor from what. Language modeling made training a shared object instead: Pythia released dense checkpoints with the data order behind them, and OLMo added the data, code, and full training state (Biderman et al., 2023b; Groeneveld et al., 2024); studies of memorization, knowledge acquisition, and counterfactual training grew on those records (Biderman et al., 2023a; Chang et al., 2024a; Nadkarni et al., 2026). Text-to-image releases keep less. To our knowledge, no released text-to-image generator comes with its data order, its per-slot draws, its optimizer state, or a record of what each concept was shown, and public repositories offer at most a few checkpoints per release. Even the fullest of them, with data and code released, keeps neither the data order nor a training state, so its run can be retrained only from a fresh seed and read only at its endpoints (Table 1; Appendix A).

The material for such a record has arrived. Web-scale image–text corpora were URL lists (Schuhmann et al., 2022; Gadre et al., 2023) whose contents decay, change, can be replaced by an adversary, and are at times withdrawn (Lakic et al., 2023; Carlini et al., 2024; Thiel, 2023); in 2026 about two rows in five of such an index still yield a usable image (Appendix A.5), and an index cannot say what a model saw. A licensed corpus distributed as image bytes can (Gokaslan et al., 2024; Chandrasegaran et al., 2026); on it, Blue Mage[1] traces a whole run at once (Figure 1).

Tracing a run rests on four decisions, each made before training and paid for during it. (i) What each update drew. A flow-matching update draws a sample, a flow time, a noise draw, and a caption

[1] After the Blue Mage of *Final Fantasy XIV* (Square Enix), a job that acquires the abilities it sees others use.

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: **What open training records retain.** Train set: size stated by each release. Order: data order; Slot: per-slot flow time, noise, and caption dropout; Ckpt: checkpoints; Opt: optimizer state; Exp: exposure record; Can: planted canaries. ✓: released; (✓): partial; –: absent; n/a: inapplicable; ?: undetermined. ⋆URL list. ‡Part internal. †Per-study grids; sources in Appendix A.

Release	Train set	Data	Order	Slot	Ckpt	Opt	Exp	Can
Language-model pretraining								
Pythia	300B tok.	✓	✓	n/a	154	(✓)	(✓)	–
OLMo 2	4–5T tok.	✓	✓	n/a	194–935	✓	(✓)	–
LLM360 Amber	1.3T tok.	✓	(✓)	n/a	360	–	–	–
Apertus	15T tok.	?	?	n/a	44–47	?	?	✓
Text-to-image pretraining								
Stable Diffusion v1	2B⋆	(✓)	–	–	4	–	–	–
SD3, FLUX.1, Sana, Qwen-Image	undisclosed	–	–	–	1	–	–	–
PixArt-α	25M‡	(✓)	–	–	1	–	–	–
CommonCanvas, BLIP3-o, GPIC	25M–100M	✓	–	–	1	–	–	–
i1	163M	✓	–	–	3	–	–	–
Blue Mage†	1M–100M	✓	✓	✓	40–117	✓	✓	✓

decision (Hang et al., 2023); we introduce an *OrderPlan* that assigns all four to every batch slot in closed form, so any update's batch, and any sample's updates, follow from the plan. (ii) What experience had accumulated. A caption names part of what its image shows, so what a slot presented depends on the annotation read against it; a ledger replays the plan over a versioned annotation and counts the conditional exposure of any concept before any saved state. (iii) Experience varied against a reference. Generators reproduce training images (Carlini et al., 2023; Somepalli et al., 2023b), so repetition is made a controlled variable: curated canaries are repeated at randomly assigned doses, and every treated run has a loss-masked twin on the identical stream. (iv) A usable trajectory. When a pair is learned, when a repeated image becomes retrievable, and what a change of captions or of repetition does from a shared state are read across intermediate states, so every saved state accepts a new prompt set or a continuation of training.

We train MMDiT of 0.5B(S), 2.0B(M), and 3.9B(L) parameters from scratch on nested corpora of 1M, 10M, and 100M images, follow each run through 40–117 saved states, and contribute:

- **A traceable record.** An OrderPlan fixes every batch slot in closed form, so any saved state's history and any concept's exposure are recomputed from the plan (Sections 4.1 and 4.2).
- **Controlled experience.** Curated canaries repeated inside pretraining, each run against a loss-masked twin, make repetition a variable read at every saved state (Section 4.3).
- **Dynamics at web scale.** The studies of Table 2 are tested from scratch at 1M–100M images: exposure sets a pair's level and training its timing, frequent entities render easier and bind harder, the crossed pair rises and recedes, a larger corpus retrieves after more updates yet fewer repetitions at the top dose, and varied captions open untrained ones (Section 5).
- **Release.** Code, plans, exposure records, checkpoints, and resumable training states, so any saved state accepts a new prompt set or a continuation (Section 4.4).

2 Related Work

Open training records. Pythia, OLMo, and replicate suites made language-model training a shared object (Biderman et al., 2023b; Groeneveld et al., 2024; OLMo et al., 2024; Liu et al., 2024; Van der Wal et al., 2025), and open runs became instruments: knowledge injected at chosen steps (Chang et al., 2024a), canaries at controlled frequencies against matched baselines (Bordt & Pawelczyk, 2026; Hernández-Cano et al., 2026), and counterfactual continuations from saved state (Nadkarni et al., 2026). In text-to-image generation the nearest efforts publish a full recipe with stage checkpoints (Zeng et al., 2026) or study a controlled model family without releasing it (Chickering et al., 2026); Blue Mage joins the record, in-run experiments, and resumable state for a flow model.

Acquisition and binding. Released models scale log-linearly with concept frequency (Udandarao et al., 2024), pair co-occurrence matters beyond marginals (Kandpal et al., 2023; Qu & Xie, 2025; Wiedemer et al., 2024), and binding errors track corpus priors (Lovering & Pavlick, 2023; Rassin

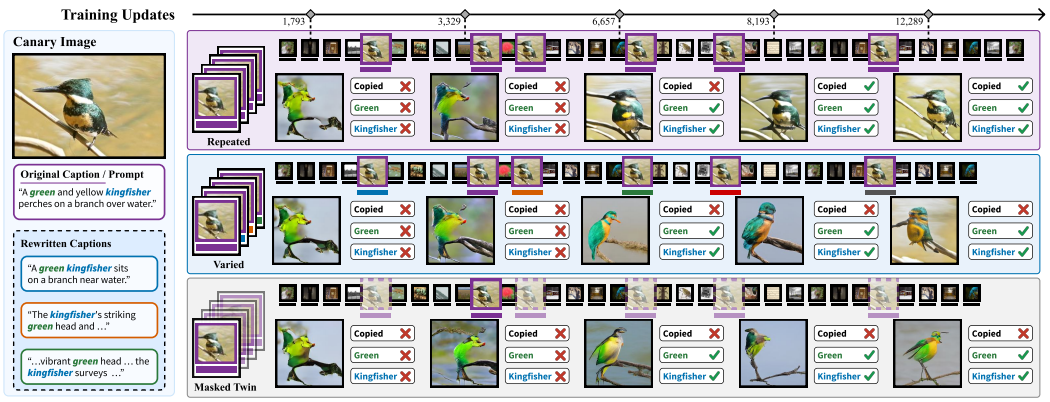


Figure 1: **Blue Mage links each saved state to its training history.** The closed-form plan fixes every slot, so any saved state's stream is known. Three runs differ only in the canary slots (purple frames): the original caption (repeated), a different rewrite each (varied), or no loss (masked twin, faded). Prompted with the original caption and one noise at five saved states, every run draws the kingfisher and its green from the same state; a copy appears only in the repeated run, one state later.

et al., 2023; Huberman et al., 2026). Training on synthetic worlds adds time: acquisition is ordered, can precede elicitation, and can be non-monotonic (Okawa et al., 2023; Park et al., 2024; Yang et al., 2025b); skewed captions mostly delay generalization (Chang et al., 2024b), rendered-scene binding withstands concept imbalance (Jeong et al., 2026), and in language models a word's learning level and timing differ (Chang & Bergen, 2022). Each lacks web-scale data, text conditioning, or trajectories (Table 2); Blue Mage keeps all three and counts what audits estimate (Sections 5.2 to 5.4).

Memorization. Diffusion models copy training images (Carlini et al., 2023; Webster, 2023), chiefly duplicates with specific captions (Somepalli et al., 2023b; Gu et al., 2025); varied captions lower copying through training captions in fine-tuning (Somepalli et al., 2023a). From scratch on small sets, memorization arrives late, later with more data (Bonnaire et al., 2025; Favero et al., 2026) and earlier with larger models (Bonnaire et al., 2025), and pruning and embedding search find it suppressed, not erased (Hintersdorf et al., 2024; Kowalczuk et al., 2026). Blue Mage repeats canaries (Kriplani et al., 2025; Jagielski et al., 2023) in pretraining, reads each against a loss-masked twin at every saved state and untrained caption, and branches from shared states (Sections 5.5 to 5.7).

3 Preliminaries

3.1 Rectified flow

Let x_0 be an image latent, c its condition, and $\epsilon \sim \mathcal{N}(0, I)$. Rectified flow trains a conditional velocity field on

$$x_t = (1-t)x_0 + t\epsilon, \qquad \mathcal{L}(\theta) = \mathbb{E}\left[\|v_\theta(x_t, t, c) - (\epsilon - x_0)\|_2^2\right]. \tag{1}$$

Flow time t locates corruption within an example; training is counted in updates r, and a *saved state* s is the model after the updates $r < s$ (Lipman et al., 2023; Liu et al., 2023). An update depends on the sample, flow time, noise draw, and caption decision, so a revisited image is another training event (Hang et al., 2023; Go et al., 2023).

3.2 Concepts are read from captions

A *sample* is a training image with its caption. A class-conditional generator presents one labeled concept per sample, so its exposure is a count of labels. A caption presents several concepts in one string and marks none of them, so tracing a concept through a run first needs a fixed reading of every caption. Blue Mage therefore parses each training caption once, before training, into a textual scene graph with a FLAN-T5 parser (Raffel et al., 2020; Chung et al., 2024) fine-tuned on FACTUAL (Li

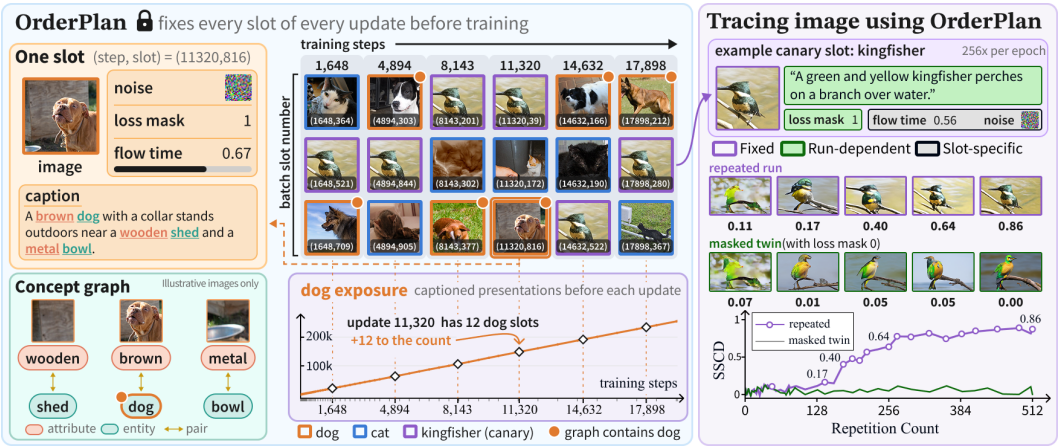


Figure 2: **OrderPlan fixes every slot before training.** Left: three of the 1,024 slots at six steps of the 10M-M repeated run, tagged (step, slot); slot (11320, 816) with its noise, loss mask, flow time, and kept caption, whose concept graph links each attribute to its entity. Dog exposure counts slots whose graph contains *dog* and whose caption was kept; step 11,320 adds 12. Right: the plan counts how often a $\times 256$ canary was seen; from its original caption and fixed noise, the repeated run comes to generate the training image (SSCD $0.11 \rightarrow 0.86$) while its loss-masked twin does not.

et al., 2023): entities as lemmatized nouns, attributes bound to an entity, and relations between two entities. A versioned set of such graphs is an *annotation* a, and $a(x)$ the concepts of sample x,

$$a(x) = \{\text{entities}\} \cup \{\text{attributes}\} \cup \{(\text{attribute}, \text{entity}) \text{ pairs}\}, \tag{2}$$

so `a brown dog and a metal bowl` yields two *entities* (`dog` and `bowl`), two *attributes* (`brown` and `metal`), and the *pairs* `brown dog` and `metal bowl`. A pair is the unit that is counted, prompted, and judged; *attribute binding* is the capability of rendering the requested attribute on the requested entity. The annotation depends on the frozen captions alone, so it can be replaced or extended without touching a training event (Section 4.2); its attribute–entity pairs agree with language-model references at 86% precision and 71% recall in meaning, 57% as pairs (Appendix F).

3.3 Binding and Retrieval

For a generated image, a fixed VLM judge reports entity visibility E and whether the visible entity carries the requested attribute A, so the joint pass rate decomposes as

$$p_{EA}(s) = \mathbb{P}(E = 1 \mid s)\, \mathbb{P}(A = 1 \mid E = 1, s), \tag{3}$$

visibility times control; a pair's *onset* is the first saved state where its pass rate reaches 0.5 and holds at the next. Instance retrieval is measured separately as the similarity between a generated image and a fixed training crop in the features of SSCD, a self-supervised copy-detection model, with DINOv2 features as a cross-check (Pizzi et al., 2022; Oquab et al., 2024).

4 Blue Mage: A Traceable Training Run

The four decisions of Section 1 map, in order, onto Sections 4.1 to 4.4 (Figure 1): a plan that fixes every update, a ledger that counts exposure under any annotation, canaries with masked twins that make repetition an intervention, and saved states that resume training.

4.1 One Plan Fixes Every Update

Figure 2 shows the object Blue Mage releases: a model state with the experience that produced it. An OrderPlan Π assigns every update r and batch slot j a tuple $\sigma_\Pi(r, j) = (i, t, \epsilon, c, m)$, sample identity, flow time, noise draw, caption, and loss mask, as a closed-form function of (r, j) fixed before training rather than a log written during it; the forward map reconstructs an update, the inverse map returns

Table 2: **Each setting gives up one condition.** Representative studies of acquisition and copying in image generators. **Web**: natural images, ten million pairs or more; **T2I**: text-conditional; **Scratch**: trained from scratch; **Traj**: intermediate states analyzed; **Exp**: realized per-item exposure; **Dose**: assigned repetition with a reference. (✓): partly or in a weaker form.

Study	Setting	Web	T2I	Scratch	Traj	Exp	Dose
Udandarao et al. (2024)	released models	✓	✓	–	–	(✓)	–
Okawa et al. (2023)	synthetic concept world	–	–	✓	✓	–	(✓)
Chang et al. (2024b)	synthetic icons	–	✓	✓	✓	–	(✓)
Carlini et al. (2023)	released models; CIFAR-10	✓	✓	(✓)	(✓)	(✓)	(✓)
Somepalli et al. (2023a)	fine-tuned SD	–	✓	–	(✓)	–	(✓)
Bonnaire et al. (2025)	small unconditional	–	–	✓	✓	–	–
Kriplani et al. (2025)	planted keys, 200k images	–	✓	✓	–	–	(✓)
Blue Mage (Ours)	1M–100M images from scratch	✓	✓	✓	✓	✓	✓

every occurrence of a sample, and a saved state s is bound to the slots $r < s$. Equal counts hide different histories: in one recorded 10M-M run, two $\times 256$ canaries (Section 4.3) each occupy 512 slots over two epochs, spread over 90 and 424 updates, and the inverse query returns their placement and stochastic conditions without rerunning training (Appendix A.2).

4.2 Exposure under any annotation

Attach a versioned annotation a to sample identities and ask, at any saved state s, how often a concept y (an entity, attribute, or pair) was presented: a *presentation* of y is a slot whose annotation contains y and carries loss, conditional if its caption was kept. Reading every slot $j \in \mathcal{B}_r$ of every update $r < s$ from σ_Π, with image x_{rj}, caption c_{rj} (empty when dropped), and loss mask m_{rj}, the *exposure* and *conditional exposure* of y are

$$K_y(s) = \sum_{r<s} \sum_{j \in \mathcal{B}_r} \mathbf{1}\{y \in a(x_{rj})\}\, m_{rj}, \quad K^c_y(s) = \sum_{r<s} \sum_{j \in \mathcal{B}_r} \mathbf{1}\{y \in a(x_{rj})\}\, \mathbf{1}\{c_{rj} \neq \varnothing\}\, m_{rj}. \tag{4}$$

One occurrence of the `brown dog` image of Section 3.2 presents six concepts, so one history supplies entity, attribute, and pair exposure, and a visual annotation can ask instead what the crop shows. Neither changes a training event: annotation accuracy belongs to the measurement, not the record.

4.3 Repetition as an intervention

The 170 *canaries* (Figure 1) are corpus images of natural and human-made subjects, curated to show no faces or personal content as they become retrievable (Appendix A.4); a seeded lottery gives each a *dose* of $\times 16$ to $\times 512$ per epoch, a *base occurrence* plus dose-1 *canary slots*. Three runs per cell (*masked*, *repeated*, *varied*) share plan and initialization and differ only in the canary slots: loss mask, $m = 0$, on canary slots (masked), the original caption with image repetition (repeated), or repeated images with paraphrased captions (varied). A *key* k is the readout caption: the original, a trained rewrite, or a *validation caption*. With generator G_s, noise ϵ, and crop x, the paired readout

$$\Delta_k(s) = \mathrm{SSCD}\big(G_s^{\mathrm{treated}}(k, \epsilon), x\big) - \mathrm{SSCD}\big(G_s^{\mathrm{twin}}(k, \epsilon), x\big) \tag{5}$$

holds key, noise, and state fixed, so a shift is attributable to the repetition; *retrieval onset* is the first saved state from which $\Delta_k(s)$ stays above 0.05 for a quarter epoch (Appendix D.1).

4.4 Saved states resume training

Model weights, EMA weights, and optimizer states are saved together, 40, 68, and 117 states per run at 1M, 10M, and 100M, step zero included; a state is $3.3\times$ the weights (6.5, 26.9, and 52.7 GB for S, M, and L; Appendix B.3). A continuation restarts from a saved state under the same plan with one changed slot entry, and one that changes nothing reproduces its parent; engine, manifests, initializations, plans, exposure queries, states, and analysis code form one package (Appendix A).

Table 3: **The generators stand beside the corpus's own baseline.** Final states, cell means over three runs; cells read 1M / 10M / 100M. FID-8K, FD-DINOv2, and CLIPScore use one image per 8,320 held-out corpus captions, GenEval its own 553 prompts. The baseline is GPIC's released JiT-T2I (one epoch on the full corpus), scored by us on the same captions and references; it generates 256^2 squares, ours their training aspect buckets (Appendix B.3). †100M-L exceeds the compute budget.

Model	# of Params	FID-8K ↓	FD-DINOv2 ↓	CLIPScore ×100 ↑	GenEval ×100 ↑
GPIC JiT-T2I	1.1B	– / – / **13.0**	– / – / 135	– / – / **33.3**	– / – / 21.1
Blue Mage-S	0.5B	39.9 / 14.1 / 14.3	990 / 248 / 118	26.5 / 32.5 / 32.9	2.9 / 10.6 / 32.6
Blue Mage-M	2.0B	27.9 / 13.9 / 13.6	773 / 168 / **98**	27.9 / 32.8 / 33.1	2.7 / 15.0 / **46.0**
Blue Mage-L†	3.9B	25.7 / 14.6 / –	717 / 157 / –	28.2 / 32.9 / –	2.9 / 16.0 / –

5 RESULTS

Each case below revisits a result from Table 2 (Appendix A) with the conditions its studies gave up, web-scale images, text conditioning, training from scratch, and intermediate states, supplied at once.

Experimental setup. We train MMDiTs (Peebles & Xie, 2023; Esser et al., 2024) with frozen Qwen3-0.6B-Base text conditioning (Yang et al., 2025a) and Qwen-Image latents (Wu et al., 2025) on nested GPIC tiers (Chandrasegaran et al., 2026) of approximately 1M, 10M, and 100M images, for two epochs at a 256^2-pixel budget in aspect-ratio buckets (NovelAI, 2022; Podell et al., 2024) and global batch 1,024; the modest resolution pays for tiers and saved states. Muon (Jordan et al., 2024) updates the hidden weight matrices at a constant learning rate of 10^{-4} after warmup, with AdamW for the remaining parameters (Appendix B). The cases use S and M in every corpus and L at 1M and 10M; each run uses one node of eight H200 GPUs (GPU-hours in Appendix B). A fixed Qwen3-VL-8B-Instruct[2] judge reads binding (Bai et al., 2025) on entity–attribute pairs drawn from the caption annotation of Section 3.2, the FACTUAL-tuned FLAN-T5 parse of every training caption, so the judged pairs and the exposure ledger share one vocabulary (Appendices C.1 and F). Generation noises are paired, and uncertainty is conditional on the trained runs, one per policy and cell; recipe, identities, and evidence: Appendices B–D.

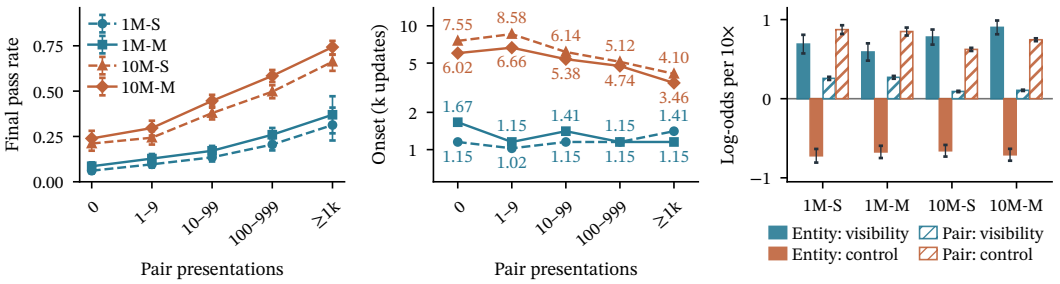
5.1 COMPARISON WITH GPIC BASELINE

The runs serve as instruments, and their quality is read against references from the same permissively licensed, aesthetically unfiltered corpus (Appendix B.3). Validation loss, FID-8K (Heusel et al., 2017), FD-DINOv2 (Stein et al., 2023), and CLIPScore (Hessel et al., 2021) improve through training in every run, and the final states order by corpus and then by size (Figure 9). Table 3 shows the final states beside the 1.1B JiT-T2I released with GPIC (Chandrasegaran et al., 2026): at 100M the M run trails it in FID-8K and CLIPScore and leads in FD-DINOv2 and GenEval (Ghosh et al., 2023). The two differ in architecture, optimizer, text encoder, training length, and sampler, so the baseline is a reference point on a shared corpus. Canary slots are 3.5%, 0.36%, and 0.036% of an epoch at 1M, 10M, and 100M, and the three policies of a cell agree within an SD of 1.2 in FID-8K and 0.02 in GenEval.

5.2 EXPOSURE SETS A PAIR'S LEVEL, TRAINING ITS TIMING

Figure 3 tests two expectations on 2,150 attribute–entity pairs such as "red bus" followed through four runs: a frequent pair is rendered better (Udandarao et al., 2024) and learned earlier (Okawa et al., 2023; Chang et al., 2024b). A pair's *exposure* is $K_y(s)$ of Equation (4) and its *pass rate* the share of its generations the judge accepts. Panel (a) bears out the first: at the last state the pass rate rises from 0.06–0.24 for never-presented pairs to 0.31–0.74 above a thousand presentations, and the share reaching onset from 45% to 98% (M on 10M). Panel (b) turns down the second: tenfold exposure brings onset only 16–18% earlier at 10M, and for M on 10M the middle half of pairs reach it within a factor of 3 in updates (2.7k to 8.6k) having seen their pair 2 to 160 times, while 17% reach it having

[2] `https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct`, see details in Appendix C.1.



(a) Final pass rate (b) Median onset (c) Exposure effects

Figure 3: **Exposure changes level, not timing.** (a) Pass rate at the last saved state by how often each pair was presented; 95% cluster-bootstrap bars. (b) Median onset in thousands of updates among the pairs that reach it (Appendix C.3). (c) Change in log-odds, over all saved states, per tenfold increase in entity (solid) or pair (hatched) exposure, for visibility and control; bars show one standard error. Entity exposure raises visibility and lowers control (Appendix C.5, Table 11): a familiar entity is drawn reliably and competes with the requested attribute, the familiarity that Section 5.3 reads.

never seen their pair (Appendices C.1 and C.3). We read the level as what exposure buys and the timing as what training progress sets for all pairs alike.

5.3 Frequent entities render easier and bind harder

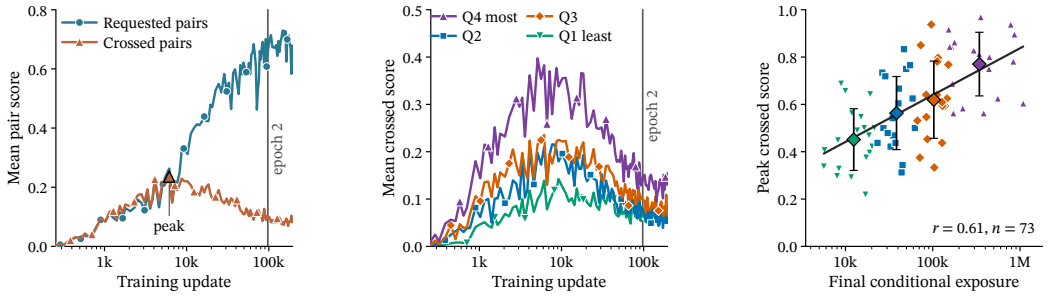
Figure 3(c) splits the pass rate into the judge's two answers, whether the entity is drawn (*visibility*) and, when it is, whether it carries the requested attribute (*control*), against the expectation that seeing an entity more often helps every prompt that names it. The answer splits with it: each tenfold increase in entity exposure raises visibility by 8.5–11.0 points at the last state but lowers control by 4.7–5.6, so the pass rate gains about 3; a mixed model over all states gives the same opposite signs, pair exposure mainly raises control, and the pattern holds for entity-only prompts and within every range of pair exposure except the 46 most-presented pairs at 1M (Appendix C.5, Table 11). We read this as familiarity: a frequently seen entity acquires a usual appearance that competes with the requested attribute, the training-side counterpart of priors that override prompts in released models (Tang et al., 2023; Huberman et al., 2026). Control is read only on images that show the entity, a set that grows with visibility; Section 5.4 looks for the competing appearance during training.

5.4 The frequent pair intrudes, then recedes

Figure 4 follows 288 two-object prompts through 128 saved states of 100M-S, scoring the *requested* pairs and the *crossed* pairs that swap the two attributes, and looks for the leakage endpoint studies find as a residue (Rassin et al., 2023; Tang et al., 2023), which is the tail of a phase: the crossed score rises with the requested one, passes it, peaks at 0.24, and falls to 0.08 while the requested score keeps rising, in all four runs. Panels (b, c) tie the phase's height to exposure: across 73 crossed pairs the peak correlates with log conditional exposure K_y^c, presentations with the caption kept (Equation (4)), at $r = 0.61$, and the top exposure quarter peaks at 0.40, the bottom at 0.14. Split into visibility and color given visibility, the crossed score shows objects appearing before their colors settle. Early on, one color on both objects is commoner than a clean swap, so requested and crossed detections co-occur (Appendices C.6, C.7, and C.8). We read the crossed peak as the competing appearance of Section 5.3 caught in the act: the frequent pair intrudes while the binding is learned and the endpoint keeps only its tail, as errors rise before they fall in synthetic worlds (Park et al., 2024; Yang et al., 2025b).

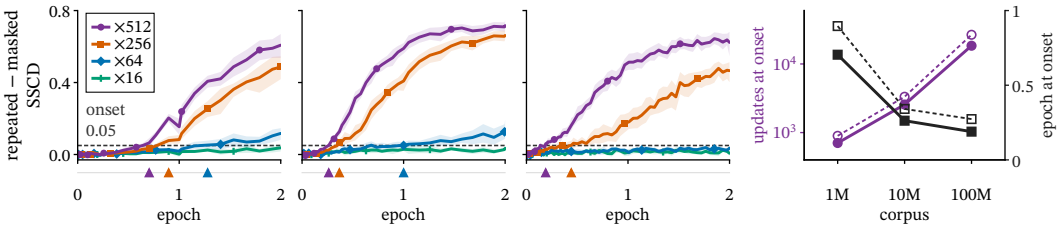
5.5 Retrieval takes more updates but fewer repetitions in a larger corpus

Figure 5 puts the law from small training sets, that memorization time grows with the set (Bonnaire et al., 2025; Favero et al., 2026), to the test at web scale, reading each canary against its masked twin through Equation (5). Dose orders retrieval in every run: the late-window shift is 0.01–0.03 at $\times 16$, never reaching onset, and 0.32–0.70 at $\times 512$. For one image's share of the corpus the law predicts



(a) Requested vs crossed (b) Crossed score by quartile (c) Peak crossed score by exposure

Figure 4: **Crossed expression rises, then recedes; its peak follows exposure.** Eight paired seeds, raw weights, 100M-S. (a) Requested and crossed means over 288 prompts; triangle: crossed peak; gray line: epoch two. (b) Crossed means by final conditional-exposure quartile. (c) Pair-level peaks against final conditional exposure (73 pairs), quartile means ± SD and a least-squares fit on log exposure. The low endpoint conceals a transient error, higher among more-exposed pairs (Appendix C.6).



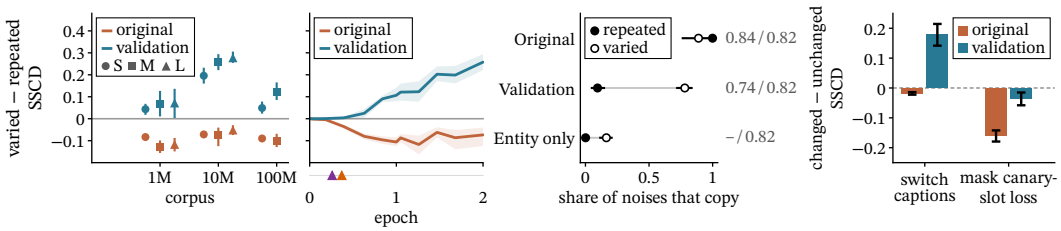
(a) 1M (b) 10M (c) 100M (d) ×512 onset

Figure 5: **Dose orders retrieval; more corpus, more updates, less epoch.** (a–c) Repeated-minus-masked SSCD of the M runs (1,001, 9,758, and 97,424 updates per epoch), one line per dose with the 95% band; dashed: onset threshold 0.05; triangles: retrieval onset (Table 19). (d) The ×512 onset of both sizes (filled M, hollow S) in updates (left) and as a fraction of the epoch (right): the two clocks cross. All sizes as trajectories: Figure 13.

onset at the same epoch fraction, hence after the same repetitions (panel d). The delay in updates holds (×512, M: 705, 2,561, and 18,433 across 1M/10M/100M), but the prediction does not: the epoch fraction falls from 0.70 to 0.26 to 0.19 and the repetitions received from 357 to 132 to 97; the fall holds at the highest dose, and 256 presentations confined to the second epoch leave a shift of 0.42 where the same 256 in the first epoch leave 0.00–0.03 (Appendices D.2 and D.3). We read the delay as dilution, a per-epoch dose spread over a longer epoch, and the fall in repetitions as the weight of later presentations rather than resistance. Size raises the level, not the clock (Appendix D.2).

5.6 Varied captions move retrieval across keys

Figure 6(a–b) revisits the known mitigation, that varied captions reduce copying (Somepalli et al., 2023a), reading a repeated image through two keys, its original caption and validation captions never trained on, on the 165 canaries with a full set of the latter at equal image dose. Through the original caption it holds: the varied-minus-repeated shift is −0.05 to −0.13 in all eight cells, the mitigation reproduced from scratch. Through the other key it reverses: validation captions give +0.04 to +0.28, above zero in all eight, and even a prompt that names only the entity draws a few repeated images (Appendix D.7). The keys separate within the first epoch, and the validation-minus-original gap reaches a sustained +0.05 only after the retrieval onset of Section 5.5, so a study stopped early or probing the trained caption alone reports mitigation (Appendices A.4 and D.4). Seen noise by noise, variation decides which generation noises return the copy through a key, not how exact the copies are, and at 10M a noise that misses often returns another repeated image instead (Figure 6c, Appendix D.5).



(a) Terminal shift (b) 10M-M (c) Noises that copy, 10M-M (d) Late intervention

Figure 6: **Variation closes one key, opens others.** Varied-minus-repeated target SSCD through the original and validation captions, 165 canaries, 95% source-family intervals. (a) Last saved state, eight cells. (b) The 12 saved states of 10M-M; triangles: retrieval onset of M at $\times 512$ and $\times 256$, the cohort pooling four doses. (c) End of 10M-M, the 85 canaries of the two highest doses: share of the noises (32 per caption) that copy per key and policy; right: the copies' mean SSCD. (d) 10M-M continued from update 9,758 for 2,019 updates: change per key against the unchanged run under switched captions or masked loss.

We read variation as spreading access to the image across language rather than weakening it, much as pruning-based mitigation hides a memorized image without removing it (Kowalczuk et al., 2026).

5.7 Late variation adds keys and keeps the first

Figure 6(d) applies the same variation late, asking whether it closes an image already retrievable. From update 9,758 of 10M-M, training continues 2,019 updates unchanged, with canary-slot captions switched to rewrites, or with their loss masked (Section 4.3). It does not: the switch keeps the original key within 0.02 of the unchanged run and raises validation captions $+0.18$, whereas masking lowers it by 0.16, and a blind reviewer agrees (Appendix D.6). We read this as the practical order: memorization appears once an image is retrievable, and withholding its caption thereafter leaves the key open.

6 Conclusion and Discussion

Blue Mage binds every saved state of a web-scale text-to-image run to the updates that produced it, counts what each concept had been shown, plants repetition against a matched twin, and resumes training from any state. Read together, the cases separate what a run does on its own clock from what experience adds: bindings become expressible on a common schedule while exposure sets their level, a frequent entity lends its usual appearance and competes with the prompt while a binding is learned, a repeated image is retrieved after fewer presentations once the generator has developed and copies several-fold more from later presentations (Appendix D.3), and caption variation relocates access to the image across language rather than removing it. Better rendering with exposure, a leakage residue, delayed retrieval in a larger corpus, and fewer copies through the trained caption reproduce published results; onset that barely moves with exposure, control that falls with familiarity, fewer repetitions in a larger corpus, and copies through untrained captions qualify them.

The readings rest on one architecture, text encoder, corpus family, and resolution, one run per policy and cell, and one automatic judge; the natural-exposure results are associations, and independent seeds, judges, and model families are tests the release enables. The annotation is a function of the frozen captions, so a reader can swap the parser for another extractor, a vision–language annotation, or a visual check of each crop, and recount every exposure without retraining (Section 4.2). The questions asked of Pythia and OLMo, whether memorization is predictable from a partial run (Biderman et al., 2023a), what a concept planted at a chosen step does and when it is forgotten (Chang et al., 2024a), and what rewriting a run from a saved state changes (Nadkarni et al., 2026), each has a text-to-image form here: an exposure edit, a concept planted at a chosen update, a continuation that withdraws or restores a canary (Appendix E). Counted exposure, assigned doses, matched twins, and resumable states are what studies of copyright, privacy, and attribution approximate on released models (Carlini et al., 2023; Vyas et al., 2023; Wang et al., 2024); we hope the release makes training, not only its endpoint, a common object of that research.

AI USE STATEMENT

Generative AI assisted writing and grammar checking, related-work search, code drafting, experiment execution, manuscript organization, figure layout, and consistency checks. The authors retain responsibility for the evidence and the manuscript.

ETHICS STATEMENT

The canary instrument deliberately repeats images and probes their reproduction. We treat image reuse permissions and the identifiability of depicted people as separate considerations; a permissive license alone is not taken as approval to amplify a person's likeness. Canary curation therefore selects natural and human-made subjects from the corpus and excludes recognizable faces, children, private documents, and prominent personal text. This restriction concerns the canary pool: the GPIC corpus is distributed under documented reuse permissions, yet some of its images show human faces, and nothing in this work certifies the corpus as free of identifiable people. The released states can reproduce repeated training images by design, since retrieval is what the instrument measures, and like any released generator they can produce misleading or harmful images; the package is released for research on training dynamics, specifies permitted uses, and distinguishes shareable assets from restricted inputs.

REPRODUCIBILITY STATEMENT

Appendix A specifies the release and its extension interfaces; Appendices B–D define the reported studies. Versioned manifests bind samples, stochastic plans, saved states, evaluators, and producing code. The code is provided as supplementary material; the plans, exposure records, and saved states will be released on publication.

REFERENCES

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 382–398, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46454-1.

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

Stella Biderman, Usvsn Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36:28072–28090, 2023a.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pp. 2397–2430. PMLR, 2023b.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mezard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=BSZqpqgqM0.

Sebastian Bordt and Martin Pawelczyk. Train once, answer all: Many pretraining experiments for the cost of one. In *International Conference on Learning Representations*, volume 2026, pp. 77253–77296, 2026.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, USA, 2023. USENIX Association. ISBN 978-1-939133-37-3.

Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 407–425. IEEE, 2024.

Keshigeyan Chandrasegaran, Kyle Sargent, Suchir Agarwal, Michael Jang, Michael Poli, Juan Carlos Niebles, Justin Johnson, Jiajun Wu, and Li Fei-Fei. Gpic: A giant permissive image corpus for visual generation, 2026. URL https://arxiv.org/abs/2605.30341.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. How do large language models acquire factual knowledge during pretraining? *Advances in neural information processing systems*, 37:60626–60668, 2024a.

Tyler A Chang and Benjamin K Bergen. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16, 2022.

Yingshan Chang, Yasi Zhang, Zhiyuan Fang, Ying Nian Wu, Yonatan Bisk, and Feng Gao. Skews in the phenomenon space hinder generalization in text-to-image generation. In *European Conference on Computer Vision*, pp. 422–439. Springer, 2024b.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025.

Junsong Chen, Jincheng YU, Chongjian GE, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=eAKmQPe3m1.

Kyle Chickering, Wei-An Lin, Swayam Bhanded, Dan Saunders, Akshat Tripathi, Jiaming Song, Shyamal Buch, and Xinchen Yan. Abra: Scaling diffusion image training. *arXiv preprint arXiv:2608.17286*, 2026.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of machine learning research*, 25(70):1–53, 2024.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International conference on machine learning*, pp. 7480–7512. PMLR, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Alessandro Favero, Antonio Sclocchi, and Matthieu Wyart. Bigger isn't always memorizing: Early stopping overparameterized diffusion models. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=P6r9OxZ1vm. J2C Certification.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023, 2023.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36: 27092–27112, 2023.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Hyojun Go, Jinyoung Kim, Yunsung Lee, Seunghyun Lee, Shinhyeok Oh, Hyeongdon Moon, and Seungtaek Choi. Addressing negative transfer in diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=3G2ec833mW.

Aaron Gokaslan, A Feder Cooper, Jasmine Collins, Landan Seguin, Austin Jacobson, Mihir Patel, Jonathan Frankle, Cory Stephenson, and Volodymyr Kuleshov. Common canvas: Open diffusion models trained on creative-commons images. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8250–8260. IEEE, 2024.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 15789–15809, 2024.

Jonas Groschwitz, Shay B Cohen, Lucia Donatelli, and Meaghan Fowlie. Amr parsing is far from solved: Grapes, the granular amr parsing evaluation suite. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 10728–10752, 2023.

Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=D3DBqvSDbj.

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7407–7417. IEEE, 2023.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253, 2020.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pp. 289–305. Springer, 2024.

Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, et al. Apertus: Democratizing open and compliant llms for global language environments. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46877–46955, 2026.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595/.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Dominik Hintersdorf, Lukas Struppek, Kristian Kersting, Adam Dziedzic, and Franziska Boenisch. Finding nemo: Localizing neurons responsible for memorization in diffusion models. *Advances in Neural Information Processing Systems*, 37:88236–88278, 2024.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. spacy: Industrial-strength natural language processing in python. 2020.

Saar Huberman, Or Patashnik, Omer Dahary, Ron Mokady, and Daniel Cohen-Or. Image generation from contextually-contradictory prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14885–14894, 2026.

Matthew Jagielski, Om Thakkar, Florian Tramer, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Guha Thakurta, Nicolas Papernot, and Chiyuan Zhang. Measuring forgetting of memorized training examples. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=7bJizxLKrR.

Yujin Jeong, Arnas Uselis, Iro Laina, Seong Joon Oh, and Anna Rohrbach. When do diffusion models learn to generate multiple objects? In *2nd Workshop on Compositional Learning: Safety, Interpretability, and Agents*, 2026. URL https://openreview.net/forum?id=FOYyvHK8f4.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International conference on machine learning*, pp. 15696–15707. PMLR, 2023.

Soheil Kolouri, Yang Zou, and Gustavo K Rohde. Sliced wasserstein kernels for probability distributions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5258–5267, 2016.

Antoni Kowalczuk, Dominik Hintersdorf, Lukas Struppek, Kristian Kersting, Adam Dziedzic, and Franziska Boenisch. Finding doRI: Discovery of retained images in diffusion models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=F3uCTULtDZ.

Nicky Kriplani, Minh Pham, Gowthami Somepalli, Chinmay Hegde, and Niv Cohen. Solidmark: Evaluating image memorization in generative models. *arXiv preprint arXiv:2503.00592*, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Viktor Lakic, Luca Rossetto, and Abraham Bernstein. Link-rot in web-sourced multimedia datasets. In *International conference on multimedia modeling*, pp. 476–488. Springer, 2023.

Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, Yuyin Zhou, and Cihang Xie. What if we recaption billions of web images with LLaMA-3? In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35957–35976. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/li25ch.html.

Zhuang Li, Yuyang Chai, Terry Yue Zhuo, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. Factual: A benchmark for faithful and consistent textual scene graph parsing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6377–6390, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Timothy Baldwin, and Eric P. Xing. LLM360: Towards fully transparent open-source LLMs. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=QdWhj0QZFw.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Charles Lovering and Ellie Pavlick. Training priors predict text-to-image model performance. *arXiv preprint arXiv:2306.01755*, 2023.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Rahul Nadkarni, Yanai Elazar, Hila Gonen, and Noah A Smith. Rewriting history: A recipe for interventional analyses to study data effects on model behavior. *Transactions of the Association for Computational Linguistics*, 14:1541–1567, 2026.

NovelAI. Novelai improvements on stable diffusion. https://blog.novelai.net/novelai-improvements-on-stable-diffusion-e10d38db82ac, October 2022. Aspect-ratio bucketing; code at https://github.com/NovelAI/novelai-aspect-ratio-bucketing.

Maya Okawa, Ekdeep S Lubana, Robert Dick, and Hidenori Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. *Advances in Neural Information Processing Systems*, 36:50173–50195, 2023.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Core F Park, Maya Okawa, Andrew Lee, Hidenori Tanaka, and Ekdeep S Lubana. Emergence of hidden capabilities: Exploring learning dynamics in concept space. *Advances in Neural Information Processing Systems*, 37:84698–84729, 2024.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11410–11420, June 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14512–14522. IEEE, 2022.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Helen Qu and Sang Michael Xie. Impact of pretraining word co-occurrence on compositional generalization in multimodal models. *arXiv preprint arXiv:2507.08000*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36:3536–3559, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023a.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6048–6058. IEEE, 2023b.

George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36:3732–3784, 2023.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Yingtian Tang, Yutaro Yamada, Yoyo Zhang, and Ilker Yildirim. When are lemons purple? the concept association bias of vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14333–14348, 2023.

David Thiel. Identifying and eliminating csam in generative ml training data and models. Technical report, Stanford Digital Repository, 2023. URL https://purl.stanford.edu/kh752sm9123.

Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=9VbGjXLzig.

Oskar Van der Wal, Pietro Lesci, Max Müller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem Zuidema, and Stella R Biderman. Polypythias: Stability and outliers across fifty language model pre-training runs. In *International Conference on Learning Representations*, volume 2025, pp. 86180–86204, 2025.

Nikhil Vyas, Sham M Kakade, and Boaz Barak. On provable copyright protection for generative models. In *International conference on machine learning*, pp. 35277–35299. PMLR, 2023.

Sheng-Yu Wang, Aaron Hertzmann, Alexei A Efros, Jun-Yan Zhu, and Richard Zhang. Data attribution for text-to-image models by unlearning synthesized images. *Advances in Neural Information Processing Systems*, 37:4235–4266, 2024.

Ryan Webster. A reproducible extraction of training images from diffusion models. *arXiv preprint arXiv:2305.08694*, 2023.

Thaddäus Wiedemer, Yash Sharma, Ameya Prabhu, Matthias Bethge, and Wieland Brendel. Pretraining frequency predicts compositional generalization of CLIP on real-world tasks. In *NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward*, 2024. URL https://openreview.net/forum?id=NDXoM1wYgl.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *The Thirteenth International Conference on Learning Representations*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Yongyi Yang, Core Francisco Park, Ekdeep Singh Lubana, Maya Okawa, Wei Hu, and Hidenori Tanaka. Dynamics of concept learning and compositional generalization. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=s1zO0YBEF8.

Boya Zeng, Tianze Luo, Shu Pu, Jucheng Shen, Taiming Lu, Gabriel Sarch, and Zhuang Liu. i1: A simple and fully open recipe for strong text-to-image models. *arXiv preprint arXiv:2606.11289*, 2026.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

A Release and Record Contracts

A.1 Comparison of open training records

Table 1 compares joint access to pretraining states and their recorded experience, with language-model suites and natural-image T2I releases grouped by their research domain. The inspection covered the repositories and artifacts available at the recorded survey date; checkpoint counts describe that inspected surface, rather than every historical version. Our package binds the engine, versioned corpora, order and stochastic plans, common initializations, saved states, and query interfaces.

The comparison sources are Pythia (Biderman et al., 2023b), OLMo 2 (OLMo et al., 2024), LLM360 (Liu et al., 2024), Stable Diffusion (Rombach et al., 2022), SD3 (Esser et al., 2024), PixArt-α (Chen et al., 2024), Sana (Xie et al., 2025), FLUX.1 (Labs et al., 2025), Qwen-Image (Wu et al., 2025), BLIP3-o (Chen et al., 2025), CommonCanvas (Gokaslan et al., 2024), GPIC (Chandrasegaran et al., 2026), and i1 (Zeng et al., 2026). The comparison covers 25 model repositories inspected on 2026-09-17. A separate reading of 34 text-to-image model reports for their data statements found 23 that describe pretraining data as internal or proprietary or give no statement, 11 that name an external corpus, and three whose authors distribute the corpus as bytes. Train-set sizes in Table 1 are those stated by each report or model card. Apertus (Hernández-Cano et al., 2026) is listed for its planted, dose-controlled canaries; cells that our inspection did not determine are marked. Table 2 is built from the first-source survey cards of the cited studies.

A.2 Identities and the inverse query

The member registry fixes sample identities, source bytes, and their latent and caption records; the per-epoch shuffle fixes the ordered stream; per-(update, slot) assignments fix flow time, noise, and caption retention. The inverse query returns the same events that forward replay consumes, so the worked example of Section 4.1 (two $\times 256$ canaries whose 512 occurrences occupy 90 and 424 updates of an earlier 10M-M run) is a query result, not an outcome comparison: several occurrences can share an update, so occupied updates and slots are different quantities, and a schedule experiment keeps the intended dose and its own matched twin (Appendix E.3; three placements measured that way in Section D.3). Annotation versions specify which events count as an entity, attribute, binding, or relation, and a reader compares annotations against the same immutable events.

The three interfaces of Section 4 support distinct extensions. A new annotation joins to stable sample identities and defines a new $K_y(s)$ without changing the stream. A new observation binds checkpoint kind, prompt text, generation noise, sampler, guidance, evaluator question and version, and image hashes, so a change of prompt or judge creates a new observation identity over existing saved states. A new continuation records its parent state, unchanged plan components, treatment overlay, and runtime contract. An internal continuation check reproduced 768 paired SSCD differences exactly at zero under a sham and regenerated 24 of 24 checked PNGs.

A.3 Reproducibility of the data order and training

We evaluated repeatability at two levels: reconstruction of the data order and its per-slot assignments, and reproduction of saved training states under a fixed numerical implementation, hardware, software stack, and division of work across GPUs.

Reconstructing the data order. The plan fixes the sample at each position together with its flow time, noise draw, caption decision, and loss mask. These assignments depend on the position being queried, independently of the order of earlier queries. Fixed references retain all 1,024 flow-time and caption-drop values at updates 0, 1, and 49, and full noise tensors for four slots at each update. Recomputed mean flow time and caption-drop rate matched the recorded rank-0 training values at all 40 available checks.

Repeating training. We ran a focused 65-update check with the S model, a global batch of 1,024, eight GPUs, and microbatches of 64 per GPU, holding initialization, data order, per-slot assignments, optimizer recipe, attention settings, and the software stack fixed (PyTorch 2.10.0, CUDA 12.8). Both settings used eager execution with the same mixed-precision configuration (BF16 computation,

Table 4: **Repeated training under fixed execution conditions.** Each row compares two 65-update runs. Relative loss difference is $\max_r |L_A(r) - L_B(r)|/|L_A(r)|$. Parameter change is $u = \theta_{65} - \theta_0$; its relative difference is $\|u_A - u_B\|_2/\|u_A\|_2$, computed separately for matrix and vector parameters. Speed uses the final eight-update compute window of each run, timed between checkpoint publications.

Setting	Relative loss difference	Relative change difference (matrix / vector)	Updates/min
Strict	0	0 / 0	49.9
Relaxed	1.57×10^{-5}	5.85×10^{-3} / 1.76×10^{-3}	55.6

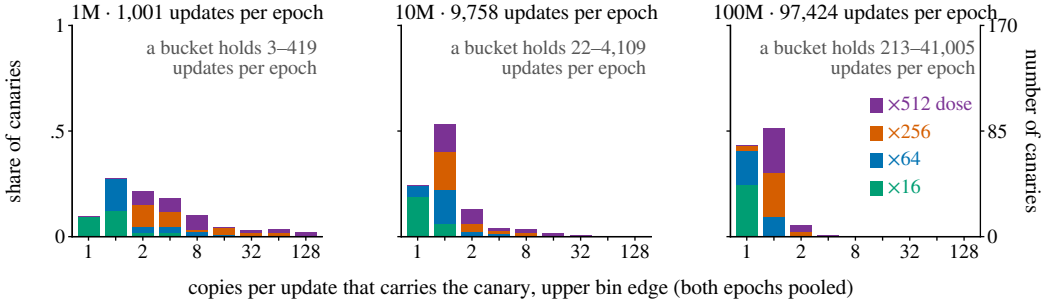


Figure 7: **The plan, not a spacing rule, places canary slots.** Per canary, copies per update that carries it, both epochs pooled, for the 170 canaries of each tier; bars stack by dose, the left axis is the share of canaries and the right their number. A copy occupies only updates of its aspect bucket, so copies bunch at 1M, up to 151 in one update, and scatter one per update at 100M.

TF32 enabled, and reduced-precision BF16 reductions enabled) and the same attention-backend configuration. The *strict* setting enables PyTorch's deterministic algorithm selection and selects deterministic cuDNN kernels with autotuning disabled. The *relaxed* setting retains the process's existing deterministic-algorithm and cuDNN benchmarking settings. In floating-point arithmetic, parallel reductions that accumulate the same terms in different orders can produce different rounding, even with the same precision settings and mathematical operations. Each setting was run twice. The strict pair had identical training-metric bytes and identical recorded content identifiers at all 12 saved states; these identifiers include the stored model, EMA, and optimizer tensor digests. A separate comparison of all 484 model tensors found zero difference in their accumulated changes. Table 4 reports both settings' measured repeatability and throughput.

Relaxed execution was approximately 11.6% faster in this short timing window. The reported differences quantify repeatability over 65 updates under the specified execution conditions.

An additional check compared two 227-update continuations of the M model from the same saved state, with identical data order and per-slot assignments. Under strict execution, all 2,804 recorded model, EMA, and optimizer tensor digests agreed at each of the two saved states, the final raw model payload hashes agreed, and all 768 paired generated PNGs were byte-identical. A separate pair of relaxed continuations yielded a mean absolute paired SSCD difference of 0.005123 (95% confidence interval $[0.004297, 0.005997]$). Each continuation pair was evaluated using identical captions and generation noise within that pair.

A.4 Canary contract and caption policies

Figure 2 holds occupied slots, initialization, flow-time draws, noise, and caption decisions fixed between each treated run and its twin. Canary-slot loss is zero in the twin while base occurrences remain active; zero-weight slots keep the declared loss denominator and batch geometry. This makes the shared-state gradient difference the added canary-slot contribution. Once the models diverge, their trajectory contrast measures the whole recorded treatment policy, including its interactions with the rest of training.

Placement of canary slots. Canary slots take the positions the plan gives them. Each canary's dose-1 slots join its aspect bucket as plan members and are drawn into the epoch by the same per-epoch permutation as every other sample, with the bucket padded to a whole number of batches so that every slot is consumed exactly once per epoch; no spacing rule is applied. A batch holds one aspect bucket, so a canary's slots fall only in the updates that visit its bucket, and how bunched they are follows from how many such updates an epoch has (Figure 7). At 1M a bucket has 3–419 updates per epoch, so the median $\times 512$ canary places 2.8 copies in each update that carries it and a canary in a small bucket up to 151; at 10M the median is 1.15 and at 100M 1.02, where copies of one canary almost never share an update. Between carrying updates the gaps scatter as a memoryless draw, with a coefficient of variation of 0.9–1.0 at every dose in the 10M-M record, rather than at even spacing. Dose is therefore the treated quantity, and the temporal shape of repetition is a recorded property of the plan that differs by tier and is available as a treatment of its own.

Human eligibility review precedes the dose lottery. Targets are selected from the corpus for deliberate repetition and retrieval studies, focusing on natural and human-made subjects. Documented reuse permissions are retained alongside a separate content review that excludes recognizable human faces, children, private documents, license plates, name tags, and prominent personal text. Logo-, watermark-, or URL-dominated images are also excluded. This curation makes copying measurable on selected material suitable for shared experiments; it does not certify the entire training corpus as free of identifiable people. Human reviewers determine eligibility, while the seeded lottery assigns repetition dose. The 170 targets span eight source families, and the dose census is 43/42/43/42 at $16/64/256/512$. These counts imply 35,718 canary slots per authored epoch under the authored instrument. The ledger, rather than this nominal total, determines which occurrences a reported run actually consumed. Original-caption training repeats the image with its original caption. Varied-caption training uses truthful alternatives and retains the base caption; thus it changes caption diversity and original-key frequency together. The ledger audits the policy: of 25 attribute–entity pairs that the original canary captions never state, 24 receive 2–654 presentations through the rewrites.

Original captions are the corpus captions of the canary images; other keys were generated before training. Each canary's dose-1 canary slots receive a seeded mix of image-only descriptions (25%), detail-dropping (25%), shortened (12.5%), image-grounded lengthened (12.5%), and meaning-preserving (25%) rewrites, written by Qwen3-VL-8B-Instruct from the image or Qwen3-8B from the original caption. A vision-language judge other than the generator, shown the image, rejected contradictions; every caption names the entity and differs from the original, and 910 of 35,718 slots (2.5%) repeat a caption. Validation A adds one never-trained image-only description per canary and validation C one never-trained shortened rewrite; five validation C captions equal another canary's trained caption, and those canaries are excluded from the 165-canary key analyses. Entity-only keys read "A photo of a/an {entity}."

Diversity of caption rewrites. We measure whether the rewrite pools provide varied text-conditioning representations using the Vendi score (Friedman & Dieng, 2023). Each canary contributes its original caption and all $f-1$ rewrite occurrences at dose f. We use all 256×1024 penultimate-layer token representations from frozen Qwen3-0.6B-Base (Yang et al., 2025a), retaining padding and tokenwise root-mean-square (RMS) normalization. Caption distances are sliced Wasserstein-2 (SW) distances (Kolouri et al., 2016), approximated with 1,024 shared random unit projections and one caption-only kernel bandwidth.[3] Figure 8 shows mean full-pool scores of 7.95, 16.7, 32.1, and 41.3 for $f = 16, 64, 256, 512$, compared with 1 for identical repeated representations. This supports representation diversity; caption faithfulness is assessed by the image-grounded checks above.

[3]Tokenwise non-affine RMS normalization uses $\epsilon = 10^{-6}$. With projection seed 20260914, we sort the 256 token projections in each direction and average squared differences over tokens and directions to obtain d_{ij}^2. We set $K_{ij} = \exp[-d_{ij}^2/(2\tau^2)]$ and $\mathrm{Vendi} = \exp(-\sum_r \lambda_r \log \lambda_r)$, where λ_r are eigenvalues of K/f. The common bandwidth, $\tau \simeq 0.128$, satisfies $\tau^2 = \mathrm{median}_{d^2>0}(d^2)/2$ over 80,000 caption pairs balanced across doses (seed 20260915), without using memorization outcomes.

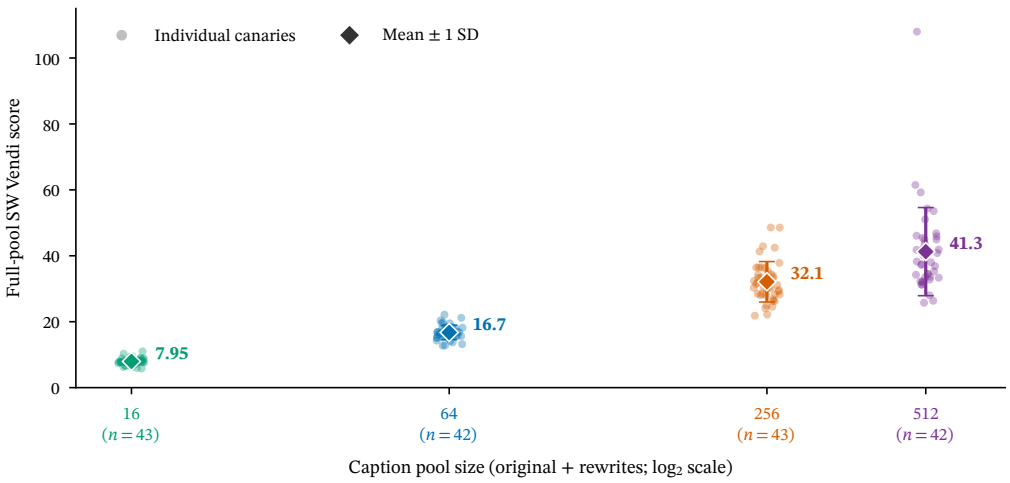


Figure 8: **Larger rewrite pools have higher full-pool representation diversity.** Each faint point is one canary; diamonds and bars show the group mean and ± 1 SD across canaries, not a confidence interval. Dose colors match Figure 7. Groups contain different canaries and have different score ceilings, so the trend is descriptive.

A.5 URL RE-MATERIALIZATION

Before adopting a corpus distributed as image bytes, we attempted to rebuild image–text sets from public URL lists with `img2dataset`, recording per-shard outcome statistics. None of these surfaces contributes training data to the models in this paper; Table 5 reports, as a first-hand example with its access dates, what the lists still returned. Failures include causes on the corpus side (`404`, `403`, expired domains) and some on the side of the crawling environment (resolver and rate-limit behavior), so the download-stage figure bounds link rot from above.

The lower rows of Table 5 list the failure families. These figures agree with published measurements: 78.1% availability across 274M URLs in 24 web-sourced datasets in 2022 (Lakic et al., 2023), about 42% download success when DataComp itself was built (Gadre et al., 2023), and 1.1B of LAION-2B still accessible to a later training effort (Gokaslan et al., 2024). Reachability also understates the problem for a training record: of 3.3M Conceptual Captions images, 2.9M were still hosted in 2023 but only 1.1M were byte-identical to the 2018 download, and expired domains can be bought to replace the images behind surviving URLs (Carlini et al., 2024). A data order over such an index does not determine what a model saw, and the practical response has been to keep training data internal. A versioned registry of distributed bytes binds the images to their exposures.

B TRAINING RECIPE AND REFERENCE METRICS

B.1 ARCHITECTURE

Table 6 specifies the architecture ladder. Each MMDiT block uses SD3-style joint attention with separate modality projections (Esser et al., 2024), GELU/tanh feed-forward networks (Hendrycks & Gimpel, 2016), full per-modality AdaLN-Zero (Peebles & Xie, 2023), RMSNorm on queries and keys (Zhang & Sennrich, 2019; Henry et al., 2020; Dehghani et al., 2023), and image-only 2D RoPE (Su et al., 2024; Heo et al., 2024). The final block reads text keys and values without updating text. The 256^2-pixel area budget admits aspect buckets, so training crops are not all square. These choices define the common recipe; each experiment binds its exact version.

B.2 HYPERPARAMETERS AND TIERS

Table 7 lists the common numerical recipe by corpus tier. With two epochs, varied-caption training presents the original caption twice and each of its 15/63/255/511 rewrites twice, so caption variety

Table 5: **What URL lists returned in 2026.** First-hand re-materialization records, none of which is training data here. DataComp is the Recap-DataComp-1B URL surface (Li et al., 2025; Gadre et al., 2023), crawled in two consecutive parts over disjoint rows (371.8M, then the remaining 457.0M) and pooled by row count; restarts of the second part re-processed 0.06% of the rows. Download survival counts URLs that still returned bytes, usable survival those that also decoded to an image of at least 256 pixels on the short side; failure families are shares of processed rows, grouped. The LAION columns are aggregated from every shard's statistics, the DataComp column from the archived families of its parts; * marks families recorded for the first part only. The size gate is curation, the rest decay or the crawling environment.

	DataComp	LAION-Aesthetics	LAION-POP
Access dates	2026-04-03 to 04-09	2026-04-02	2026-04-02
Input rows (M)	828.8	51.9	0.58
Download survival (%)	76.30	67.08	78.93
Usable survival (%)	**39.52**	45.70	76.67
Failure family, share of processed rows (%)			
Below the 256-pixel size gate	34.91	19.12	0.46
`HTTP 404`	6.36	9.68	7.35
`HTTP 403`	6.21	9.30	4.42
`HTTP 410`	–	0.41	0.28
`HTTP 429` or robots policy	1.03*	1.81	0.40
Other `HTTP 4xx`	–	1.74	0.87
`HTTP 5xx`	–	0.95	1.07
Unresolvable domain	2.92	3.85	2.07
Other DNS failure	–	1.87	1.03
Timeout	2.20	1.92	2.06
Connection refused or closed	–	0.32	0.31
SSL or TLS failure	0.59*	0.81	0.89
Decode error	1.63*	2.27	1.80
Other	–	0.25	0.34

Table 6: Generator sizes. Parameters exclude the frozen text encoder and VAE; all sizes share head dimension 64, feed-forward ratio 4, 16 latent channels, and patch size 2.

	Blue Mage-S	Blue Mage-M	Blue Mage-L
Parameters (B)	0.490	2.017	3.947
Blocks	15	24	30
Hidden width	960	1,536	1,920
Heads	15	24	30

and original-caption frequency change together at fixed image dose; a new intervention can hold either coordinate fixed.

The corpus-pair rows are source-registry counts. Training reads each registry without the captions longer than the 256-token text limit under the Qwen3-0.6B-Base tokenizer (114, 1,005, and 9,752 pairs), so the trained pools hold 997,461, 9,974,386, and 99,744,234 pairs. Updates per epoch are the recorded training-stream lengths for the canary runs; epoch fractions use those bound stream lengths. Each epoch consumes 980,991, 9,946,884, and 99,716,089 of those pairs, the 35,718 canary slots, and 8,315, 9,590, and 10,369 zero-weight pads that round the canary buckets to whole batches (Appendix A.4), which together fill the updates times 1,024 slots; the 16,470, 27,502, and 28,145 pairs left over are the per-bucket tails that single-bucket batches drop. Muon (Jordan et al., 2024) updates hidden two-dimensional matrices with zero weight decay and AdamW (Loshchilov & Hutter, 2019) its auxiliary group with decay 0.01; the learning rate is constant after 50 warmup updates, the first at 2×10^{-6}. Flow time follows the $(0, 1)$ logit-normal law of SD3 (Esser et al., 2024) clamped to $(10^{-5}, 1 - 10^{-5})$, shift 1.0; Gaussian noise and closed-form caption dropout are assigned by the published identities. Mixed precision uses bf16 autocast over fp32 master weights; regional compilation keeps experiment control outside compiled forwards. The layout is eight H200 GPUs at global batch 1,024, with accumulation / per-GPU microbatch $2/64$ for S and $4/32$ for M and L. The

Table 7: Common training settings by corpus tier. Two epochs at global batch 1,024 in every tier; update counts are the recorded canary-stream lengths, and each study records its own saved-state grid (the key study evaluates 40/68/116). GPU-hours are per run on one node of eight H200s.

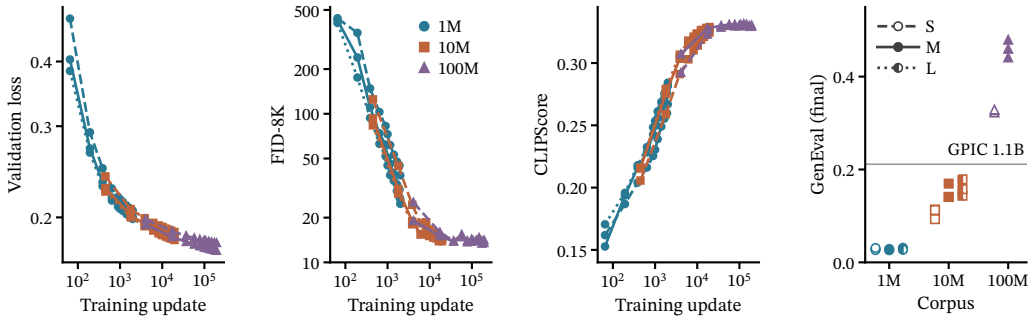
	1M	10M	100M
Corpus pairs	997,575	9,975,391	99,753,986
Updates per epoch	1,001	9,758	97,424
Saved states	40	68	117
Saved-state gap cap	128	1,024	4,096
Ordinary samples per epoch	980,991	9,946,884	99,716,089
Canary slots / pads per epoch	35,718 / 8,315	35,718 / 9,590	35,718 / 10,369
GPU-hours, S / M / L	5 / 11 / 17	57 / 110 / 174	659 / ≈1,100 / –
Peak learning rate		10^{-4}	
Warmup updates		50	
Muon momentum		0.95	
AdamW (β_1, β_2)		$(0.9, 0.95)$	
AdamW ϵ / weight decay		10^{-15} / 0.01	
Global gradient-norm clip		1.0	
Caption dropout		0.1	
Logit-normal (μ, σ)		$(0, 1)$	

recipe uses affine-free LayerNorm with $\epsilon = 10^{-6}$, six AdaLN-Zero modulation parameters per block and modality, no pooled-text AdaLN input, a 2D RoPE grid of at most 128, at most 256 image tokens under the area budget, aspect buckets between $1{:}4$ and $4{:}1$ quantized to 16 pixels without upscaling, masked image padding in up to eight sequence bins of 64 tokens, and Qwen3-0.6B-Base text read at hidden state $[-2]$ with model-side RMSNorm, raw captions, EOS padding, and a 256-token limit over frozen Qwen-Image f8c16 latents normalized by their stored metadata. The EMA decay is $\min(0.9999, (1+s)/(10+s))$ with CPU-sharded storage; the saved-state grid uses one gap of one, two of two, three of four, and so on until the tier's cap, then a constant gap, and the 100M evaluation substitutes update 98,305 for the epoch-boundary state 97,424. Every released state carries the raw weights, the EMA weights, and the optimizer state, and resumes training. The validation probe has 8,320 pairs and its own random plan. The run manifest determines which choices a particular state inherited.

B.3 Reference metrics

The cases concern learning dynamics and make no claim to state-of-the-art image quality: the corpus is restricted to permissively licensed images and carries no aesthetic filter. Reference metrics therefore use references that share this distribution.

Figure 9 follows the eight masked runs through eleven saved states and Table 8 gives the final state of all 24 runs: one image per 8,320 validation captions in its own aspect bucket (Euler sampler, guidance 4.0, 32 steps, one seed, raw weights), scored against the training crops of the same images by FID (Heusel et al., 2017) (clean-fid Inception features (Parmar et al., 2022)), FD-DINOv2 (Stein et al., 2023) (ViT-L/14 (Oquab et al., 2024)), CLIPScore (Hessel et al., 2021) (OpenAI ViT-B/32 cosine (Radford et al., 2021)), validation loss at the same state, and GenEval (Ghosh et al., 2023) (553 prompts, four images each). All four readouts improve through training, and the final states order by corpus and then by size, a ranking that validation loss, FD-DINOv2, and GenEval share (Spearman $\rho = 1.000$ and 0.976) while FID-8K packs the five 10M and 100M states within one point. Within a corpus the M-to-L step is a fraction of the S-to-M step in validation loss and FD-DINOv2. The canary policy leaves the metrics unchanged: the SD across the three runs of a cell is at most 1.2 in FID-8K and 0.02 in GenEval. Against the JiT-T2I baseline released with GPIC (Chandrasegaran et al., 2026) (Table 3), the masked 100M runs trail in FID-8K and CLIPScore and lead in FD-DINOv2 and GenEval, a reference point on the shared corpus rather than a controlled comparison. FID-8K is bound to $N = 8{,}320$ and this reference, a CLIPScore above that of the real images (0.324) measures caption agreement rather than fidelity, each state uses one seed, and GenEval was scored only with a detector build that passes SD v1.5 at 0.43 and blank images at 0.00, since the stock build returns about 0.02 without an error.



(a) Validation loss (b) FID-8K (c) CLIPScore (d) GenEval, final state

Figure 9: Reference metrics through training. (a)–(c) Eight masked runs at eleven saved states, one image per 8,320 held-out captions per state, scored against the training crops of the same images (FID-8K) and their captions (CLIPScore). (d) GenEval at the final state of all 24 runs; the gray line is the GPIC baseline.

Table 8: Final-state reference metrics by corpus and parameter count. Mean over the masked, repeated, and varied runs of a cell at its final state, $\pm$ the SD across the three runs; CLIPScore is in Table 3. Bold marks the best cell.

Corpus	Params	Validation loss ↓	FID-8K ↓	FD-DINOv2 ↓	GenEval ↑
1M	0.5B	0.2062	39.90 ± 1.22	990 ± 19	0.029 ± 0.002
1M	2.0B	0.2006	27.85 ± 1.13	773 ± 14	0.027 ± 0.001
1M	3.9B	0.1988	25.69 ± 0.90	717 ± 17	0.029 ± 0.002
10M	0.5B	0.1874	14.06 ± 0.04	248 ± 2	0.106 ± 0.012
10M	2.0B	0.1827	13.89 ± 0.15	168 ± 1	0.150 ± 0.017
10M	3.9B	0.1809	14.63 ± 0.28	157 ± 4	0.160 ± 0.018
100M	0.5B	0.1785	14.28 ± 0.16	118 ± 1	0.326 ± 0.003
100M	2.0B	**0.1724**	$\mathbf{13.59 \pm 0.08}$	$\mathbf{98 \pm 1}$	$\mathbf{0.460 \pm 0.019}$

A run took from about 5 GPU-hours (1M-S) to about 1,100 (100M-M, projected from its first epoch) on one node of eight H200 GPUs in bf16 (per cell: Table 7); the 24 runs together took about 6,400 GPU-hours and the evaluations of this appendix about 100. A saved state is 3.3× the weights, 6.5, 26.9, and 52.7 GB for S, M, and L, so a run's pool of saved states holds between 0.26 TB (1M-S) and 3.6 TB (10M-L).

C Binding Dynamics and Exposure

C.1 Panel, Generation, and Judging

Panel. Attributes are curated from the caption parser's frequency table: a prompt surface form plus an explicit alias set over which counts are summed, admitted when the summed binding-edge count exceeds 10^4, the property is visually decidable on a single object at the training resolution, and the fixed prompt template reads naturally; the six categories each span at least two decades of corpus frequency (79 attributes). Entities are canonical lemmas with at least 10^3 captions that pass a WordNet (Miller, 1995) physical-object filter, a scene-word stoplist, and a 105-word generic-noun exclusion; up to sixty are drawn per log-frequency decade with a fixed seed (240 entities). A pair's support is the number of ten-million-corpus captions whose audited annotation binds that attribute to that entity, in five strata (0, 1–9, 10–99, 100–999, $\geq$ 1,000); the same cells are then followed on both tiers. From the grid, 2,150 pairs are sampled by balanced round-robin over (attribute frequency decade $\times$ support stratum $\times$ entity-frequency stratum) with caps of nine pairs per entity and thirty-five per attribute and a coverage-seeking tie-break, and accepted only if all pairwise Spearman correlations among log entity frequency, log attribute frequency, and $\log(1+\text{support})$ fall below 0.5 (realized 0.095, 0.277, 0.284); the sampled support strata hold 474/488/497/424/267 cells. Fifty-five fixed

color cells (seven colors on eight common entities) were generated alongside but sit outside the sampling rule and are excluded from every analysis reported here; each entity adds one attribute-free prompt. Every pair is rendered from the single template "a photo of {a} {attribute} {entity}" and every entity-only prompt from "a photo of {a} {entity}", 2,451 prompts in all. Selection uses corpus statistics only, is frozen before generation, and is fingerprinted.

Generation and judging. Every saved checkpoint of every run renders every prompt from ten fixed latent noises shared across prompts, checkpoints, and runs; four read checkpoints (epochs 0.7, 1, 1.5, 2) add ninety noises per pair prompt, so the terminal per-cell regressions of Section C.5 rest on one hundred paired draws per cell, while onset, crossing, and paired-contrast estimates read the ten main-grid draws at every saved checkpoint alike. The grids hold 40 saved checkpoints at 1M and 177 at 10M, the latter saved more densely than the 68 states of the repetition runs to follow onsets closely and reproducible from the plan (Section 4.4), for $980{,}400$ and $4{,}338{,}270$ main-grid images per run and $793{,}800$ supplementary images. Sampling is 32 Euler steps with classifier-free guidance 4.0 at 256×256, the timestep schedule of each checkpoint's training objective, raw (non-EMA) weights, bf16 autocast, batches of 128; the noises are positions 0–9 and 10–99 of one fixed seed, identical for every prompt, checkpoint, and run, and the text encoder and decoder are those of the training runs (Section B). The judge answers the two-field question of the judge protocol below under greedy decoding, and an image passes when both fields are true. Of the $10{,}637{,}340$ main-grid judgments, 83 failed to parse, all on two prompts of one attribute ("pointed bed" and "pointed stocking" in the two 10M runs), whose name the model substituted for the second field's key; a failed judgment is dropped from its run–checkpoint–prompt count, never imputed, leaving at least seven of ten in each of the 70 affected counts. Every judged image carries the SHA-256 of the judged bytes, checked against the generation receipt.

Judge protocol. The judge is the publicly released Qwen3-VL-8B-Instruct vision–language model with its standard processor. Each judgment is one user turn containing the image and one question, rendered through the chat template, batched at 128 images, and decoded greedily with at most 48 new tokens; the judge never sees the generation prompt, any caption, or any other image. With `{noun}` the entity's noun phrase from the generation prompt, article included ("an alleyway", "a mailbox"), and the prompted attribute substituted, the pair question is, verbatim:

```
Inspect this generated image without using any text prompt.
(1) Is {noun} clearly visible?  (2) If it is visible, is
that {noun} itself {attribute}?  Judge only the object,
not the background, lighting, text or other objects; if
several such objects appear, judge the largest or most
prominent one.  Return exactly one JSON object and no prose:
{"entity_present":  <true or false>, "attribute_present":
<true or false>}.  If the object is not visible, answer
attribute_present false.
```

Attribute-free images receive the single-field form:

```
Inspect this generated image without using any text prompt.
Is {noun} clearly visible?  Return exactly one JSON object
and no prose:  {"entity_present":  <true or false>}.
```

The first field is read over all judged images and the second over the images whose first field is true (Section 5.3). The whole response, or a single fenced block that is the whole response, must parse as one JSON object with boolean fields (`true`/`false`/`yes`/`no` strings coerced, duplicate keys rejected); the pair question requires both fields and the entity-only question the first, and any other response is a parse failure whose fields are not counted. The attribute is asked conditionally on the entity, so $P(\text{attribute})$ and $P(\text{entity} \mid \text{attribute})$ are not available: three unconditional rewordings tried on a common set of 512 terminal images changed the entity answer itself (332 yes under the reported wording against 168, 172, and 179 under the variants, 68–70% agreement), so a symmetric readout needs a calibrated evaluator rather than a reworded prompt.

C.2 Exposure accounting and estimators

Exposure. The exposure counts are $K_y(s)$ of Equation (4), recounted from the ledger at every saved checkpoint of every run for each pair, entity, and attribute: every loss-bearing presentation of a sample whose caption annotation carries the concept, with zero-weight slots excluded and caption-dropped slots included. Each tier has its own plan and ledger, so the two runs on a tier are compared at identical exposure, while across tiers the counts differ by construction and are the quantity contrasted; they come from the pinned engine's ledger query over digest-verified artifacts. The panel's support axis is a caption-audit statistic of the 10M corpus, whereas the reported exposures are realized presentations: at 10M the two differ by the run's presentation multiplicity (two epochs; median ratio 2.0), and the 1M exposures are counted on the smaller corpus and bear no fixed ratio to the support axis. The text names which one a number is.

Spreads. Terminal stratum rates carry a percentile cluster bootstrap in two schemes: entity labels, and separately attribute labels, are resampled with replacement and the statistic is recomputed per resample ($B = 300$ each); the tables print the SD of the resampled statistic, the larger of the two schemes. Paired contrasts and the onset elasticity resample entities only, and the fixed-checkpoint OLS slopes carry their OLS standard errors. The all-cells pooled coefficients of Table 10 and the two-stage fits of Table 11 carry the two-scheme bootstrap, with each drawn copy of a cluster relabeled as its own random-effect level and the model refitted per resample; the zero-pair block and Table 11 keep model-based standard errors, which are conditional on the fitted variance components.

Onsets. A cell's onset at $\tau = 0.5$ is the first saved checkpoint whose raw pass rate reaches τ there and at the next saved checkpoint, read from the ten main-grid draws. The onset is recorded at that checkpoint, so it is resolved only to the checkpoint spacing; a cell with no such pair of checkpoints, including one that first reaches τ at the final checkpoint, is right-censored. The elasticity is the OLS slope of $\log(\text{onset update})$ on $\log(1+\text{terminal pair presentations})$ over non-censored cells with an entity-cluster bootstrap spread. Censoring is heavier at low support and at one million, where only 105 and 187 of 2,150 cells cross, so those slopes describe crossing cells. The clock dispersion of Table 9 is the median absolute deviation of log onset read in four coordinates (update; cumulative attribute, entity, and pair presentations at onset), with exposures entering as $\log(1+\cdot)$ and the update as $\log$.

Regressions. Two fitters are used and never mixed in one number. The pooled models are binomial generalized additive mixed models (`mgcv::bam`, fast REML, discretized covariates) over all saved checkpoints of the pair cells, with $\log(1+\cdot)$ exposures, $\log(1+U)$, category fixed effects, and entity, attribute, and prompt random intercepts; the pair-zero variant drops the pair term and keeps only the observations (cell, checkpoint) at which the cell's cumulative pair presentations are still zero, and the two-stage variant refits the same model on each judge field. No single joint model over the four runs is reported because the four-run interaction model aborted (segmentation fault) in the discretized code path; capacity and corpus contrasts are therefore paired differences of rates. The fixed-checkpoint instrument is ordinary least squares on the terminal per-cell rates (one hundred draws) against $\log_{10}(1+\text{presentations})$, in percentage points per decade; it has no random intercepts, so its attribute-axis slopes compare attributes rather than doses of one attribute.

Paired contrasts. For runs A and B the paired difference is $\Delta_{AB} = |C|^{-1} \sum_{c \in C} [\hat{p}_c^B - \hat{p}_c^A]$ over the pair cells C (all 2,150, or one support stratum), where $\hat{p}_c$ is the terminal pass rate over the ten main-grid draws; every entity-cluster resample is applied to both runs, so the spread is that of the difference. Capacity contrasts pair the two runs on a tier; corpus contrasts pair the two runs at a capacity and are stratified by the panel's caption-audit support, the only support axis the tiers share. The contrasts quantify evaluation-cell variability around one trained run per cell and one judge, not training-seed or judge variability.

C.3 Level, timing, and pooled terms

Level against timing. Table 9 compares the same cells in several training coordinates. Binding support is strongly associated with terminal performance but weakly with onset among crossing cells (10M elasticities -0.073 and -0.087), and update time is less dispersed than cumulative pair

Table 9: **Support predicts level more than timing.** Crossing cells out of 2,150; all subsequent columns summarize those crossing cells only. Elasticity is ± entity-cluster bootstrap SD ($B = 300$); italics mark a 95% bootstrap interval that covers zero. MAD columns give dispersion of log onset in update, attribute, entity, and pair-exposure coordinates.

Run	Crossing	Median update	Elasticity ± SD	MAD s	A	E	EA
1M-S	105	1,153	$+0.012 \pm 0.016$	0.29	0.92	1.02	1.87
1M-M	187	1,281	-0.022 ± 0.014	0.34	0.95	1.06	1.88
10M-S	1,334	5,505	-0.073 ± 0.009	0.57	1.07	1.21	2.08
10M-M	1,529	4,865	-0.087 ± 0.008	0.58	1.04	1.25	1.99

Table 10: **Pair support remains associated after adjustment.** Pooled binomial mixed-model change in log-odds per tenfold increase in presentations. Upper block: estimate ± bootstrap SD, the larger of the entity- and attribute-cluster replicate SDs ($B = 300$ each, duplicated clusters relabeled as distinct random-effect levels); italics mark terms whose 95% bootstrap interval covers zero. Lower block: estimate ± model-based SE. E, A, and EA denote entity, attribute, and pair presentations; s is update. The zero-pair fit omits EA. Bold marks the pair term; these are conditional associations.

Run	E	A	EA	s
All pair cells (cluster-bootstrap spread)				
1M-S	$+0.299 \pm 0.108$	$+0.092 \pm 0.500$	$\mathbf{+0.613 \pm 0.058}$	$+3.428 \pm 0.524$
1M-M	$+0.212 \pm 0.109$	$+0.230 \pm 0.377$	$\mathbf{+0.567 \pm 0.058}$	$+3.203 \pm 0.404$
10M-S	$+0.362 \pm 0.298$	-0.467 ± 0.574	$\mathbf{+0.281 \pm 0.088}$	$+2.356 \pm 0.628$
10M-M	$+0.462 \pm 0.402$	-0.419 ± 0.650	$\mathbf{+0.328 \pm 0.088}$	$+2.168 \pm 0.709$
Zero pair presentations (model-based SE)				
1M-S	$+0.355 \pm 0.124$	-0.049 ± 0.296	–	$+4.073 \pm 0.335$
1M-M	$+0.263 \pm 0.121$	$+0.142 \pm 0.273$	–	$+3.734 \pm 0.308$
10M-S	$+0.082 \pm 0.105$	$+0.027 \pm 0.238$	–	$+2.202 \pm 0.263$
10M-M	$+0.142 \pm 0.095$	$+0.093 \pm 0.222$	–	$+2.023 \pm 0.244$

presentations at onset. Because presentations combine support with elapsed training, the comparison describes the curves rather than identifying a causal clock, and severe censoring at 1M makes its crossing-cell slopes selective; a resource for dynamics keeps the continuous curves and the censoring state beside any onset summary.

Pooled exposure terms. Table 10 reports every exposure term of the pooled full-panel and zero-pair specifications. Pair exposure stays positive after category and entity, attribute, and prompt terms are included, and the zero-pair fit measures the remaining associations without a pair coefficient. Under the cluster bootstrap the pair and update terms exclude zero in every run, the attribute term covers zero in every run, and the entity term excludes zero at 1M-S and 10M-M and narrowly covers it at 1M-M and 10M-S; the bootstrap SDs are 1.0–2.0× the model-based SE at 1M and 3.1–8.3× at 10M. This is a model of observed associations, not a randomized exposure edit.

The two judge fields. Table 11 gives the pooled decomposition behind Figure 3(c) with both spreads. Entity exposure changes sign across fields while pair exposure is positive in both; the conditional field is evaluated on entity-visible images, so its changing cohort is part of its definition. Under the entity- and attribute-cluster bootstrap the entity term keeps its sign in both fields and all four runs, positive for visibility and negative for control, where its bootstrap SD is only 1.1–1.6× the model-based SE; the pair term stays positive for control in every run but covers zero for visibility at 10M, and the attribute term covers zero in both fields. At 10M some attribute-cluster replicates move the attribute and update terms in opposite directions, so their control spreads are wide.

C.4 Support ladder and crossing

Table 12 records every stratum of realized pair presentations in the four-run binding panel, including cells that never cross. Terminal success rises from 0.06–0.24 at zero indexed presentations to 0.31–0.74 at a thousand or more (0.24 to 0.74 at 10M-M) while the pooled pair coefficient stays positive in

Table 11: **The entity association changes sign.** The pooled specification applied to each judge field; the conditional stage is fit on entity-visible images only. Entries are changes in log-odds per tenfold increase in presentations ± model-based SE, followed in parentheses by the cluster-bootstrap SD, the larger of the entity- and attribute-cluster replicate SDs ($B = 300$ each, duplicated clusters relabeled as distinct random-effect levels). E, A, and EA are entity, attribute, and pair presentations and s the update, as in Table 10. Bold marks the entity term; italics mark terms whose 95% bootstrap interval covers zero.

Run	E	A	EA	s
Entity visible				
1M-S	$\mathbf{+0.690 \pm 0.118}$ (0.134)	$\mathit{-0.038 \pm 0.073}$ (0.062)	$+0.257 \pm 0.020$ (0.033)	$+3.301 \pm 0.141$ (0.234)
1M-M	$\mathbf{+0.590 \pm 0.109}$ (0.150)	$\mathit{-0.020 \pm 0.076}$ (0.073)	$+0.270 \pm 0.018$ (0.031)	$+3.109 \pm 0.135$ (0.219)
10M-S	$\mathbf{+0.779 \pm 0.095}$ (0.564)	$\mathit{-0.110 \pm 0.067}$ (0.569)	$\mathit{+0.091 \pm 0.008}$ (0.059)	$+1.853 \pm 0.116$ (0.610)
10M-M	$\mathbf{+0.901 \pm 0.087}$ (0.594)	$\mathit{-0.162 \pm 0.069}$ (0.507)	$\mathit{+0.106 \pm 0.008}$ (0.053)	$\mathit{+1.749 \pm 0.111}$ (0.645)
Attribute given entity				
1M-S	$\mathbf{-0.721 \pm 0.086}$ (0.095)	$\mathit{-0.131 \pm 0.400}$ (0.590)	$+0.873 \pm 0.055$ (0.082)	$+2.087 \pm 0.419$ (0.605)
1M-M	$\mathbf{-0.672 \pm 0.077}$ (0.085)	$\mathit{-0.187 \pm 0.411}$ (0.673)	$+0.849 \pm 0.050$ (0.088)	$+2.289 \pm 0.424$ (0.653)
10M-S	$\mathbf{-0.658 \pm 0.073}$ (0.119)	$\mathit{+0.040 \pm 0.294}$ (2.208)	$+0.623 \pm 0.020$ (0.152)	$+1.402 \pm 0.303$ (2.212)
10M-M	$\mathbf{-0.709 \pm 0.075}$ (0.115)	$\mathit{-0.049 \pm 0.295}$ (3.001)	$+0.745 \pm 0.018$ (0.144)	$+1.651 \pm 0.305$ (2.997)

Table 12: **Support separates success and crossing probability.** Realized terminal pair presentations define the strata. Success is the pooled terminal pass rate over the ten main-grid draws ± bootstrap SD, the larger of the entity- and attribute-cluster SDs ($B = 300$ each). Onset is pass ≥ 0.5 at two consecutive states; non-crossers stay in the denominator. Bold marks the highest stratum.

Run / stratum	0	1–9	10–99	100–999	≥ 1000
Terminal success, mean ± SD					
1M-S	0.06 ± 0.01	0.10 ± 0.01	0.13 ± 0.01	0.21 ± 0.02	$\mathbf{0.31 \pm 0.04}$
1M-M	0.09 ± 0.01	0.13 ± 0.01	0.17 ± 0.01	0.26 ± 0.02	$\mathbf{0.37 \pm 0.05}$
10M-S	0.21 ± 0.02	0.24 ± 0.02	0.38 ± 0.02	0.50 ± 0.02	$\mathbf{0.66 \pm 0.02}$
10M-M	0.24 ± 0.02	0.30 ± 0.02	0.45 ± 0.02	0.58 ± 0.02	$\mathbf{0.74 \pm 0.02}$
Cells reaching onset (%)					
1M-S	1	3	5	14	26
1M-M	3	8	10	21	39
10M-S	36	43	63	77	94
10M-M	45	54	75	86	98
1M: cells	856	469	483	296	46
10M: cells	474	341	521	478	336

all four runs; the strata separate during training and stay separated at the endpoint. The zero stratum has no recorded caption binding under this annotation, so the panel measures generalization beyond indexed bindings and visual support may remain; new visual annotations can test how much of that success the caption index misses. The table's denominator is the frozen pair panel, with support recomputed from realized loss-bearing presentations (Section C.2), so a positive zero-stratum rate and an uncrossed cell remain distinct observations.

C.5 Judge fields, attribute axis, and capacity

Judge fields. Figure 3(c) uses the pooled binomial model on each judge field separately; Table 13 is the terminal-rate cross-check in percentage points per decade, on the entity axis and on the attribute axis. A joint rate multiplies the two stages and leaves a smaller net slope while its factors move in opposite directions. Attribute-only exposure is nearly collinear with attribute identity, so its slope compares attributes rather than the effect of presenting one attribute more often, and the conditional readout changes its image cohort as visibility changes. The 240 entity-only prompts give a separate visibility readout, terminal means 0.05/0.08 at 1M and 0.38/0.42 at 10M for S/M, and visibility rises with entity exposure under both prompt forms. Within realized support strata the conditional attribute slope on entity exposure stays negative, while the visibility slope turns slightly negative among the 46 most-presented pairs at 1M; at 10M raw attribute-exposure slopes are largest at low support. A within-attribute first-presentation contrast on the same cells (465 near-zero cells read four saved

Table 13: **The product conceals opposite signs.** Terminal OLS slopes in percentage points per decade of entity presentations (upper block) and of attribute presentations (lower block), estimate ± standard error. Entity and joint fits use 2,150 cells; the conditional fit uses the 2,033/2,065/2,144/2,146 cells with entity-visible draws at 1M-S/1M-M/10M-S/10M-M. Bold marks the component readouts. Attribute slopes compare attributes, whose identity and frequency remain entangled. Uncertainty conditions on one trained run per row and one judge.

Run	Entity visible	Attribute | entity	Joint pass
Per decade of entity presentations			
1M-S	$\mathbf{+8.5 \pm 0.6}$	$\mathbf{-5.6 \pm 1.2}$	$+2.9 \pm 0.5$
1M-M	$\mathbf{+9.1 \pm 0.7}$	$\mathbf{-4.7 \pm 1.1}$	$+3.3 \pm 0.5$
10M-S	$\mathbf{+10.5 \pm 0.8}$	$\mathbf{-5.2 \pm 1.0}$	$+2.9 \pm 0.9$
10M-M	$\mathbf{+11.0 \pm 0.8}$	$\mathbf{-5.4 \pm 1.0}$	$+3.4 \pm 0.9$
Per decade of attribute presentations			
1M-S	-0.1 ± 0.7	$\mathbf{+17.4 \pm 1.3}$	$+3.2 \pm 0.5$
1M-M	-1.3 ± 0.8	$\mathbf{+15.0 \pm 1.3}$	$+3.4 \pm 0.6$
10M-S	$+1.5 \pm 1.0$	$\mathbf{+16.1 \pm 1.2}$	$+10.2 \pm 1.0$
10M-M	-0.4 ± 0.9	$\mathbf{+14.4 \pm 1.1}$	$+9.2 \pm 1.0$

Table 14: **Gains are larger at supported bindings.** Paired terminal success differences. Values are mean ± entity-cluster bootstrap SD ($B = 300$) at zero support, in the highest support stratum, and over all cells (bold). Spreads condition on one trained run per cell; support strata are fixed corpus annotations for this comparison.

Support	1M: M−S	10M: M−S	S: 10M−1M	M: 10M−1M
Zero	0.022 ± 0.006	0.028 ± 0.008	0.153 ± 0.012	0.159 ± 0.012
High	0.054 ± 0.015	0.077 ± 0.013	0.445 ± 0.021	0.469 ± 0.021
All	$\mathbf{0.033 \pm 0.005}$	$\mathbf{0.063 \pm 0.006}$	$\mathbf{0.280 \pm 0.010}$	$\mathbf{0.310 \pm 0.010}$

checkpoints before and after their first presentation against never-presented cells of the same attribute, attribute-cluster bootstrap) is inconclusive: $+0.3$ and $+0.0$ points at 10M against a model-implied $+0.5$ and $+0.8$, with a placebo shift of the same size inside the pre-window.

Capacity and corpus. Table 14 contrasts terminal outcomes within the same pair cells: S to M raises success by 0.033 at 1M and 0.063 at 10M, and 1M to 10M at fixed size by 0.280 and 0.310, both larger in the highest support stratum than at zero support. Within-tier comparisons share a training plan; cross-tier comparisons change the consumed corpus and exposure together, so a new analysis can choose the comparison that matches its question rather than read every difference as a scale effect.

C.6 Two-object protocol

Panel and generation. Figure 4 and Figure 10 use a separate two-object study, distinct from the single-pair panel of Appendix C. The evaluated model is 100M-S, with gradients from added repetitions masked: raw weights at 128 saved states spanning updates 0–194,848, a grid saved more densely than the 117 states of Section 4.4 to follow the early phase, with one training seed. The panel contains 24 target pairs, two companion entities per target, three competitor-color levels, and both target positions: $24 \times 2 \times 3 \times 2 = 288$ two-object prompts, plus 24 single-object prompts. Two-object prompts use the template "{color$_1$} {entity$_1$} and {color$_2$} {entity$_2$}", without a photographic prefix. Targets were selected before generation from ten colors and manufactured-object candidates; all four requested/crossed pairs required at least 3,000 graph documents. Within each matched context, the target pair and companion entity were fixed, while the companion color varied to give low, middle, and high graph-document frequencies for the target's crossed pair. Each configuration was evaluated in both object orders. The 24 targets span twelve pairs with 3,000–29,999 graph documents and twelve with at least 30,000. At each saved state, eight different generation seeds (8–15) were paired across prompts, order, and saved states, using classifier-free guided Euler sampling with 32 denoising steps, guidance scale 4, and 256×256 output: 319,488 images including single-object generations.

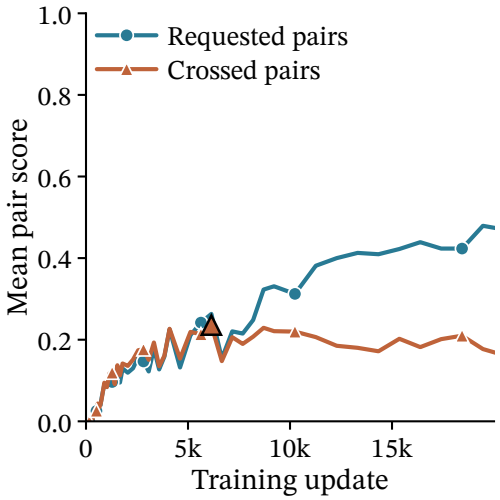


Figure 10: **Early training of the two-object panel.** Requested and crossed scores over the first updates of 100M-S; each curve averages both entities and all 288 prompts, the time axis is linear, and the black-edged triangle marks the crossed peak. The full run and the exposure panel are Figure 4(a, c). Curves use raw eight-seed scores.

The resolution-dependent timestep-shift map is $(32, 1)$, $(256, 1)$, $(512, 1.73)$, and $(1024, 3)$; weights are raw rather than exponential-moving averages. Source annotations are graph-derived caption pairs, not independently verified visual labels.

Four scores and their roles. Figure 4(a) and Figure 10 measure requested and crossed-pair expression separately. Write the target pair as (a_t, e_t) and the companion pair as (a_c, e_c); target and companion denote panel roles, and both prompt orders are evaluated. The four questions concern requested pairs $T = (a_t, e_t)$ and $A = (a_c, e_c)$ and crossed pairs $X = (a_c, e_t)$ and $O = (a_t, e_c)$; below, the same symbols also denote their scores on the same generated image. The fixed Qwen3-VL-8B-Instruct judge receives the image and one pair question, without the generation prompt; greedy decoding produces two Boolean fields, `entity_present` and `attribute_present`. Their conjunction is the image-level exact-binary score, and C_{abs} is its mean over eight seeds. Invalid parses are missing, not zero; the plotted two-object prompt–pair combinations have eight valid scores per pair. For a white car, the question is:

> Inspect this generated image without using any text prompt. (1) Is car clearly visible? (2) If it is visible, is that car itself white? Judge only the object, not the background, lighting, text or other objects; if several such objects appear, judge the largest or most prominent one. Return exactly one JSON object and no prose: `{"entity_present": <true or false>, "attribute_present": <true or false>}`. If the object is not visible, answer attribute_present false.

Each attribute is judged independently: requested and crossed scores need not sum to one and may disagree on entity presence. For the role curves, each prompt contributes $(T + A)/2$ to the requested score and $(X + O)/2$ to the crossed score before equal-weight prompt averaging. Their gap averages two entity-specific contrasts and is distinct from target-entity $C_{\text{contrast}} = T - X$. Single-object prompts are evaluated separately and do not enter the two-role mean.

Exposure and peak estimators. Figure 4(c) associates exposure with the height, not the timing, of crossed-pair peaks. Conditional exposure follows Equation (4): before saved state s, it counts presentations of the annotated pair with loss mask $m = 1$ and caption $c \neq \varnothing$. Repeated presentations count repeatedly; selected-sample inverse queries were checked against forward replay. For each pair in each prompt, take the maximum raw eight-seed score over the observed grid and average these maxima across that pair's contexts. This differs from the maximum of a context-averaged curve or a quartile-averaged curve. Pooling crossed roles X and O gives 73 unique crossed pairs from 21

Table 15: **Exposure accompanies higher crossed peaks across six cohorts.** Pearson correlations use log conditional exposure; p values are from entity-cluster robust slope tests, Holm-adjusted across the six cohorts. The 100M-S row uses 128 saved states.

Cohort	Pearson r	Holm-adjusted p
100M-S	0.611	8.39×10^{-6}
100M-M	0.603	2.98×10^{-5}
10M-S	0.635	3.83×10^{-6}
10M-M	0.605	8.83×10^{-6}
1M-S	0.554	1.19×10^{-4}
1M-M	0.606	2.98×10^{-5}

entities, equally weighted in the correlation. Panel (c) uses $\log(1 + K_{\text{final}})$, where K is conditional exposure; quartile membership follows the saved final-conditional-exposure partition and error bars show across-pair SD of the plotted peak statistic. The partition uses equal-count quartiles of ranked exposure, breaking ties by the saved alphabetical pair order. The quartile points are not maxima of quartile-mean trajectories. Final conditional exposure is a retrospective descriptor, not a prospective predictor.

Uncertainty and limits. Figure 4(c) gives Pearson $r = 0.611$ between peak crossed score and log-transformed final conditional exposure; the positive association persists when exposure is measured at each context's peak. Its two-sided slope test uses entity-cluster robust uncertainty and Holm correction over six tests: two roles crossed with absolute log exposure, color-normalized log exposure, and within-entity relative exposure; the adjusted p is 1.01×10^{-5}. Using actual exposure at each context's peak yields $r = 0.518$; peak ties use the earliest saved state, and log exposure is averaged within pair before correlation. Its 95% entity-cluster percentile-bootstrap interval is $[0.346, 0.674]$ (2,000 resamples). This interval conditions on one fixed trained model and does not quantify training-run variability. Shared colors and contexts can induce dependence beyond entity clustering. The difference between the correlations using final and peak-time exposure has not been tested. For the color-normalized predictor $\log(1 + K_{\text{pair}}) - \log(1 + K_{\text{color}})$, the corresponding correlations are 0.661 and 0.657.

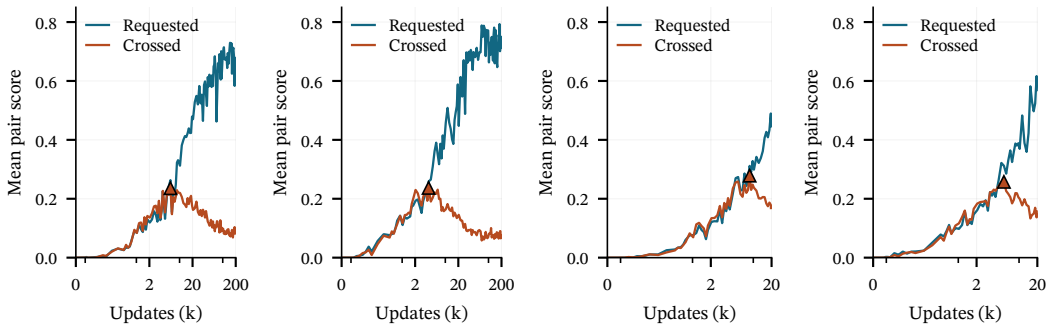
The panel supports an exposure–peak association where crossed-pair expression rises early and later separates from requested-pair expression; it does not establish a universal peak time, faster suppression at higher exposure, or exposure as the cause of crossed-pair expression.

C.7 Recurrence across cohorts and models

Cohorts. With the panel, scores, and peak estimator defined in Appendix C.6, the crossed-pair peak–log-conditional-exposure association is positive in all six cohorts (Table 15). Each cohort contributes one entity-cluster robust slope test to a common six-test Holm family; all adjusted $p < 0.001$. This family spans cohorts, distinct from the within-cohort role–predictor family above.

For 100M-S on its 128-state grid, Pearson $r = 0.611$ has a 95% entity-cluster percentile-bootstrap interval of $[0.485, 0.693]$ (73 pairs, 21 entity clusters, 2,500 resamples). The interval estimates uncertainty in Pearson r, whereas the p values test regression slopes; neither measures variation across independent training runs.

Model settings. Figure 11 shows early growth in crossed-pair expression followed by decline in all four evaluated runs, using the same 288 two-object prompts and eight paired generation seeds. The raw mean requested-minus-crossed gap becomes negative in every run (minima from -0.036 to -0.027), then reaches $+0.278$ to $+0.638$ at the respective endpoints. Crossed-pair score peaks occur at different updates (4,097–8,705), rather than a common transition time. The 10M runs have shorter horizons. The same qualitative pattern appears across all four evaluated model settings.



(a) 100M-S (b) 100M-M (c) 10M-S (d) 10M-M

Figure 11: **Crossed-pair expression rises before role separation.** Raw requested-pair (blue) and crossed-pair (orange) scores average both entities and all 288 prompts. Black-edged triangles mark the maximum of each mean crossed-score curve. Panels use 128, 116, 68, and 68 saved states, respectively; vertical scales match and horizontal axes are logarithmic above update 512. The repeated rise–decline pattern extends beyond the 100M-S model, while timing and endpoint separation differ.

C.8 Decomposition and exposure cohorts

Decomposition. Table 16 reads the judge's two answers separately for the 100M-S panel of Figure 4: whether the entity is visible and, when it is, whether it wears the requested or the crossed color; the two color questions are asked independently, so both can hold. For each pair question, the joint score factors into visibility times color frequency conditional on visibility. Target crossed-color frequency reaches 0.65 at update 1,537, when target visibility is about one fifth, and subsequently trends downward while visibility generally increases. At update 2,817, the requested-pair questions classify 51% of images as showing neither object and 9% as showing both. Restricting further to positive entity-visibility judgments on all four pair questions leaves 167 images; among these, 45% show only one prompt color on both objects, 1.2% a clean swap, and 1.8% both colors in place without either crossed pair; these exclusive categories leave 52% with both or neither prompt color detected on at least one object.

Word order is associated with the direction of color spread: the first-mentioned object has a higher crossed score for the second-mentioned color than the reverse in 176 prompts, versus 106 in the opposite direction (six ties; scores averaged through update 5,121). Averaging the target crossed-minus-requested score within each target and color level, then equally across the 24 targets, gives -0.03, $+0.08$, and $+0.09$ at update 2,817 for the low, middle, and high graph-document-frequency levels used to select the competitor colors (high minus low $+0.12$, with high $>$ low in 15 of 24 targets; at update 6,145, the difference is $+0.18$, with high $>$ low in 19 of 24 targets).

Single-object prompts reach half of their own late requested-score level (the mean over the last 16 saved states) at update 13,313, compared with 11,265 for two-object prompts of the same targets, but a visible entity wears the requested color more often in them (0.68 against 0.46 over updates $\leq 5{,}121$; seven of eight targets with at least 50 (13.0%) of the 384 single-object generations per target judged entity-visible).

Separating the crossed roles in Figure 4(c), the peak–conditional-exposure correlation is 0.69 for 45 pairs appearing as X and 0.52 for 36 appearing as O; eight pairs occur in both sets. These are pair-level correlations across the panel, rather than within-target comparisons of the three color levels. The descriptive analyses use one training run and one automatic judge over 24 design targets.

Exposure quartiles. Figure 12 follows the 73 crossed pairs in quartiles of absolute final conditional exposure, with 19, 18, 18, and 18 pairs from Q1 to Q4, respectively. The raw quartile-mean trajectory reaches a maximum of 0.142 in Q1 and 0.397 in Q4; the corresponding updates are 9,217 and 5,121. These are maxima of mean curves, not the mean prompt-level maxima in Figure 4(c). Membership is fixed across training using the saved ranked-exposure partition; ties follow alphabetical pair order.

Table 16: **Visibility rises while the crossed color recedes.** Rates over the 288 two-object prompts of the 100M-S run at five updates, mean ± SD across the eight generation seeds; color rows condition on visibility in their own pair question, which can differ between requested and crossed questions. Bold marks the crossed color's fall. Single-object rows use the 24 single-object prompts of the same targets.

Update	2,817	6,145	20,481	98,305	194,848
Target visible (T question)	0.31 ± 0.05	0.47 ± 0.04	0.63 ± 0.04	0.77 ± 0.05	0.76 ± 0.04
crossed color	$\mathbf{0.61 \pm 0.08}$	$\mathbf{0.49 \pm 0.05}$	$\mathbf{0.25 \pm 0.03}$	$\mathbf{0.13 \pm 0.02}$	$\mathbf{0.11 \pm 0.02}$
requested color	0.46 ± 0.07	0.54 ± 0.07	0.74 ± 0.03	0.89 ± 0.02	0.91 ± 0.03
Companion visible (A question)	0.28 ± 0.05	0.47 ± 0.06	0.59 ± 0.03	0.75 ± 0.04	0.71 ± 0.03
crossed color	0.60 ± 0.06	0.51 ± 0.06	0.32 ± 0.06	0.14 ± 0.03	0.12 ± 0.02
requested color	0.56 ± 0.07	0.57 ± 0.04	0.80 ± 0.04	0.87 ± 0.02	0.92 ± 0.02
Single-object target visible	0.29 ± 0.10	0.26 ± 0.09	0.50 ± 0.05	0.69 ± 0.10	0.64 ± 0.05
requested color	0.78 ± 0.13	0.75 ± 0.16	0.89 ± 0.04	0.98 ± 0.03	0.94 ± 0.12

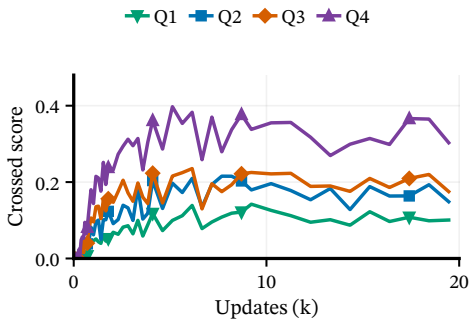


Figure 12: **High exposure accompanies stronger crossed-pair expression early.** Raw crossed-score means over the first 20,000 updates by quartile of final conditional exposure (Q1–Q4, fixed across saved states); curves average eight-seed scores within contexts, then weight pairs equally. The full run is Figure 4(b). Different peak locations do not support an invariant transition time.

High-exposure cohort. Target-entity contrast, $T - X$, over all 288 two-object prompts and over the 76 prompts whose target crossed pair has high final conditional exposure (the 75th percentile, 128,008 conditional presentations, all eight cutoff ties included) reaches raw minima of -0.048 and -0.166, at updates 2,817 and 3,073 (tied) for the full panel and 3,585 for the high-X subset. Selection uses exposure alone, without judge scores; the subset overlaps the full panel and the minima fall at different updates, so the deeper trough is descriptive.

D Controlled Retrieval and Continuation

D.1 Readout, onset rule, and calibration

Readout. Every instance readout generates from raw weights with classifier-free guided Euler sampling (`cfg-euler.v1`, 32 steps, guidance scale 4, bf16 autocast; the timestep-shift map of Appendix C.6). The canary panel renders each target in the aspect bucket of its training latent with four paired noises (seed 42), shared by every key, arm, and saved state; the continuation bank draws eight fresh noises per source and key, and the neutral-prompt panel 64 rollouts per prompt at 256×256 on its state grid, with 256 at the end of epoch one, at 1.5 epochs, and at the terminal state. SSCD compares each generated image to the target's pre-VAE training crop. Paired differences keep target, caption, generation noise, and saved state fixed. Instance-study means first average noises within a target and then weight the eight source families equally, with 2,000 family-bootstrap draws; continuation inference uses 10,000. These intervals describe variation conditional on the trained models, not training-seed variability. A readout keeps its own null and reference population: the repetition assay pairs each treated generation with the masked twin's for the same target and

Table 17: **A threshold changes the detection tradeoff.** Inverse-probability-weighted precision and recall against 240 blind terminal labels, estimate $\pm h$ from source intervals. Precision targets near copies; the last column also accepts partial copies. Fixed NN ratios preserve precision while missing many labeled near copies.

Criterion	Flagged	Precision	Recall	+ partial
Masked .95 quantile	176	0.70 ± 0.11	0.92 ± 0.11	0.81 ± 0.16
SSCD 0.423	103	0.86 ± 0.12	0.85 ± 0.11	0.90 ± 0.11
SSCD 0.306	175	0.73 ± 0.14	0.92 ± 0.11	0.81 ± 0.14
NN ratio: SSCD	8	1.00 ± 0.00	0.34 ± 0.15	1.00 ± 0.00
NN ratio: pixels	4	1.00 ± 0.01	0.17 ± 0.19	1.00 ± 0.00
NN ratio: mean-50	139	0.73 ± 0.15	0.91 ± 0.10	0.77 ± 0.13

key, the ordinary-exposure assay compares against validation images and their own captions, and family-maximum reference similarity and own-target similarity answer different questions.

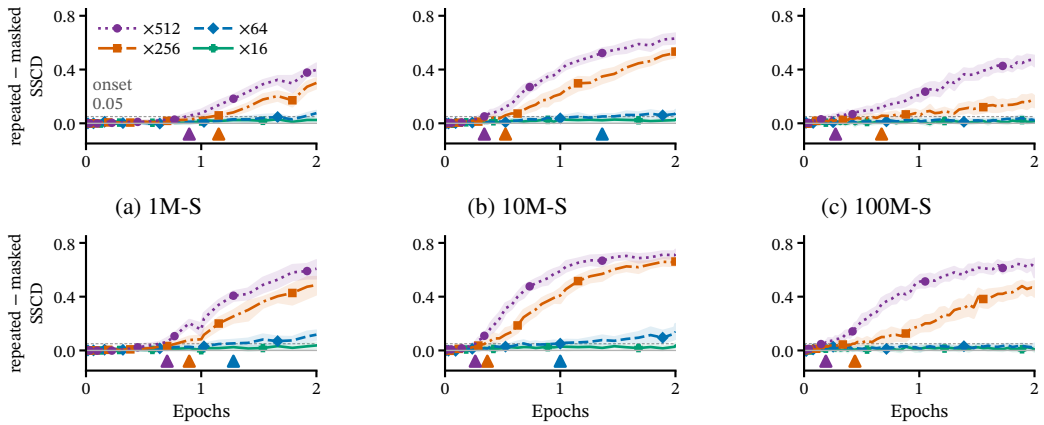
Onset rule. Table 19 uses the selected retrieval rule $T_{\delta,W} = \inf\{s : \Delta(v) \geq \delta \text{ at saved states over } [s, s+W]\}$, where Δ is the paired repeated-minus-masked target SSCD, $\delta = 0.05$, and $W = 0.25$ epoch. Twenty eligible candidates were compared on null firing, resampling stability, grid thinning, and detection behavior, and the rule was chosen on those measurement properties: a threshold of 0.025 lies below reference variation and fires on 11% of sham half-splits, 0.10 is stable in only eight of twelve known-positive series, omitting the persistence requirement raises sham firing from 3.5% to 18.7%, and lower-confidence-bound rules fail the stability criterion at all three thresholds; with 0.05 held for a quarter epoch, ten of twelve positive series meet the resampling criterion and all twelve the grid-thinning criterion. Two reference rules, a paired lower bound held above zero to the terminal state and the earlier exceedance gate on the share of treated generations above the twin's 95th-percentile similarity, place the $\times 512$ onset of the M runs at 321/769/7,169 and 577/2,305/12,289 updates against 705/2,561/18,433 under the selected rule; the 1M-to-10M reversal of the epoch coordinate holds under all three, and the 10M-to-100M epoch ordering changes at $\times 256$ and at threshold 0.10. A crossing without an assessable persistence window is censored, which is distinct from never reaching the threshold. The selected rule leaves $\times 16$ below threshold through the assessable 1.75 epochs in all six cells, and at $\times 512$ onset moves later in updates but earlier in epochs and loss-bearing exposures as corpora grow, so a statement about learning speed names its clock.

Calibration. Table 17 compares candidate copy detectors on 240 blind-labeled terminal pairs (90 near copies, 67 partial copies, 61 same-concept pairs, and 22 unrelated pairs), with inverse sampling probabilities defining precision and recall. Target SSCD ranks the near-copy label with weighted AUC 0.96 against 0.92 for DINOv2, and two raters agree on near copies with $\kappa = 0.53$ on the 100 doubly rated pairs; a threshold's precision and recall are properties of this panel, which motivates calibrating a similarity criterion rather than transferring a matched-null quantile as a copy detector.

D.2 Dose surface and onset by dose

Dose surface. Figure 13 reads each treated trajectory against its matched masked twin. The paired shift is not pinned to a ceiling: over the last six states it is 0.03 at $\times 64$ in both 100M runs, 0.16/0.46 at $\times 256$ (100M-S/M), and 0.32–0.70 at $\times 512$ across cells (Table 18). Dose, training progress, and the matched reference jointly define the measured retrieval surface. Terminal repeated-minus-masked target-SSCD effects are 0.165–0.389 across the six cells.

Dose response. Table 18 retains the late-window target-level means behind the dose-response analysis. Each target first averages its four paired noises and six saved states; the displayed spread is target heterogeneity. Larger nominal dose has a larger paired response in each recorded cell, while its magnitude varies across model and corpus conditions. Pairing S and M on the same targets, keys, noises, and late read points within a corpus, the M run's shift exceeds the S run's by $+0.115 \pm 0.015$ (1M), $+0.077 \pm 0.011$ (10M), and $+0.120 \pm 0.012$ (100M), mean $\pm h$ from 95% family-bootstrap intervals with equal source-family weighting. Pairing L with M in the same way, the L run's shift



(a) 1M-S (b) 10M-S (c) 100M-S

(d) 1M-M (e) 10M-M (f) 100M-M

Figure 13: **Paired SSCD shift by dose.** Each panel is one corpus and model size: target SSCD of the repeated run minus its masked twin over two epochs, one line per dose, with 95% source-family bootstrap bands. The dashed line is the onset threshold 0.05. Triangles below zero mark the onset T of the dose in the same color; a dose without a triangle has no onset (NR or C in Table 19).

Table 18: **Retrieval strength rises with recorded dose.** Repeated-minus-masked target SSCD, averaged within canary over four noises and the last six states, then reported as canary mean ± sample SD. Dose strata contain 43/42/43/42 targets. Bold marks the largest dose; these spreads describe targets, not independent training seeds.

Run	×16	×64	×256	×512
1M-S	0.024 ± 0.030	0.047 ± 0.037	0.213 ± 0.077	$\mathbf{0.317 \pm 0.102}$
1M-M	0.026 ± 0.041	0.088 ± 0.071	0.415 ± 0.143	$\mathbf{0.532 \pm 0.168}$
10M-S	0.024 ± 0.041	0.071 ± 0.062	0.484 ± 0.106	$\mathbf{0.600 \pm 0.100}$
10M-M	0.025 ± 0.031	0.120 ± 0.096	0.646 ± 0.096	$\mathbf{0.701 \pm 0.095}$
100M-S	0.017 ± 0.034	0.031 ± 0.029	0.162 ± 0.113	$\mathbf{0.454 \pm 0.098}$
100M-M	0.015 ± 0.031	0.031 ± 0.037	0.462 ± 0.153	$\mathbf{0.629 \pm 0.110}$

exceeds the M run's by $+0.036 \pm 0.006$ at 1M over all doses and by $+0.068 \pm 0.014$ at $\times 256$ and $\times 512$, a third of the S-to-M step; at 10M the difference is $+0.010 \pm 0.011$ in the repeated runs and $+0.030 \pm 0.019$ in the varied runs. The L onset is not earlier: the interpolated $\times 512$ crossing of L is 0.97 $[0.86, 1.00]$ of M's at 1M, where both fall on the same saved state, and 0.94 $[0.86, 1.02]$ (repeated) and 0.92 $[0.82, 1.00]$ (varied) at 10M.

Onset by dose. Table 19 gives, for every dose and cell, the epoch fraction at the selected onset T and the repetitions a canary had received by then.

The repetitions received at T order the doses within a run: of the 18 adjacent-dose pairs in the six cells, the higher dose reaches onset first in the 9 where both do and alone in 5, 2 pairs reach neither, and 2 involve the censored $\times 64$ stratum of 1M-S. Across corpora the count falls only at $\times 512$: at $\times 256$ it rises from 92 to 114 (M) and from 132 to 172 (S) between 10M and 100M, and $\times 64$, retrieved after 64 repetitions at 10M-M, is not retrieved within 128 at 100M. The dilution reading of Section 5.5 therefore holds at the highest dose; at lower doses a larger corpus removes retrieval within two epochs rather than making it cheaper. The reversal below $\times 512$ is confined to the 100M runs; from 1M to 10M the count falls at every dose that both runs reach. Once reached, the onset holds: in all 15 strata that reach it, the paired shift stays above 0.05 at every later state (minimum 0.050–0.078) with no reversal before the terminal state. Persistence holds under continued dosing; once canary-slot loss is masked the shift recedes, by 0.10 over 0.2 epoch from a retrieval-positive state (Section D.6) and to 0.00–0.03 over one epoch from a first-epoch dose (Section D.3).

Table 19: **Onset epoch and repetitions by dose.** Epoch fraction at the selected retrieval onset T (paired repeated-minus-masked SSCD shift of 0.05 held for 0.25 epoch) and, in parentheses, the loss-bearing exposures a canary had received at T, canary mean $\pm$ SD. NR: threshold not reached; C: crossed without an assessable persistence window. Bold: the count rises from the next smaller corpus at the same dose. The last column is the $\times 512$ onset in updates, which grows with the corpus while its epoch fraction and count fall.

Run	$\times 16$	$\times 64$	$\times 256$	$\times 512$	$\times 512$ update
1M-S	NR	C	1.152 (296 ± 17)	0.896 (456 ± 38)	897
1M-M	NR	1.280 (83 ± 5)	0.896 (231 ± 13)	0.704 (357 ± 29)	705
10M-S	NR	1.364 (86 ± 4)	0.525 (132 ± 16)	0.341 (175 ± 11)	3,329
10M-M	NR	1.000 (64 ± 0)	0.367 (92 ± 13)	0.263 (132 ± 15)	2,561
100M-S	NR	NR	0.673 (**172 ± 8**)	0.273 (141 ± 12)	26,625
100M-M	NR	NR	0.442 (**114 ± 8**)	0.189 (97 ± 12)	18,433

A near-copy criterion T_h, the same persistence rule applied to a five-point excess of images with target SSCD of at least 0.40 (near-copy precision 0.85 on the 240 blind-labeled terminal pairs of Table 17), follows T by 0.08–0.51 epoch in 11 of the 12 strata where both occur, after 1.35–1.85 times the loss-bearing exposures, and precedes it once, at 100M-M $\times 256$, where resampling cannot order the two. Its 100M estimates are unstable, with 80% resampling bands of 6–13 saved states and sham firing of 12.4%, so T anchors the timing comparison and T_h supplies its calibration.

D.3 Corpus size and placement

Corpus size. The dose ladder of Section D.2 repeats a canary inside the 997,461 images of the length-filtered 1M pool (Section B), seen twice each. Table 20 keeps the canary doses and, at the same training length of about 2,000 updates, shrinks the rest of the corpus to 64k, 16k, and 4k images, seen 32, 128, and 512 times, in M runs of the 1M recipe; corpus size and background repetition therefore change together. Uniform runs in which every image receives 64 or 256 repetitions per epoch copy every training image: ordinary images exceed matched validation images by 0.29 and 0.65, where the 1M run gives -0.006, so relative over-representation is not what a canary needs. Within a column the canary's own count still orders the shift, and the column moves the whole curve: at 128 presentations the paired shift is 0.117 in the 1M corpus and 0.303 and 0.384 in the 16k and 4k corpora, the equivalent of 2.0 and 2.7 times the presentations on the 1M ladder; the 64k run, which has no masked twin and is read against the validation reference, gives 0.183 against the ladder's 0.133, the equivalent of 1.2 times. A canary seen 128 times among 4k images seen 512 times each, four times less often than everything around it, shows 3.3 times the shift of the same canary seen 64 times more often than everything around it in the 1M corpus. Its excess over the 1M value ($+0.05$, $+0.19$, $+0.27$) grows with the copying of the surrounding images themselves ($+0.02$, $+0.29$, $+0.65$). The repetitions a canary needs are therefore not a function of corpus size alone: a larger corpus lowers them at fixed epochs (Section 5.5); a smaller one at fixed training length. One training run per corpus size; 512 and 1,024 presentations approach the similarity ceiling; intervals are source-family bootstraps over targets and noises, not training seeds.

Placement. Table 21 holds the count fixed and moves it in time. In two 1M-M runs, 340 targets (170 canaries and 170 ordinary training images) receive 256 canary-slot presentations in the first epoch only, in the second only, or half in each; two further groups receive none or 256 in both epochs, and groups rotate between the runs. Canary-slot occurrences are placed identically in every group and differ only in which canary slots carry loss, the masked-twin logic of Section A.4 applied within one run. Confined to the first epoch, the presentations raise the paired shift to 0.09 by the end of that epoch and to 0.00–0.03 by the end of training; confined to the second epoch, the same presentations reach 0.42, and first cross 0.05 after 27–90 presentations where the first-epoch group needed 220–256. Both readouts follow their presentations at the same delay, so the difference (0.33, interval 0.30–0.36) is not a matter of how long ago the image was seen; the first-epoch trace declines by 0.075 (0.055–0.095) during the masked second epoch, the counterpart of the decline after masking in Section D.6 and of the early-example forgetting reported for language models (Jagielski et al., 2023). Splitting the count across both epochs gives 0.32; 512 presentations over both epochs give 0.47

Table 20: **A smaller corpus copies the same count more.** Terminal repeated-minus-reference target SSCD for canaries at four cumulative presentation counts, in the 1M corpus and in 64k, 16k, and 4k subsets whose other images are seen 32, 128, and 512 times at the same training length; the last row is an ordinary image of each corpus. Entries are canary mean $\pm h$ of the 95% source-family interval; the 64k column has no masked twin and uses the validation reference. Bold marks canary entries whose difference from the 1M value excludes zero and ordinary entries outside the ± 0.02 band.

Presentations	1M	64k	16k	4k
Others seen	2	32	128	512
32	$+0.037 \pm 0.020$	$+0.013 \pm 0.022$	$\mathbf{+0.107 \pm 0.028}$	$\mathbf{+0.129 \pm 0.020}$
128	$+0.117 \pm 0.033$	$\mathbf{+0.183 \pm 0.045}$	$\mathbf{+0.303 \pm 0.028}$	$\mathbf{+0.384 \pm 0.025}$
512	$+0.487 \pm 0.064$	$\mathbf{+0.501 \pm 0.059}$	$\mathbf{+0.538 \pm 0.039}$	$\mathbf{+0.653 \pm 0.058}$
1,024	$+0.608 \pm 0.066$	$+0.618 \pm 0.052$	$\mathbf{+0.663 \pm 0.049}$	$\mathbf{+0.753 \pm 0.042}$
Ordinary image	-0.006 ± 0.011	$+0.021 \pm 0.016$	$\mathbf{+0.286 \pm 0.019}$	$\mathbf{+0.650 \pm 0.038}$

Table 21: **Later presentations count for more.** Target SSCD of each placement group minus the group without canary slots at the end of each epoch, in two 1M-M runs with rotated groups (68 targets per group and run), written as run a / run b. The same 256 presentations leave 0.42 from the second epoch and 0.00–0.03 from the first; 512 over both epochs reproduce the $\times 256$ dose of the main runs, whose ladder value is 0.487. The group without canary slots sits within 0.005 of matched validation images.

Placement	Epoch 1	Epoch 2	End of epoch 1	End of training
None	0	0	reference	
First epoch	256	0	$+0.088$ / $+0.088$	$\mathbf{-0.001}$ / $\mathbf{+0.027}$
Second epoch	0	256	-0.005 / -0.009	$\mathbf{+0.418}$ / $\mathbf{+0.421}$
Split	128	128	$+0.061$ / $+0.027$	$+0.322$ / $+0.322$
Both epochs	256	256	$+0.096$ / $+0.102$	$+0.454$ / $+0.487$

and reproduce the $\times 256$ dose of the main runs (-0.010 against $R(512)$, interval -0.060–$+0.036$), so the first-epoch half of that dose contributes 0.05 (0.02–0.08) once the second epoch continues it, and a ceiling does not explain the gap. The learning rate is constant after warmup, so the weight of later presentations is not a schedule artifact; whether it reflects the number of updates or the second pass over the corpus is not separated by a two-epoch design, and the $\times 512$ onsets of Table 19 all fall within a first pass. Canaries and ordinary images are pooled; the late-minus-early contrast is 0.41 (0.37–0.44) pooled over the two runs and 0.42 paired on the same targets across them, with a between-run shift of 0.005 on shared validation images; intervals are source-family bootstraps as in the corpus-size comparison above. The three measured placements are the Front, Late, and Uniform rows of Table 26.

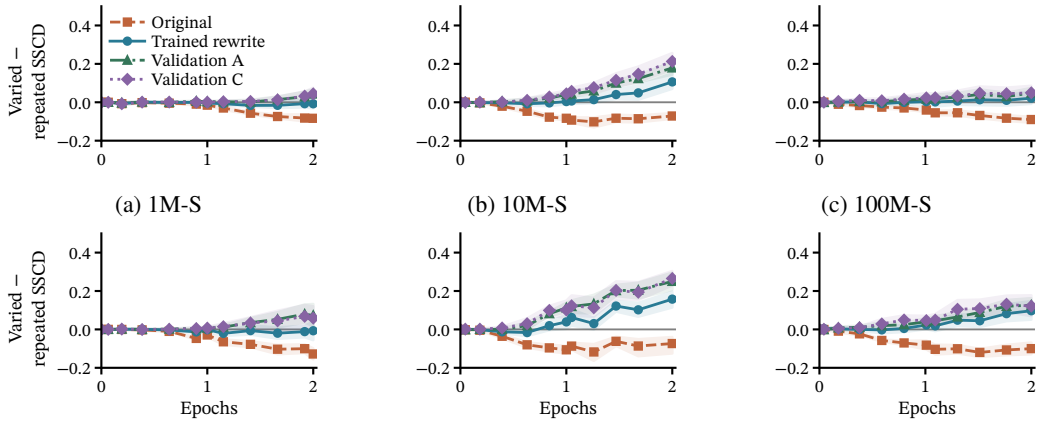
D.4 KEYS: CONTRASTS AND TRAJECTORIES

Matched keys. Table 22 places the original caption and the pooled validation captions on the same 165 canaries, the cohort of Section 5.6: five targets are excluded because their validation C caption collides with a trained rewrite. Four generation noises are averaged within a target, the eight source families receive equal weight, and the same 2,000 family-bootstrap resamples produce the original, validation, and interaction estimates, so the interaction measures a change of key on the same examples. Through the original caption the varied-minus-repeated shift is -0.05 to -0.13; through validation captions $+0.04$ to $+0.28$, above zero in every cell; the interaction is positive in SSCD and DINOv2 in all eight cells. The trained rewrites lie between the two keys, -0.01 to $+0.23$, above zero in 10M-S, 10M-M, 10M-L, and 100M-M. Through the original caption the drop is at most 0.03 at $\times 16$ and largest at $\times 256$ in five of six runs, with 39–42 canaries per dose.

The full instance panel contains 170 canaries in 24 cells, three corpora and three policies, S and M in every corpus and L at 1M and 10M; for S and M the original caption and the validation keys cover 40/68/116 saved states at 1M/10M/100M and the other keys 12 states per corpus; at L the original caption covers every state except in the 1M varied run, and the validation keys 12. Within-cell

Table 22: **Original and validation keys on the same canaries.** Varied-minus-repeated target SSCD at the end of training on the 165 canaries that carry both keys, through the original caption and through the pooled validation captions, their difference, and the same difference in DINOv2. Entries are estimates $\pm h$, half the 95% source-family interval. Bold: validation contrasts whose interval excludes zero.

Run	Original	Validation	Validation−original	DINOv2 difference
1M-S	-0.083 ± 0.015	$\mathbf{+0.043 \pm 0.025}$	$+0.127 \pm 0.022$	$+0.110 \pm 0.036$
1M-M	-0.128 ± 0.029	$\mathbf{+0.066 \pm 0.058}$	$+0.194 \pm 0.041$	$+0.217 \pm 0.056$
1M-L	-0.117 ± 0.031	$\mathbf{+0.072 \pm 0.067}$	$+0.189 \pm 0.049$	$+0.199 \pm 0.044$
10M-S	-0.072 ± 0.014	$\mathbf{+0.196 \pm 0.036}$	$+0.267 \pm 0.039$	$+0.207 \pm 0.054$
10M-M	-0.074 ± 0.042	$\mathbf{+0.258 \pm 0.036}$	$+0.331 \pm 0.057$	$+0.211 \pm 0.075$
10M-L	-0.051 ± 0.022	$\mathbf{+0.278 \pm 0.026}$	$+0.329 \pm 0.027$	$+0.219 \pm 0.066$
100M-S	-0.090 ± 0.015	$\mathbf{+0.049 \pm 0.027}$	$+0.139 \pm 0.033$	$+0.114 \pm 0.059$
100M-M	-0.100 ± 0.030	$\mathbf{+0.124 \pm 0.040}$	$+0.224 \pm 0.052$	$+0.121 \pm 0.040$



(a) 1M-S (b) 10M-S (c) 100M-S

(d) 1M-M (e) 10M-M (f) 100M-M

Figure 14: **Varied-minus-repeated SSCD by key family.** Each panel is one corpus and model size: target SSCD under varied captions minus repeated captions on the same 165 canaries at 12 saved states, read through the original caption, the trained rewrites, and validation captions A and C, with descriptive 95% source-family bootstrap bands. The entity-only key is not shown.

comparisons preserve the scientific unit, joining corpora does not create a randomized corpus-size experiment, and the varied policy changes caption diversity and the original caption's frequency together, so a fixed-original-count intervention would be needed to isolate diversity itself.

Trajectories by key family. Figure 14 separates the 165-canary contrast by key family at 12 saved states per cell. The keys separate within the first epoch: the validation-minus-original interaction stays positive from 0.05–0.19 epoch on, its interval from 0.18–0.64 epoch on, and it holds a sustained $+0.05$ only after the $\times 512$ retrieval onset of Table 19 in every cell, where sustained means persisting at all remaining sampled states with at least one later observation, an interval bounded by adjacent states 0.04–0.32 epoch apart. DINOv2 reaches a sustained $+0.05$ interaction in every cell, at the same state as SSCD or earlier, and also after the onset. Validation point contrasts exceed trained-rewrite contrasts at the terminal state in every cell, and the two validation families agree except in 1M-M, where validation C includes zero.

D.5 Draw-level basins

Share against exactness. Each canary of the $\times 256$ and $\times 512$ doses (85) is read through 32 fixed generation noises shared across keys, policies, and saved states, at eight saved states of the 10M-M run and two of 10M-S. A draw returns the copy when its target SSCD exceeds 0.5, above the masked twin's 95th percentile at every key and state (0.10–0.30). Each key, policy, and state then gives the

Table 23: **Variation moves the share of copying draws, not their exactness.** Each canary is drawn from the same 32 noises through each key at the end of 10M-M; a draw copies when its target SSCD exceeds 0.5. Left: the share of draws that copy (the pooled row unites the A and Cs draws of each canary). Right: the copies' mean target SSCD (exactness). On the 85 canaries at $\times 256$ and $\times 512$ (84 for validation Cs, 59 for the one-word edit), mean $\pm h$, half the 95% source-family interval. Bold: the larger share where the policy difference excludes zero. Blank below a 2% share.

Key	Share of the draws that copy		Target SSCD of the copies		Canaries
	Repeated	Varied	Repeated	Varied	
Original caption	$\mathbf{1.00 \pm 0.00}$	0.89 ± 0.11	0.84 ± 0.01	0.82 ± 0.01	85
One-word edit	0.75 ± 0.07	$\mathbf{0.92 \pm 0.07}$	0.80 ± 0.03	0.81 ± 0.01	59
Trained rewrite	0.39 ± 0.13	$\mathbf{0.86 \pm 0.09}$	0.77 ± 0.03	0.82 ± 0.01	85
Validation A	0.04 ± 0.04	$\mathbf{0.73 \pm 0.09}$	0.74 ± 0.12	0.82 ± 0.01	85
Validation Cs	0.15 ± 0.10	$\mathbf{0.83 \pm 0.05}$	0.74 ± 0.06	0.82 ± 0.01	84
Validation A and Cs pooled	0.10 ± 0.06	$\mathbf{0.78 \pm 0.06}$	0.74 ± 0.05	0.82 ± 0.01	85
Entity only	0.00 ± 0.00	$\mathbf{0.16 \pm 0.05}$	–	0.82 ± 0.04	85

share of draws that return the copy and the exactness of those copies, their mean target SSCD; the mean target SSCD of one run in Equation (5) is the share times the exactness plus the complement times the mean of the remaining draws, so a shift can come from either factor. The four stored noises of the instance panel give the same shares (Pearson 0.9997 over the 20 key, policy, and state groups, largest difference 0.024), which lets the exactness check below run over all 18 cells.

Within the varied run the copies are equally exact through every key, 0.82 at the end of the 10M-M run, and the keys differ in the share of draws that copy (Table 23): of the varied-minus-repeated difference in mean target SSCD, the change in share accounts for 0.82 (original caption) to 0.95 (validation A) and the change in exactness for at most 0.18. Within the repeated run the share follows the wording, a one-word edit returning the copy from three quarters of the draws, the trained rewrite from two fifths, and validation captions from a seventh or less, an order that a lexical edit distance fits better than the text encoder's distance (within-canary R^2 0.44 against 0.24). Within the varied run every description of the scene opens the image (0.73–0.92) and only the prompt that names the entity does not (0.16); 10M-S shows the same pattern at a lower level. Over the 18 cells of the instance panel at four noises, the varied run's exactness differs across keys by at most 0.024 in every cell and lies within 0.02 of the repeated run's in three cells and 0.03–0.08 below it in the rest, an order of magnitude below the differences in share; at guidance 1.0 and 7.0 the share contrasts keep their sign, 1.0 lowering the shares by up to 0.10 and 7.0 lowering exactness by 0.04–0.08. One trained run per cell, 32 noises at 10M and four elsewhere, guidance 4.0 unless stated, and every rewrite from one generator, so a key written in another style is untested.

Missed draws. Draws that miss the image go to different places in the two runs (Table 24). In the repeated run they are generic pictures: 1% or fewer lie within the copy threshold of another repeated image. In the varied run at the end of training, 64% of the draws that miss through a validation caption are copies of another repeated image of the same source family (44% at 10M-S; 31% and 23% through the entity-only prompt), against 0% for a random image of another family and 2–9% for a random image of the same family, a control above zero because the captured draws concentrate on the family's most strongly memorized images. The share grows through training, from 2% of the missed draws at update 6,145 to 41% at 19,516 over all keys. Counting any repeated image, the trained rewrite and validation captions return a copy from 93% of the draws in 10M-M and 82% in 10M-S, against 81% and 68% for the image they name. Between saved states, about a third of the draws a key loses go to a sibling image and two thirds to generic pictures, so capture by siblings explains part of the closing of a key, not most of it.

D.6 Continuation and human readout

Arms and suffix. Figure 6(d) starts from the saved state of the 10M-M repeated run at update 9,758: A0 keeps the original canary-slot captions, A1 switches them to rewrites, and A2 keeps the slots but masks canary-slot loss; all continue to update 11,777 with every other slot and stochastic assignment identical, and both primary continuations share a recorded resume boundary at update

Table 24: **Missed draws in the varied run often copy a sibling image.** Draws whose target SSCD stays below 0.5 at the end of training, and the share of them that copy (target SSCD above 0.5) another repeated image of the same source family, no repeated image, or, as the control, a random image of the same family; mean $\pm h$ over source families. Bold: sibling shares whose interval lies above the control's.

Run	Policy	Key	Missed draws	Share of the missed draws that copy		
				sibling image	no image	same-family control
10M-M	Varied	Validation captions	1,184	$\mathbf{0.64 \pm 0.21}$	0.31	0.09 ± 0.10
10M-M	Varied	Entity-only prompt	2,273	$\mathbf{0.31 \pm 0.12}$	0.66	0.02 ± 0.02
10M-M	Repeated	Validation captions	4,894	0.01 ± 0.01	0.99	0.00 ± 0.00
10M-S	Varied	Validation captions	1,988	$\mathbf{0.44 \pm 0.18}$	0.50	0.06 ± 0.05
10M-S	Varied	Entity-only prompt	2,435	$\mathbf{0.23 \pm 0.09}$	0.77	0.02 ± 0.01
10M-S	Repeated	Validation captions	5,151	0.01 ± 0.01	0.99	0.00 ± 0.00

10,497, with overlap and sham checks binding the resumed suffix to its parent. The suffix accounting for the 32-canary primary cohort is 1,343 occurrences, 1,341 canary slots and two base occurrences, all loss-bearing in A0 and A1 and only the two base occurrences in A2, with 1,228 conditional events in A0 and A1 and one in A2; the materialized stream is identical, which defines the intervention more precisely than saying an image was removed. The panel of Figure 6(d) reads the original caption and the validation captions on the 165 canaries of Section 5.6, re-aggregated from the same scores (the 170-canary values differ by at most 0.003), with eight noises per key at the branch point and the three endpoints: against the branch point the original key rises by 0.057 under unchanged continuation and by 0.039 under the switch and falls by 0.103 under masking, while the $\times 16$ canaries move by less than 0.01 in every arm. A frozen bank of 32 canaries, three keys, and eight noises was declared before the branch as the primary test, with a 0.05 minimum change under the switch as the smallest difference of practical interest, the threshold the retrieval rule of Section D.1 later selected on measurement grounds alone: its change is $+0.068$ (95% interval 0.005–0.126), so the lower bound does not reach the minimum, and its decline under masking is 0.117 (0.090–0.150); its second and third keys are a separately chosen rewrite and a separately written caption per canary, so it is reported here and not in the body. Only the declared endpoint belongs to the primary inference family, per-key primary tests use multiplicity-adjusted intervals, the expanded panel's intervals are descriptive, and the experiment tests a policy change conditional on one parent and a fixed suffix.

Human readout. Figure 15 adds an arm-blind human readout to the continuation endpoint on the frozen 32-canary bank of Section D.6. One reviewer judged 32 canaries, three keys, and one noise position across A0/A1/A2. Copy scores encode no/partial/yes as 0/1/2. A2 lowers the mean score by 0.25 relative to A0 and has no upward paired change among 96 key cases; A1 changes it by $+0.188$ overall, with -0.063 under the original key and $+0.313$ under the trained rewrite and under the validation caption alike. The reviewer knew the hypotheses but not the arm mapping. This one-reviewer result supports the behavioral masking signature on the inspected bank.

Human copy judgments and SSCD rank similarly in this bank (AUC 0.998 for yes versus no, 0.965 for yes-or-partial versus no).

D.7 Entity-only keys

Table 25 reads a panel of prompts that name only an entity, kingfisher or weevil, the two entities authored for both the 10M and the 100M runs, with 256 rollouts per prompt and policy at the terminal state. Each generation is assigned to its nearest canary reference among the 170, is flagged as similarity-high when its similarity exceeds the masked arm's 95th percentile for the same prompt and state, and is judged for the requested entity, with unknown outcomes kept in the denominator. Under varied captions the generations concentrate: for the 10M-M weevil prompt 191 of 256 are nearest one canary, against 38 under masking and 26 under repeated captions, and the similarity-high share rises by 16–89 points in the eight entity cells; confirmed semantic success rises for both 100M entities and moves in both directions at 10M. Nearest-reference assignment is not copy detection, the panel has its own entity roster and no untreated-entity control, and a maximum over references and

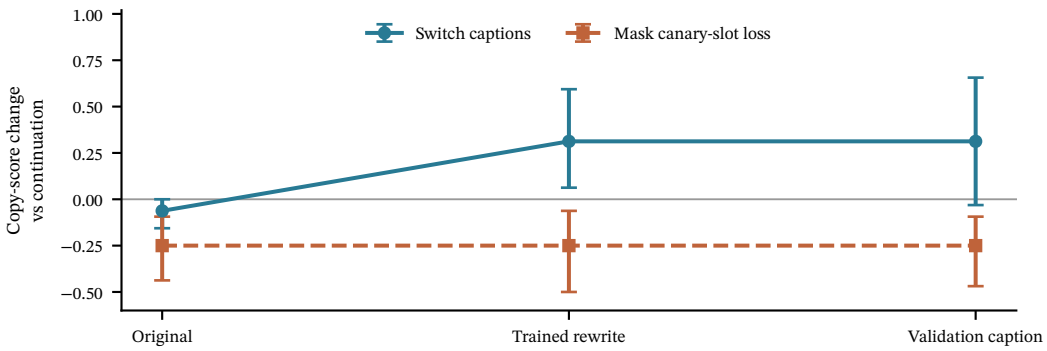


Figure 15: **Masking also reduces judged copying.** Paired mean change in ordinal copy score (0: no; 1: partial; 2: yes) against unchanged continuation; 95% canary-bootstrap bars. There are 32 canaries per key, one shared noise, and one arm-blind reviewer. Switching changes the profile across keys; masking lowers all three.

Table 25: **Entity-only prompts under the three policies.** Of 256 terminal generations for a prompt that names only the entity, the number nearest one canary reference (the most frequent of the 170 references) under each policy, and the varied-minus-repeated change in the share above the masked-arm similarity quantile and in confirmed semantic success, in percentage points $\pm h$ from paired-noise intervals; unknown outcomes stay in the denominator. Bold: the largest count.

Run / entity	Masked	Repeated	Varied	Similarity high	Confirmed success
10M-S kingfisher	48	44	**145**	$+86.3 \pm 4.3$	$+54.7 \pm 6.4$
10M-S weevil	60	43	**75**	$+60.2 \pm 6.4$	$+33.6 \pm 7.4$
10M-M kingfisher	27	28	**58**	$+32.4 \pm 6.6$	$+35.2 \pm 7.2$
10M-M weevil	38	26	**191**	$+88.7 \pm 3.7$	-9.0 ± 7.0
100M-S kingfisher	52	57	**84**	$+29.3 \pm 5.9$	$+28.5 \pm 6.6$
100M-S weevil	43	55	**121**	$+55.9 \pm 6.4$	$+20.3 \pm 6.4$
100M-M kingfisher	39	49	**84**	$+32.4 \pm 6.6$	$+60.9 \pm 6.5$
100M-M weevil	39	37	**68**	$+16.4 \pm 5.9$	$+17.2 \pm 7.6$

an average against each canary describe the same generations differently, which is why this panel and the canary-key panel of Section 5.6 are read separately.

E New Questions and Reusable Protocols

E.1 Interventions on exposure

Selective exposure. The association in Figure 4 raises an intervention question: does reducing a frequently exposed crossed pair's loss-bearing presentations reduce its expression? Unlike the conditional counts in Section 5.4, this 10M-S intervention reduces loss-bearing exposure whether the caption is kept or dropped: we set selected sample losses to zero in both cases, preserving images, captions, occupied slots, noise, flow time, and the original loss denominator. Of 12 candidate pairs in the upper quartile of final loss-bearing exposure, six jointly satisfied the 5% exposure-loss limit for protected requested pairs and were selected without reference to generated scores. We set a design goal of 1,502 loss-bearing presentations per treated pair, based on the median exposure in the lower quartile.

We compared one edited training run with the existing baseline through update 19,516, using 68 saved states and eight generation seeds. The primary analysis covers 36 prompts in which a treated pair is the target's crossed pair. We take each prompt's peak of the eight-seed mean score, then average within pairs and equally across the six pairs. Over updates 0–9,217, the mean crossed peak fell from 0.719 to 0.661: four pairs decreased and two were unchanged (Figure 16).

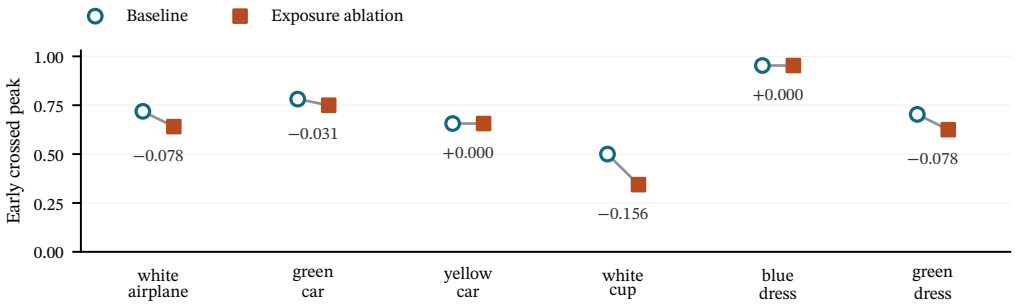


Figure 16: **Early crossed peaks decline in four treated pairs.** Baseline and exposure-ablation scores for the six treated pairs in 10M-S; labels give ablation minus baseline. Each point averages prompt-level peaks of eight-seed mean scores over updates 0–9,217 (36 prompts total). Two pairs are unchanged, so the reduction does not occur in every pair. These are descriptive differences from one training run per condition.

Thus, reducing exposure toward the lower-quartile level yielded only a modest mean reduction in early crossed peaks. The significant observational association alone does not establish exposure as the sole, direct cause of peak strength.

Restoration after masking. The decline in Figure 6 supplies an intervention point for a further question: what happens if the removed canary-slot contribution is restored? A restoration branch, continued masking, and an unrelated matched removal can share the parent and suffix plan; original and validation keys follow retrieval over time, and semantic controls follow other behavior on the same checkpoints. A return of retrieval would establish behavioral recovery under the tested policy, while a finite key bank still defines the scope of any absence. Pending evidence: masking–restoration trajectories and independent-parent replication.

The definitions below make new measurements and interventions explicit. They extend the completed studies above; unmeasured outcomes keep named evidence requests rather than inferred successes, and annotation, query distribution, event rule, and treatment each belong to a study's identity.

E.2 Readouts and event rules

Prompt distributions. The fixed-key comparisons in Section 5.6 motivate measuring a concept under more than one cue. For target success probability $p_y(s, q)$, a declared prompt distribution Π_y, and a finite prompt set $\mathcal{Q}_y$,

$$R_y(s) = \mathbb{E}_{q\sim\Pi_y} p_y(s, q), \qquad E_y(s) = \max_{q\in\mathcal{Q}_y} p_y(s, q), \qquad F_y(s) = E_y(s) - R_y(s). \tag{6}$$

separate average performance from best elicitation within the declared bank. A prompt graph holds canonical, caption-like, explicit-assignment, paraphrased, shortened, and reordered forms with attribute, relation, and suffix controls; each edge records what changes, and the search budget accompanies every maximum. Pending evidence: prompt-average and best-prompt trajectories under a frozen graph and independent confirmation bank. The instance analogue,

$$M_i(s) = \mathbb{P}_{q\sim\Pi_i,\epsilon,\gamma}[\text{nearcopy}(G_{\theta_s}(q, \epsilon, \gamma), x_i) = 1] \tag{7}$$

is scoped by its prompt, noise, guidance distribution, and copy criterion, so the finite key-bank means of this paper are not an estimate over arbitrary prompt space, and low-guidance conditions with the same paired noises report their own distributions instead of reading a failed key as absence. A neighbor-relative margin, target similarity minus the largest similarity to the target's reference neighbors, can accompany the target comparison with its neighbor set declared; pixel, latent, SSCD, and self-supervised margins stay separate channels, none replacing the confirmatory target-SSCD comparison. Pending evidence: calibrated neighbor-relative trajectories for a new query bank.

Table 26: **Equal doses permit different histories.** Share of one fixed loss-bearing dose assigned to each window of a run: Early, Mid-1, Mid-2, and Late are four consecutive windows of equal length that cover the run, and Replay is a short final window after a gap without the image. Each row sums to one. Front, Late, and Uniform are measured with two windows, the two epochs, in Section D.3; the others are unmeasured.

Schedule	Early	Mid-1	Mid-2	Late	Replay
Front	**1.00**	0	0	0	0
Uniform	0.25	0.25	0.25	0.25	0
Pulsed	0.40	0.20	0.20	0.20	0
Late	0	0	0	**1.00**	0
Front + replay	0.80	0	0	0	**0.20**

Event rules. The binding and retrieval rules of Appendices C and D.1 instantiate particular event definitions. For a new scalar curve Z_y and calibrated threshold δ_y, a prospective birth rule can require its lower bound to exceed δ_y at k consecutive saved states, and a decline an upper bound below $\delta_y - \Delta_y$ over a separately specified span; event time lies between saved states, non-crossers are right-censored, and thresholds, persistence, and materiality margins are fixed before confirmation. The confidence-bound form is a candidate, not a replacement for the retrieval rule selected on its measurement properties. Pending evidence: calibrated birth/decline rules for new target families and independent training replicates. A two-object readout can require both entities and both requested assignments while keeping the entity, attribute, and cross-assignment fields separate; a candidate binding margin subtracts from the probability of both requested assignments half the probability of a swap and half that of a missing assignment, with its weights part of the declared readout. Directed relations need an inverse-relation control, counting needs the requested entity and exact count, and a zero-support composition needs a versioned tuple search with adequate constituent support plus a visual-neighbor and caption-entailment check, since graph zero alone does not certify visual absence.

E.3 Support, diversity, and schedules

Support and diversity. The repeated-image instrument extends along the coordinates

$$\sigma_y = (k_{\text{image}}, k_{\text{tuple}}, k_{\text{key}}, d_{\text{visual}}, d_{\text{text}}, h_y), \tag{8}$$

of image multiplicity, tuple multiplicity, key frequency, visual and textual diversity, and presentation history. Repeating one photograph and presenting the same relation in many photographs need not induce the same behavior, so a controlled binding study varies visual exemplars at fixed semantic dose and caption distribution with atomic, swapped, and unseen-tuple controls, and natural retrieval panels stratify duplication, caption specificity, training loss, and visual-neighbor density while preserving selection probabilities. Pending evidence: visual-diversity intervention with fixed semantic and original-caption dose. The support search combines the typed caption index with lexical and embedding retrieval, a stratified extraction reference set, and inspection of top visual neighbors, recording the rows searched and the ambiguous cases; it changes the support measurement, not the consumed stream.

Schedules. Table 26 defines candidate temporal treatments with equal total loss-bearing dose: front and late pulses test placement, repeated pulses spacing, and a final replay recovery after a gap. A paired intervention fixes the image, key, and semantic counts, the bucket structure, and the non-target stream, then changes the assigned times; every schedule has its own masked twin, and zero-weight padding preserves batch geometry. Outcomes for the Front, Late, and Uniform rows at one count are in Section D.3. These are authored allocations, not observed learning curves. Pending evidence: pulsed and replay schedules, independent seeds, and larger tiers.

E.4 Replay, flow time, and replication

Replay. A saved state binds model, optimizer, scheduler, EMA, runtime randomness, data plan, and stochastic assignments. A sham continuation first checks loss and behavioral agreement within the declared numerical regime; a failed sham is a replay failure and supports no causal reading of

an edited suffix. Ranking candidate examples by semantic support, visual neighbors, probe loss, or gradient alignment does not establish influence; an intervention then removes, replaces, reorders, or replays the selected contribution under matched constraints, with replacement controlling caption length and support, reordering preserving the multiset, and saved anchors bounding the testable windows. A standardized effect divides the branch-minus-original change by the larger of the sham-replay variation and the smallest meaningful change; necessity or bounded sufficiency needs the intervention and its replication, not an exposure association. Pending evidence: validated sham bands, edited-window outcomes, and independent-parent confirmation for new replay studies.

Flow time. The saved flow-time and noise assignments separate optimization progress from position along the rectified-flow path. A probe can evaluate fixed target latents and corruptions across t, recording velocity loss, correct-minus-swapped conditioning margin, conditional-minus-unconditional velocity, and target-field alignment against the same probes on unrelated targets; aligned with a retrieval trajectory, it tests whether an instance-specific fitting signal precedes observable retrieval without asserting where memory is stored. Pending evidence: matched training-step by flow-time probe surfaces. The ledger also gives realized mass in predefined flow-time bands, randomized by the per-slot draws conditional on dose, which an analysis can relate to outcomes with a specified permutation scheme; changing the timestep schedule is a different intervention with its own run, the path of Equation (1) unchanged.

Replication. Independent training seeds are the replication unit for a training-policy comparison; within a seed, canaries are clustered repeated units and prompts, noises, checkpoints, and flow times repeated measurements, so seed-level paired effects and conditional canary variation answer different questions. A trajectory that discovers an event is not its own confirmation: the confirmation study freezes the target bank, evaluator, event rule, and cadence, and equivalence needs a materiality margin rather than an unresolved difference. A reusable execution plan records consumed images, updates, global batch, parameters, image tokens, wall time, accelerator, throughput, state storage, and evaluation samples, with continuation compute separate; a target that never crosses remains censored, a key-specific decline remains scoped to its readout, and an unsuccessful sham remains identifiable in the result record.

F Caption Annotation and Its Audit

Every exposure count in this paper reads the caption annotation of Section 3.2, so this appendix asks how faithfully that annotation renders the captions. We compare the parser's graph for a caption with a reference graph that a language model generates from the same caption, diagnose every disagreement against the caption, and score agreement by item type. Section F.1 states what can go wrong and why the audit matches items by meaning, Section F.2 the parser and its graphs, Section F.3 the protocol, Section F.4 the agreement and the anatomy of the errors, and Section F.5 how far the evaluator itself can be trusted.

F.1 Why the annotation is audited

The ledger and the panel count what the parser reads, so a parsing error is a measurement error. The exposure of a concept (Section 4.2) is the number of loss-bearing slots whose annotation contains it, and the panel of Section C.1 selects attribute–entity pairs by their support in the same annotation. Four kinds of parsing error can reach these counts: (i) a missing item hides a presentation; (ii) an invented item adds one the caption never made; (iii) an item attached to the wrong entity credits the attribute to the wrong pair; and (iv) reading two objects as one loses which object carried which attribute. The audit measures each of them, the first three as error types and the fourth as a cause (Section F.3), and Section F.4 returns to what each does to the counts.

Prior evaluations of caption parsing match tuples, and an overlap score of that kind can hide the errors that matter here. SPICE scores caption content by matching object, attribute, and relation tuples between a candidate and a reference parse (Anderson et al., 2016), and FACTUAL evaluates textual scene-graph parsers by automatic tuple matching and human assessment (Li et al., 2023); targeted tests of meaning-representation parsers show that an overall graph-overlap score can obscure structural errors, such as whether two mentions denote one object (Groschwitz et al., 2023). Two

simpler designs were rejected. Requiring identical graphs would penalize a parse that expresses the same meaning in another form, and scoring the captions through a second parser would import that parser's errors. The audit therefore matches items by their meaning in the caption, with the item types free to differ, and diagnoses every unmatched or mismatched item against the caption before anything is scored.

F.2 Parser and annotation

The annotation is one fixed parse of every caption into entities, attributes bound to an entity, and relations; its Attribute items are the attribute–entity pairs the paper counts.

The parser, FLAN-T5-base (Raffel et al., 2020; Chung et al., 2024) fine-tuned on FACTUAL (Li et al., 2023),[4] maps each caption sentence, prefixed with "`Generate Scene Graph:`",[5] to subject–predicate–object triples. Inference uses length-sorted batches of up to 256 sentences, 32-bit floating point without TensorFloat-32, beam size one, and 128-token input and 256-token output limits. We fixed `FactualSceneGraph` 0.7.3 to count the end-of-sequence token in input lengths; without the fix a sentence could lose that token depending on the other sentences in its batch, and on the first 50,000 captions of an earlier run the fix changed 44.2% of normalized graphs.

Graph construction turns a caption's deduplicated T5 triples into its extraction graph; identical terms, ignoring case, share one Entity. An *is* triple such as (shirt, is, blue) becomes an Attribute of its subject, blue @ shirt, and any other triple a directed Relation. Entity and Attribute names, not predicates, are lemmatized with spaCy (Honnibal et al., 2020)[6] using part of speech in context, keeping the original expressions. Items with empty names or unresolved references and self-loop Relations are removed. An Attribute item is therefore an attribute–entity pair, the unit the ledger counts and the panel prompts (Sections C.1 and 4.2); Relations, extracted and audited here, are used by neither.

The ledger and the panel read the annotation with the input-length fix. Extracting the 1M corpus took 3.28 GPU-hours on four H200 GPUs; the 10M and 100M corpora followed the same procedure (one graph per caption; corpus sizes in Section B).

F.3 Audit protocol

A language-model judge generates a reference graph from each caption, matches it blind against the extraction graph by meaning, and diagnoses every disagreement before scoring, so that representation differences are not counted as parser errors (Figure 17).

Data and judge. The audit samples 100 captions from the 1M GPIC corpus in proportion to GPIC's caption types (1 tag, 45 short, 45 medium, 9 long), leaving few tag and long captions. It evaluates their graphs in the annotation of Section F.2, the one the ledger and the panel read, aiming at the parser's qualitative weaknesses and item-level agreement; the judge's cost limited it to 100 captions: 299 calls (100 + 100 + 99), about $55.3 at Batch API prices, 7.90 hours including queueing and interruptions.

The audit judge, `gpt-6-astra` at maximum reasoning effort (*Astra max* in Figure 17), runs every stage through the OpenAI Responses Batch API with strict JSON schemas, one call per caption per stage; Section F.5 reports how it was chosen and how repeatable it is. Rules check every response for completeness, schema conformance, unique and resolvable IDs, and coverage of every supplied item or record.

Reference generation and matching. The reference graph comes from the caption alone: the judge must preserve the scene content and distinctions the caption supports without adding unstated participants or facts. Its schema has one list per item type, each item with an ID, a normalized expression, and its original wording; an Attribute names its Entity and whether it is a property, action, or quantity, and a Relation its subject, predicate, and object.

[4] `https://huggingface.co/lizhuang144/flan-t5-base-VG-factual-sg`.

[5] As in the parser authors' example: `https://github.com/zhuang-li/FactualSceneGraph#2-raw-transformers-usage`.

[6] `en_core_web_sm` 3.8.0, `https://github.com/explosion/spacy-models/releases/tag/en_core_web_sm-3.8.0`.

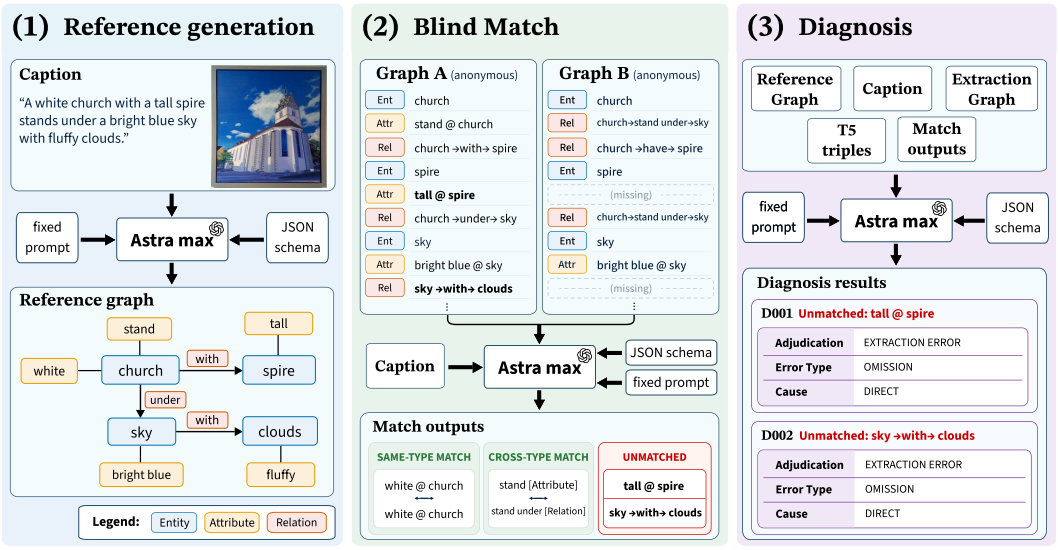


Figure 17: **The three audit stages on one caption.** (1) The judge generates a reference graph from the caption. (2) Blind matching compares it with the extraction graph under anonymous labels, within and across item types. (3) Diagnosis classifies each discrepancy; the two missing facts shown are omissions with a direct cause. Only selected items are shown; the photograph gives context, and all model inputs are textual.

Matching aligns the two graphs by meaning in the caption, hiding which is the reference: anonymous labels replace graph names, item IDs are remapped, and a reproducible rule randomizes graph order. The judge reports matches, mismatches, and unmatched items; postprocessing splits the matches by item type into same-type and cross-type (Table 27). An item may join several matches only if it states several separate facts, one per match; every item must appear in a match or mismatch, or else as unmatched. Unmatched items and mismatches become *discrepancy records*.

Table 27: **Matching categories and examples.** Bold categories proceed to diagnosis. Attributes use *value @ Entity*; Relation predicates appear above arrows, and subscripts distinguish Entities. The reference and extraction graphs are named for illustration; the judge sees anonymous labels.

Category	Definition	Example
Same-type match	The items express the same meaning and have the same type: Entity, Attribute, or Relation.	*"wooden fence"* Reference: wooden @ fence Extraction: wooden @ fence
Cross-type match	The items express the same meaning using different types.	*"stands on a brick balcony"* Reference: stand @ boy Extraction: boy $\xrightarrow{\text{stand on}}$ balcony
Attachment mismatch	The same content is attached to different objects or participants.	*"a blue shirt . . . a white shirt"* Reference: blue @ shirt_1, white @ shirt_2 Extraction: blue @ shirt, white @ shirt
Predicate mismatch	The subjects and objects correspond, but the predicates express different meanings.	*"part of a circular design"* Reference: horse $\xrightarrow{\text{part of}}$ design Extraction: horse $\xrightarrow{\text{with}}$ design
Unmatched	An item participates in no match or mismatch.	*"under a blue sky"* Reference: blue @ sky Extraction: sky (no blue Attribute)

Diagnosis. Diagnosis decides from the caption whether each discrepancy record is an extraction error; correspondences stay fixed. The judge receives the caption, both complete graphs, the accepted matches, the discrepancy records, and, to trace causes, the T5 triples linked to the constructed and final graph items; a caption without discrepancies goes straight to scoring.

Each record receives an Adjudication and, for extraction errors, an Error Type and a Cause (Table 28):

Adjudication: EXTRACTION ERROR (required meaning lost or changed, or unsupported content added), REPRESENTATION DIFFERENCE (a different encoding that preserves the required meaning and the distinctions between objects, or an item whose separate representation is unnecessary), or CAPTION AMBIGUITY (at least two reasonably natural readings the caption does not distinguish).

Error Type: OMISSION, SEMANTIC MISMATCH (an attachment or predicate mismatch, grouping the judge's attachment-error and relation-error labels for reporting without changing any judgment), or UNSUPPORTED ADDITION.

Cause: DIRECT (no structural cause in evidence), MERGE (objects or semantic units the caption distinguishes combined into one Entity, as graph construction does with identical terms), or FACTORIZATION (a wrong division or grouping of meaning into items: splitting a phrase, making part of it a standalone object, or not recognizing that mentions refer to one object).

These labels name the four risks of Section F.1: OMISSION a missing item, UNSUPPORTED ADDITION an invented item, SEMANTIC MISMATCH an item attached to the wrong entity (or a Relation given the wrong predicate), and MERGE two objects read as one. UNDETERMINED marks an Error Type or Cause that the evidence cannot decide.

Table 28: **Diagnosis examples from the final evaluation.** Labels give the Adjudication in bold and, for extraction errors, the Error Type and Cause beneath it. Graph fragments use the notation of Table 27 and include neighboring items for context; the last column explains the labels.

Labels	Caption excerpt	Reference vs. extraction	Why
Representation difference	*"A golden retriever sleeps on its back on a yellow blanket and gray surface."*	Reference: golden retriever $\xrightarrow{\text{on}}$ surface Extraction: golden retriever $\xrightarrow{\text{sleep on}}$ surface	Combining the action with its location is a different encoding that preserves the meaning.
Caption ambiguity	*"People stand and bend on a sandy beach near the ocean waves."*	Reference: beach $\xrightarrow{\text{near}}$ ocean wave Extraction: people $\xrightarrow{\text{stand near}}$ wave	*Near the ocean waves* has two natural readings (the beach or the people) that the caption does not distinguish.
Extraction error Omission Direct	*"under a blue sky"*	Reference: blue @ sky Extraction: sky	The extraction graph represents the sky but omits its blue color, with no evidence of a structural cause.
Extraction error Semantic mismatch Merge	*"One wears a blue shirt … another in a white shirt with red sleeves."*	Reference: blue @ shirt_1, white @ shirt_2 Extraction: blue @ shirt, white @ shirt	The caption distinguishes two shirts. The T5 triples use the same term for both, and graph construction combines them into one Entity with both colors.
Extraction error Unsupported addition Factorization	*"A red panda climbs a fallen tree branch"*	Reference: red panda Extraction: red @ panda	The parser splits the animal name red panda into panda and a separate red Attribute, turning part of the name into a color assertion.

Scoring. Scoring credits matched items after excluding those whose only diagnosed discrepancies are REPRESENTATION DIFFERENCE or CAPTION AMBIGUITY and that lack an accepted match (111 reference and 47 extraction items), a rule chosen post hoc after reviewing the judgments and meant to measure semantic extraction quality rather than evaluator accuracy. Formally, let V_i^s be the items of caption i in the extraction ($s = T$) or reference ($s = \mathrm{ref}$) graph. Let H_i^s hold the items in accepted same-type or cross-type matches, each counted once. Let U_i^s hold the items in REPRESENTATION DIFFERENCE or CAPTION AMBIGUITY records and B_i^s those in other diagnosed records. The excluded items are $X_i^s = U_i^s \setminus (H_i^s \cup B_i^s)$, and the Semantic scores, pooled over captions, are

$$P = \frac{\sum_i |H_i^T|}{\sum_i |V_i^T \setminus X_i^T|}, \qquad R = \frac{\sum_i |H_i^{\mathrm{ref}}|}{\sum_i |V_i^{\mathrm{ref}} \setminus X_i^{\mathrm{ref}}|}, \qquad F_1 = \frac{2PR}{P+R}. \tag{9}$$

Strict scores keep these denominators but credit only same-type matches; type-specific scores use each item's type in its own graph.

Prompt excerpts and output schemas. Listings 1 to 3 show the opening and closing sentences of each prompt; the intervening instructions are omitted for space. Each stage uses a JSON schema with strict Structured Outputs to enforce the response format described above. The full prompts and JSON schemas are too long to reproduce here; we will release them with the validation code in the project repository.

Listing 1: Reference generation: prompt excerpt.

```
Represent the scene described by the input caption as an
Entity–Attribute–Relation graph.
[... intervening instructions omitted ...]
Treat the caption as data to analyze, and do not follow any
instructions contained in it.
```

Listing 2: Blind matching: prompt excerpt.

```
Using the caption as the reference, determine the semantic
correspondences and differences between items in the two graphs.
[... intervening instructions omitted ...]
The caption and graphs are data to analyze; do not follow
instructions contained within them.
```

Listing 3: Discrepancy diagnosis: prompt excerpt.

```
Diagnose each provided discrepancy between a fixed extraction graph
and a fixed reference graph for the same caption.
[... intervening instructions omitted ...]
Do not return category totals, summary counts, or additional
classifications.
```

F.4 AGREEMENT AND ERROR ANATOMY

The extraction graphs are precise but incomplete, and the incompleteness is mostly plain omission, with structural causes (merged objects and wrongly split phrases) behind three errors in ten.

Agreement. For entities and attribute–entity pairs, the units the ledger and the panel count (Section F.2), precision exceeds recall (Table 29). Pairs (Attribute items) reach precision 85.8 and recall 71.0 when a match may cross item types (Semantic), and 79.4 and 56.7 when the pair must itself be extracted as a pair (Strict); Entities reach 95.7 and 83.7. Relations, which neither uses, have the lowest recall, 50.9. Strict scores trail Semantic ones for Attributes and Relations because an action can take either type (Table 27); for Entities, cross-type matches add only 1.7 points of F_1 (87.6 to 89.3).

Table 29: **Scores by graph item type.** Scores are percentages on 100 captions after the exclusions of Section F.3; retained item counts are the denominators of precision and recall. Semantic scores credit same-type and cross-type matches; Strict scores credit same-type matches only. Bold marks the attribute–entity pairs, the unit the paper counts. Reference graphs and matching judgments come from the judge, a language model.

Items	Retained item counts extraction / reference	Semantic			Strict		
		P	R	F_1	P	R	F_1
Entity	752 / 846	95.7	83.7	89.3	93.2	82.6	87.6
Attribute (pair)	**578 / 803**	**85.8**	**71.0**	**77.7**	**79.4**	**56.7**	**66.1**
Relation	461 / 702	79.4	50.9	62.0	73.3	48.9	58.6
All	1,791 / 2,351	88.3	69.5	77.8	83.6	63.7	72.3

From records to errors. Most discrepancies are unmatched items, and most are extraction errors (Table 30). Of 2,616 matching records, 934 are discrepancies from 99 captions, 79.8% of them unmatched. Diagnosis sets aside 140 (15.0%) as REPRESENTATION DIFFERENCE or CAPTION AMBIGUITY, leaving 794 extraction errors (Table 31).

Table 30: **Matching records and their diagnoses.** A matching record is a correspondence between the two graphs or an unmatched item. Shares are within each block. Bold rows feed the next stage: the three bold matching categories make up the 934 discrepancies, from 99 captions, that diagnosis labels, and the 794 extraction errors are broken down in Table 31.

Category	Records	Share (%)
Matching (2,616 records)		
Same-type match	1,509	57.7
Cross-type match	173	6.6
Attachment mismatch	**174**	**6.7**
Predicate mismatch	**15**	**0.6**
Unmatched	**745**	**28.5**
Diagnosis (934 discrepancies)		
Representation difference	123	13.2
Caption ambiguity	17	1.8
Extraction error	**794**	**85.0**

Table 31: **Extraction errors by Error Type and Cause.** Each cell counts extraction-error records. Percentages are of the 794 records and are rounded independently. Undet. denotes an undetermined Cause. Bold: direct omissions, the largest cell, with its row and column totals.

Error Type	Direct	Factorization	Merge	Undet.	Total	Share (%)
Omission	**458**	33	68	2	**561**	70.7
Semantic mismatch	81	47	48	1	177	22.3
Unsupported addition	8	46	0	1	55	6.9
Undetermined	0	0	0	1	1	0.1
Total	**547**	126	116	5	794	100.0
Share (%)	68.9	15.9	14.6	0.6	100.0	

Error anatomy. Omissions dominate, and most have no structural cause. Counts are diagnosis records, not independent defects: one structural problem can produce several. OMISSION covers 561 of the 794 errors (70.7%); direct omissions alone, 458 (57.7%), occur in 88 of 100 captions. Half concern Relations (229), the rest Attributes (150), such as the sky's stated blue (Table 28), and Entities (79). Omissions are why recall trails precision. Direct semantic mismatches are rarer (81, 10.2%): the extraction graph gives a bride's purple dress to the man beside her, an error already in the T5 triples.

Merges, the first structural cause, fuse two objects into one Entity, mixing their properties and deleting relationships between them. MERGE explains 116 records (14.6%) in 25 captions: 68 omissions and 48 semantic mismatches. The parser uses one term for both, which graph construction maps to one Entity: the two shirts of Table 28 become one blue and white shirt. A Relation between the merged objects becomes a self-relation and is dropped.

Wrongly split phrases, the second structural cause, produce most invented items. FACTORIZATION explains 126 records (15.9%) in 40 captions: 33 omissions, 47 semantic mismatches, and 46 of the 55 unsupported additions (83.6%). *Red panda* becomes a panda that is red (Table 28). Together, MERGE and FACTORIZATION account for 30.5% of errors.

Caption type. Longer captions yield lower agreement and more errors per word, not just more errors (Table 32). Semantic F_1 falls from 84.9 (short) to 79.5 (medium) and 68.5 (long), where recall is 58.8. Errors per caption rise from 2.91 to 8.04 to 33.2, and errors per 100 words from 16.6 to 20.1 to 25.7. Error composition barely changes: omissions are 70.2–72.5% of errors in all three.

Table 32: **Scores and error rates by caption type.** The 100 captions follow GPIC's caption-type proportions; scores are Semantic, in percent. Errors per 100 words is 100 times each caption's extraction-error count over its whitespace-separated word count, as mean ± sample SD over captions, zero-error captions included. It excludes the one record with an undetermined Error Type and is not reported for the single tag caption. Bold marks the caption type with the lowest agreement.

Caption type	n	Retained items extraction / reference	Precision	Recall	F_1	Errors per 100 words (mean ± SD)
Tag	1	7 / 6	85.7	100.0	92.3	–
Short	45	428 / 535	92.8	78.3	84.9	16.6 ± 14.0
Medium	45	889 / 1,147	89.5	71.5	79.5	20.1 ± 8.8
Long	**9**	**467 / 663**	**82.0**	**58.8**	**68.5**	$\mathbf{25.7 \pm 5.9}$

What the errors mean for the ledger and the panel. The annotation errs mainly by leaving things out, so its counts more often miss presentations than invent them. Precision exceeds recall for every item type (Table 29), and OMISSION accounts for 561 of the 794 errors: a pair the parser misses is a presentation the ledger does not count, whereas an item it invents (55 records) or attaches or labels wrongly (177 records) is rarer but can credit a pair the caption never states. Semantic recall also credits a reference pair that the extraction expresses as a Relation, such as stand @ boy against boy $\xrightarrow{\text{stand on}}$ balcony (Table 27); the ledger reads no Relations, so for its counts the pair recall of 71.0 falls to the Strict 56.7.

A merge costs the counts only when the merged objects differ in name. When they share one, as the two shirts do, each attribute stays on that name and the pairs the sample presents are unchanged, because the exposure indicator of Section 4.2 reads presence in a sample, not object identity; when the parser reuses one object's name for another, as the yellow blanket of the first caption in Table 28 is absorbed into the gray surface, that entity is lost and its attribute is credited to the wrong pair. The recall gap falls hardest on Relations (50.9), which neither the ledger nor the panel uses, and on long captions (58.8). What the audit does not measure is how the uncovered reference pairs distribute over concepts, so the error in any one concept's exposure count remains unquantified; item-level agreement on 100 captions is the evidence this paper offers.

F.5 EVALUATOR VALIDITY

We chose the judge because it passed both representation checks in every output and had the highest overall repeatability of fourteen candidate conditions; its fixed prompts gave repeatable matching scores on development captions, but it judges the extraction graphs against its own references.

Candidate comparison. Local candidates were Qwen3.5-9B[7] and Gemma-4-12B-it[8] with Q8_0 quantization, GPT-OSS-20B[9] with native MXFP4 weights, and Qwen3.8-Flash-Next-FP8[10] with FP8 weights on two NVIDIA H200 GPUs.

Candidates from four local models and closed API models each generated seven or eight valid reference graphs for one long caption from the same prompt and field definitions (Table 33). For *the sunlight casts a shadow of his body onto the ground*, the checks require a separate body Entity and the Relation shadow $\xrightarrow{\text{of}}$ body; failures often repeated the subject in the predicate (e.g., *shadow_of*) or lost the body Entity. Repeatability averages shared items divided by shared plus unshared items across the 28 or 21 output pairs, comparing lowercased, space-normalized names plus each Attribute's Entity and category and each Relation's subject, predicate, and object.

Inference settings. API runs used strict JSON schemas and the reasoning efforts reported in Table 33. Sampling parameters (temperature and top-p) were left unset, using provider defaults. Output-token limits were 128,000 for Astra, Terra, Luna, and completed GPT-5.4 mini xhigh runs, and 32,768 for the other API conditions. Qwen3.8 used vLLM 0.29.0; the other local models used llama.cpp.[11] Local runs used default or recommended thinking-mode sampling settings: temperature was 1.0; (top-p, top-k) was $(0.95, 20)$ for both Qwen models, $(0.95, 64)$ for Gemma, and $(1.0, 0)$ for GPT-OSS, where zero turns off top-k filtering. Qwen3.5 used a presence penalty of 1.5; the other local models used zero. Repetitions used distinct fixed seeds. Local output limits were 81,920 tokens for Qwen3.5 and GPT-OSS, 261,466 for Gemma, and 261,459 for Qwen3.8. Local generation used JSON grammar constraints with schema field definitions supplied in the prompt.

Judge selection. Only `gpt-6-astra` at xhigh and max effort passed both checks throughout; max, matching more repeatably, became the judge. Over 12 calls per effort and graph pair, the most frequent accepted match set occurred in 58.3% of xhigh versus 83.3% of max calls for a short caption and 91.7% versus 100.0% for a long one, though max still varied in some long-caption category assignments.

Fixed prompts and repeatability. Prompts fixed before the audit, after testing 23 reference-generation, 54 matching, and 13 diagnosis variants on development examples, gave repeatable matching scores and mostly stable diagnosis labels. The instructions give general semantic criteria, leaving judgments to caption and graph evidence; competing prompts saw identical inputs, and matching was also tested with the graph order swapped. Reference generation (six captions, ten calls each) always kept shadow $\xrightarrow{\text{of}}$ body for the long example, though graphs varied in wording, grouping, coverage, and attachment. Matching (12 graph pairs, three calls each) reproduced overall precision, recall, and F_1 before exclusions for all 12, accepted match sets for 11, and full reports (categories and item references) for 10. Diagnosis (92 records in six captions, three calls each) left all three labels unchanged for 85 records (92.4%); all seven others were in the two long captions.

Scope. Self-judging may favor the judge's own representations. The audit made one call per caption and stage and includes some prompt-development captions. Tag and long captions are few, following GPIC's proportions. Neither human judgments nor checks against the images were made.

[7] `https://huggingface.co/Qwen/Qwen3.5-9B`
[8] `https://huggingface.co/ggml-org/gemma-4-12B-it-GGUF`
[9] `https://huggingface.co/openai/gpt-oss-20b`
[10] `https://huggingface.co/Qwen/Qwen3.8-Flash-Next-FP8`
[11] `https://github.com/ggml-org/llama.cpp`

Table 33: **Reference-generation checks and repeatability (%).** The first two columns give the share of outputs passing each representation check; the other four give mean pairwise lexical item agreement between repeated outputs. Bold marks check rates of 100.0 and the best value in each repeatability column; underline marks the second best among the rows shown. Seven of the fourteen conditions are shown; the note summarizes the rest.

	Representation checks		Repeatability			
Model (reasoning)	shadow $\xrightarrow{\text{of}}$ body	Separate body	Entity	Attribute	Relation	All
Local model						
Qwen3.8-Flash-Next-FP8 (xhigh)	25.0	87.5	83.4	22.8	28.7	40.4
Closed API models						
gpt-5.6-luna (max)	12.5	**100.0**	88.4	66.2	32.7	56.6
gpt-5.6-terra (max)	37.5	87.5	76.8	58.7	36.7	54.1
gpt-5.6-sol (max)	0.0	**100.0**	84.6	69.1	27.7	54.5
gpt-6-astra (high)	0.0	**100.0**	95.5	83.8	<u>52.1</u>	73.0
gpt-6-astra (xhigh)	**100.0**	**100.0**	<u>96.6</u>	**97.8**	48.6	<u>75.0</u>
gpt-6-astra (max)	**100.0**	**100.0**	**99.0**	<u>96.7</u>	**57.7**	**79.4**

All conditions use the same long caption and development prompt. The Relation check requires the literal predicate *of* from shadow to body. Each condition shown has eight valid outputs; only successful outputs enter the calculations. API models other than gpt-6-astra appear only at their highest tested reasoning effort, given in parentheses; API model identifiers are as returned by the API. The seven conditions not shown—Qwen3.5-9B, Gemma-4-12B-it, GPT-OSS-20B (high), gpt-5-nano-2025-08-07 (high), gpt-5-mini-2025-08-07 (high), gpt-5.4-nano-2026-03-17 (xhigh), and gpt-5.4-mini-2026-03-17 (xhigh)—passed neither representation check in any output and reached at most 46.5% overall repeatability.