POLYGEN: FULLY SYNTHETIC VISION-LANGUAGE TRAINING VIA MULTI-GENERATOR ENSEMBLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthetic data offers a scalable solution for vision-language pre-training, yet current state-of-the-art methods typically rely on scaling up a single generative backbone, which introduces generator-specific spectral biases and limits feature diversity. We introduce **PolyGen**, a framework grounded in the *Generator-Invariance Hypothesis*: when images are rendered by an ensemble of architecturally distinct generators, generator-specific artifacts vary across sources while semantic content is shared by construction, and a model trained on the union is encouraged to encode the latter. We treat this hypothesis as a training principle with two testable implications, ensemble composition and invariance regularization, and pair it with the timing of hard-negative introduction, each examined through dedicated ablations. PolyGen operationalizes it through a multi-positive strategy exposing the model to disjoint visual manifolds rendered from the same caption, together with a Programmatic Hard Negative curriculum enforcing fine-grained compositional understanding. By reallocating the same image budget from unique captions to multi-source variations, PolyGen yields a more robust feature space, outperforming the leading single-source baseline (SynthCLIP) by $+19.0\%$ on aggregate multi-task benchmarks and $+11.3\%$ on compositional reasoning, and, under an identical protocol, exceeding a real-data CC3M-subset baseline on cross-modal retrieval while closing two thirds of the zero-shot gap.

1 INTRODUCTION

One of the major breakthroughs in artificial intelligence has been the emergence of Vision-Language Models (VLMs) capable of aligning visual and textual representations in a shared embedding space. Models such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) have unlocked remarkable zero-shot capabilities, effectively decoupling visual recognition from the constraints of fixed label sets. However, the efficacy of these models rests upon a foundation that is becoming increasingly strained: massive, web-scale datasets comprising billions of image-text pairs.

This reliance on web-scraped data presents two challenges: the supply of high-quality human-generated data may become a bottleneck for future scaling (Villalobos et al., 2022), and real-world datasets are noisy, rife with misalignment, societal biases, and privacy or copyright concerns requiring expensive curation. A principled alternative is fully synthetic data (Hammoud et al., 2024; Tian et al., 2024): given the fidelity of modern text-to-image (T2I) diffusion models, scalable, safe and controllable training corpora can be generated on demand (Fan et al., 2024).

Despite this promise, a persistent "Synthetic Gap" remains: models trained exclusively on synthetic data underperform their real-data counterparts in generalization tasks. Fan et al. (2024) identified a fundamental cause, a trade-off between *diversity* and *recognizability* inherent to single-generator training, with older diffusion models offering broader semantic variance at the cost of perceptual fidelity and modern distilled models the inverse. Recent evidence confirms the pattern: across thirteen T2I generators released between 2022 and 2025, classifiers trained purely on synthetic data transfer progressively worse to real test sets despite gains in perceptual quality and prompt adherence, with the deficit traced to texture impoverishment, high-frequency mismatch, and density-coverage collapse (Adamkiewicz et al., 2026). Even ensembles of identical models can outperform single-model

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.

approaches, suggesting individual generators are confined to local optima in the manifold space (Lampis et al., 2023). Trained on a single generator, VLMs learn to exploit low-level, generator-specific cues such as texture patterns and spectral fingerprints (Corvi et al., 2023), rather than the abstract semantic features open-world generalization requires.

To address these challenges, we introduce **PolyGen** (Polylithic Generation Framework), an end-to-end pipeline for fully synthetic vision-language pre-training. Our approach is grounded in the *Generator-Invariance Hypothesis*: when training on images rendered by an ensemble of architecturally distinct generators, generator-specific spectral artifacts and stylistic biases vary across sources while the underlying semantic content is shared by construction; a model trained on the union is therefore biased toward the latter. We treat this hypothesis as a training principle with two testable implications, and pair it with a third question specific to the hard-negative curriculum, each examined in a dedicated ablation: **(1)** at equal image budget, heterogeneous generator ensembles should outperform single-source scaling, isolating the contribution of architectural diversity from sample volume; **(2)** enforcing image-to-image invariance should yield a directional trade-off, improving semantic transfer while reducing reliance on texture-dependent shortcuts; **(3)** the timing of hard-negative introduction should affect both general and compositional performance, with a gradual curriculum recovering both.

PolyGen additionally addresses a well-known limitation of contrastive learning: standard training objectives often reduce to "bag-of-words" matching, where models identify objects through simple co-occurrence statistics rather than fine-grained compositional understanding (Yuksekgonul et al., 2022). We integrate a Programmatic Hard Negative curriculum that enforces the syntactic precision required for compositional generalization, and Section 4.3 isolates the effect of its timing.

The framework operationalizes the principle through three complementary mechanisms (Figure 1): a **Multi-Positive Ensemble** that renders each caption through a heterogeneous pool of generators (SD 1.5, SD 2, SDXL-Turbo, SANA), pairing architecturally distinct outputs as positives via multi-positive contrastive objectives; **Programmatic Hard Negatives** obtained by axis-conditioned LLM perturbations of the base captions, forcing the model beyond co-occurrence shortcuts; and a **Curriculum Scheduler** that gradually increases the proportion of hard negatives during training, allowing the model to consolidate coarse semantic structure before refining discriminative boundaries.

Empirically, PolyGen demonstrates that engineering the training distribution, rather than scaling it, yields marked improvements: $+\mathbf{19.0}\%$ in Δ_{MTL} over strong single-source baselines and $+\mathbf{11.3}\%$ on the SugarCrepe/SugarCrepe++ compositionality benchmarks. Under an identical training protocol, PolyGen also exceeds a real-data (CC3M 500k) baseline on both retrieval directions while closing roughly two thirds of the zero-shot classification gap (Section 4.2).

2 Related Works

Hybrid Synthetic Strategies. A dominant line of research mitigates data scarcity by leveraging real-world captions to generate synthetic images. StableRep (Tian et al., 2023) showed that multi-positive contrastive learning, treating multiple synthetic images generated from the same real caption as positive pairs, can outperform models trained on real data; crucially, its positives are multiple seeds of a single generator, so they share the same architectural prior. LaCLIP (Fan et al., 2023) and VeCLIP (Lai et al., 2024) instead apply LLMs to rewrite captions of real images, improving alignment and diversity while maintaining image-text grounding. These approaches remain constrained by web-scraped data, inheriting its biases, privacy risks and finite scale.

Fully Synthetic Pipelines. Recent works pursue fully synthetic pipelines, generating both captions and images without human-collected data (Tian et al., 2024). Fan et al. (2024) proposed scaling laws for synthetic data, showing that synthetic pre-training can rival real-world baselines, and SynthCLIP (Hammoud et al., 2024) introduced an end-to-end pipeline using LLM-generated captions and Stable Diffusion images, training CLIP from scratch without human data. Synthetic data has also been shown to reach parity with real datasets in image classification while providing privacy guarantees (Resmini et al., 2025), though only for single-task supervised learning. These approaches all suffer from *architectural homogeneity*: relying on a single generative model causes the downstream model to overfit to generator-specific artifacts (Corvi et al., 2023). We posit that this is a primary contributor to the persistent Synthetic Gap in vision-language tasks, as models learn to exploit low-level texture patterns rather than abstract semantic features.

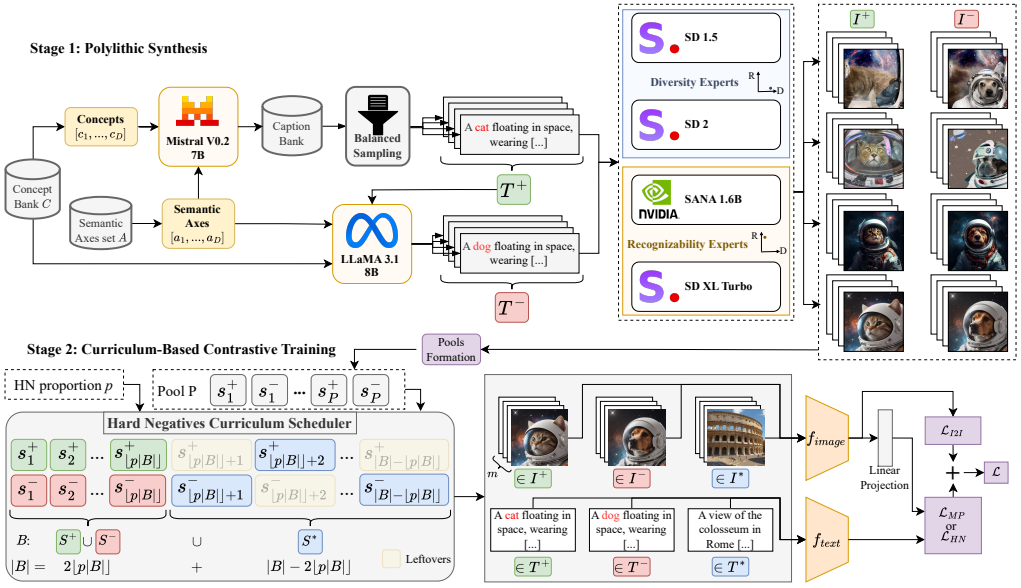


Figure 1: **PolyGen Pipeline Overview.** *Stage 1 (Polylithic Synthesis):* Concepts from the MetaCLIP Bank and semantic axes feed a dual-LLM system: Mistral V0.2 generates base captions T^+ from concept-attribute pairs, Llama 3.1 produces hard negatives T^- by modifying specific semantic dimensions. Caption pairs are rendered through four diffusion models: Diversity Experts (SD 1.5, SD 2) ensure broad manifold coverage despite lower fidelity, Recognizability Experts (SANA, SDXL-Turbo) provide high prompt adherence despite limited variance. *Stage 2 (Curriculum-Based Training):* A scheduler increases the hard negative proportion p during training; unused samples populate a leftover queue. Batches combine multi-positive samples (I^+, T^+) and hard negatives (I^-, T^-).

Hard Negative Mining. Contrastive vision-language models tend to rely on lexical co-occurrence rather than compositional reasoning (Yuksekgonul et al., 2022), a failure mode systematically evaluated by ARO, SugarCrepe (Hsieh et al., 2023) and SugarCrepe++ (Dumpala et al., 2024) through attribute binding, spatial reasoning and object-relation tasks. Hard Negative Mining has evolved from importance weighting (Robinson et al., 2020; Radenovic et al., 2023) to hard negative generation via rule-based swapping (Yuksekgonul et al., 2022) and LLM-driven rewriting. TripletCLIP (Patel et al., 2024), the closest related work, trains on real CC3M pairs augmented with LLM-generated synthetic negatives, a hybrid approach that depends on web-scraped anchors and does not address generator-specific biases in the visual encoder; intrinsic data mining (Wang et al., 2025) remains computationally expensive and bound by the static data distribution. We propose a *constructive* alternative: synthesizing structurally grounded base captions first, we programmatically constrain the LLM to perturb specific semantic axes, yielding controlled counterfactuals. To isolate our framework from the loss alone, we include a fully-synthetic TripletCLIP-style baseline (Section 4.3).

3 Methodology

We introduce **PolyGen** (Polylithic Generation Framework), an end-to-end pipeline for training vision-language models on fully synthetic data. The framework consists of two stages (Figure 1): *Polylithic Synthesis* combines structured caption generation with multi-generator image rendering, followed by *Curriculum-Based Contrastive Training*. The design directly addresses the two implications and the curriculum question stated in Section 1: ensemble composition (Section 3.2), invariance regularization, and the timing of hard-negative introduction (Section 3.3).

3.1 Structured Caption Pairs

The first component of *Polylithic Synthesis* (Stage 1, Figure 1) generates semantically controlled caption pairs. Prior work on synthetic data generation typically relies on unconstrained prompting, producing descriptive captions without semantic control. We instead adopt a constructive approach, generating training pairs (T^+, T^-) through controlled semantic modifications.

Attribute-Conditioned Caption Generation. We use Mistral-V0.2-7B (Jiang et al., 2023) to generate the base captions T^+, conditioning generation on a control tuple $(c \in \mathcal{C},\ a \in \mathcal{A})$. Here, c is the base *Concept*, i.e., the primary subject sampled from the MetaCLIP Concept Bank (Xu et al., 2023), and a is the *semantic modification axis* along which the hard negative will be generated. $\mathcal{A}$ includes visual attributes (*Lighting, Material, Perspective, Color, Background, Position, Style*, ablated in Appendix D) and an explicit *Concept* axis for subject-substitution negatives. The attribute conditioning additionally increases variability when repeated concepts arise during large-scale generation (full prompts in Appendix A). T^+ is obtained by downsampling a broader caption pool using MetaCLIP's balanced sampling algorithm (Appendix G).

Hard Negative Generation. We use Llama-3.1-8B (Dubey et al., 2024), selected based on an axis-recovery ablation across five 7B/8B-parameter LLMs (Appendix C.1), to generate the counterfactual captions T^-. Given $t \in T^+$, the corresponding axis a, and concept c, the model is instructed to modify only the specified axis while preserving syntactic structure. Compared to rigid programmatic substitution, which can produce logical absurdities by modifying terms in isolation, LLM-based generation dynamically adjusts contextually dependent attributes to align with the target modification. Qualitative examples across generators and axes are provided in Appendix K.

3.2 Multi-Generator Synthesis Ensemble

The second component of Stage 1 renders each caption pair through multiple diffusion models. Relying on a single generator risks learning a biased representation of the data manifold: each generator captures only a subset of the visual-semantic spectrum, introducing architecture-specific frequency patterns and compression artifacts that the downstream model may exploit as spurious features rather than learning robust semantic content. All image generation is performed offline prior to CLIP training, incurring no additional cost at training or inference time; per-generator inference costs are reported in Appendix C.4.

Metric-Driven Selection. To systematically evaluate candidate generators, we establish a quantitative framework measuring two complementary properties of synthetic datasets in open-vocabulary contrastive settings. While Fan et al. (2024) defined *Recognizability* and *Diversity* for supervised classification datasets with fixed label sets, we adapt these for the self-supervised CLIP paradigm, computing both with a CLIP model pre-trained on real data:

Recognizability (Fidelity) is measured, in the absence of class labels, as the expected CLIPScore (Hessel et al., 2021) over dataset $\mathcal{D}$, $R(\mathcal{D}) = \mathbb{E}_{(x,t)\sim\mathcal{D}}[\mathrm{CLIPScore}(f_{\mathrm{img}}(x), f_{\mathrm{txt}}(t))]$; it indicates how reliably the generator respects the semantic constraints of the prompt.

Diversity is a cluster-based metric capturing intra-concept variance: the dataset is partitioned into $K{=}1000$ semantic clusters by caption embedding, and Diversity is the expected within-cluster dispersion of image embeddings, $D(\mathcal{D}) = \mathbb{E}_k[\sigma_k]$, where $\sigma_k = \sqrt{\mathbb{E}_{x\in\mathcal{D}_k}\|f_{\mathrm{img}}(x) - \mu_k\|^2}$ is the standard deviation of the embeddings in cluster k about their centroid μ_k. This metric captures the generator's ability to produce visually distinct interpretations of the same semantic concept or group of concepts. The pair (R, D) defines the axes of Figure 2 and is the admission test by which a generator is assigned its role below; roles follow from measured position, not model identity or release date (Appendix Q).

The Generator-Invariance Strategy. We define the *Generator-Invariance Hypothesis* as the assumption that, across architecturally distinct generators rendering the same caption, semantic content is shared by construction while spectral and stylistic artifacts vary across sources; a model trained on the union of such renderings is therefore biased toward the former. Guided by this, we construct an ensemble $\mathcal{G}$ maximizing coverage of the visual-semantic spectrum, with four architecturally distinct models in complementary regions of the Recognizability/Diversity space:

Recognizability Experts: SDXL-Turbo (Sauer et al., 2024) and SANA-1.6B (Xie et al., 2024), selected for high prompt adherence and photorealism despite lower output diversity.

Diversity Experts: Stable Diffusion v1.5 and v2 (Rombach et al., 2022), selected for high intra-concept variance despite lower perceptual fidelity.

For every caption pair (t^+, t^-), we generate corresponding images using all four models. The semantic content of the caption is consistent across the four sources by construction, whereas generator-specific spectral signatures (e.g., high-frequency attenuation in SDXL-Turbo; Appendix E)

vary across sources, so the union reinforces no single signature. The hypothesis requires this variation in aggregate, not that any two renderings of a caption be visually disjoint: two generators may depict a caption similarly, and Appendices E and L establish the aggregate claim.

3.3 Curriculum-Based Contrastive Training

Stage 2 leverages the structured dataset through a specialized training pipeline that exploits the one-to-many mapping from captions to images (up to $n^+ = 4$ positive images per caption).

Multi-Positive Objectives. To exploit the ensemble of n^+ images generated for each caption t_i, we adopt a Multi-Positive objective (Tian et al., 2023). Rather than one-hot targets, we use soft targets distributing probability mass uniformly ($1/n^+$) across all images from the same prompt. This objective $\mathcal{L}_{MP}$ aligns the text embedding with the semantic centroid of the diverse visual manifold formed by the generator ensemble.

Image-to-Image Invariance. Beyond text-to-image alignment, we evaluate the impact of explicit image-to-image regularization through an additional contrastive term $\mathcal{L}_{I2I}$ (Tian et al., 2023), which penalizes generator-specific idiosyncrasies by aligning the embeddings of images rendered from the same caption. The loss is computed as the cross-entropy:

$$\mathcal{L}_{I2I} = -\frac{1}{N}\sum_{i\in I}\left(\sum_{j\in I\setminus\{i\}} p_{i,j}\log q_{i,j}\right), \tag{1}$$

where $p_{i,j}$ is the uniform ground-truth distribution over images generated from the same caption and $q_{i,j}$ is the predicted softmax (full derivation in Appendix B.1). The term acts as a structural regularizer, directing the visual encoder away from spectral artifacts and stylistic biases toward the invariant semantic core defined by the prompt. Importantly, $\mathcal{L}_{I2I}$ is an *optional* component: as analyzed in Section 4.2, multi-generator training without $\mathcal{L}_{I2I}$ already yields consistent gains on vision-language tasks while preserving performance on texture-sensitive evaluations. The two configurations offer a controllable trade-off between semantic abstraction and low-level feature retention.

Hard Negatives Objectives. Standard InfoNCE suppresses hard negatives T^- by treating them as any other sample in the batch denominator. To enforce fine-grained discrimination, we employ the TripletCLIP loss (Patel et al., 2024), adapted here to the fully synthetic regime where both anchor images and hard negatives are produced by our multi-generator pipeline. We define the Hard-Negative NCE loss for image-to-text as:

$$\mathcal{L}_{NCE}^{I^+\to T^+;\,T^-} = -\frac{1}{N}\sum_{k=1}^{N}\log\frac{e^{s(i_k,\,t_k)/\tau}}{\sum_{t\in T^+\cup T^-} e^{s(i_k,\,t)/\tau}}, \tag{2}$$

analogously for text-to-image. The NegCLIP loss (Yuksekgonul et al., 2022) is then $\mathcal{L}_{NegCLIP}^{I^+,T^+,T^-} = \mathcal{L}_{NCE}^{I^+\to T^+;\,T^-} + \mathcal{L}_{NCE}^{T^+\to I^+}$, and the total TripletCLIP loss is $\mathcal{L}_{TripletCLIP} = \mathcal{L}_{NegCLIP}^{I^+,T^+,T^-} + \mathcal{L}_{NegCLIP}^{I^-,T^-,T^+}$, forcing the model to distinguish the base concept from its semantic counterfactual.

Hard Negatives Curriculum Scheduler. Naive injection of hard negatives causes early training instability: fine-grained verification tasks interfere with coarse semantic alignment before stable conceptual clusters can form. We introduce a curriculum scheduler that linearly increases the in-batch hard negative ratio p over the training epochs. Here $p = |T^-|/|B| \in [0, 0.5]$ is the fraction of the batch occupied by hard negatives; since each enters jointly with its positive counterpart, the fraction of samples that are paired is $2p$, and a fully paired batch corresponds to $p = 0.5$, the maximum of the range by construction. The endpoint and the ramp speed are therefore the two free parameters of the schedule, and we characterize the full feasible family in Section 4.3. To ensure full data utilization throughout each epoch, we implement a *Leftover Queue*: samples not yet paired with their corresponding hard negatives are queued and batched in subsequent steps. Used in combination with a decaying learning rate, this schedule allows the model to consolidate a robust semantic baseline before hard negatives are introduced to refine discriminative boundaries. Since batch composition varies, we adapt $\mathcal{L}_{HN}$ to balance subset contributions:

$$\mathcal{L}_{HN} = \frac{1}{|B|}\left(|T|\,\mathcal{L}_{NegCLIP}^{I,\,T,\,T^-} + |T^-|\,\mathcal{L}_{NegCLIP}^{I^-,\,T^-,\,T}\right), \tag{3}$$

with $T = T^* \cup T^+$, $I = I^* \cup I^+$, and $|B| = |T| + |T^-|$, where T^* and I^* are the unpaired captions and images drawn from the Leftover Queue to fill the batch when $p < 0.5$.

n^+	m	$\mathcal{L}_{I2I}$	*L.P.*	*Few shot*	*Img Ret.*	*Text Ret.*	*0 shot*	Δ_{MTL} vs. SynthCLIP	Δ_{MTL} vs. same n^+, $m=1$
1	–	✗	49.4 +0.3	62.5 +0.9	11.5 −0.1	14.7 −1.2	13.23 +5.08	–	–
1	1	✗	49.1	61.6	11.6	15.9	8.15	-	-
1	2	✗	47.5 −1.6	61.1 −0.5	12.1 +0.5	17.5 +1.6	8.53 +0.38	+3.1%	+3.1%
1	3	✗	48.5 −0.6	61.3 −0.3	13.3 +1.7	17.9 +2.0	8.56 +0.41	+6.1%	+6.1%
1	4	✗	48.9 −0.2	61.5 −0.1	12.8 +1.2	17.6 +1.7	8.92 +0.77	+6.2%	+6.2%
2	1	✗	49.5 +0.4	62.4 +0.8	13.8 +2.2	18.7 +2.8	9.41 +1.26	+11.0%	-
2	2	✗	47.9 −1.2	61.1 −0.5	14.2 +2.6	20.2 +4.3	9.54 +1.39	+12.7%	+1.3%
3	1	✗	49.3 +0.2	62.7 +1.1	14.7 +3.1	21.1 +5.2	9.66 +1.51	+16.1%	-
3	3	✗	48.2 −0.9	61.1 −0.5	16.2 +4.6	**22.4** +6.5	9.38 +1.23	+18.7%	+1.7%
4	1	✗	**49.8** +0.7	**62.8** +1.2	14.8 +3.2	19.4 +3.5	9.56 +1.41	+14.1%	-
4	4	✗	48.9 −0.2	62.0 +0.4	16.1 +4.5	21.4 +5.5	9.86 +1.71	**+19.0%**	+3.8%
2	1	✓	43.4 −5.7	57.3 −4.3	13.3 +1.7	17.6 +1.7	10.14 +1.99	+6.3%	-
2	2	✓	41.5 −7.6	55.7 −5.9	14.4 +2.8	19.8 +3.9	11.19 +3.04	+12.3%	+4.8%
3	1	✓	44.4 −4.7	57.3 −4.3	14.2 +2.6	18.9 +3.0	10.88 +2.73	+11.8%	-
3	3	✓	41.3 −7.8	55.2 −6.4	**16.5** +4.9	21.7 +5.8	10.98 +2.83	+17.4%	+4.1%
4	1	✓	45.2 −3.9	57.7 −3.9	15.0 +3.4	19.9 +4.0	11.24 +3.09	+15.6%	-
4	4	✓	42.9 −6.2	56.4 −5.2	15.7 +4.1	20.6 +4.7	**11.34** +3.19	+16.6%	+0.4%

Table 1: Heterogeneity (m) vs. Density (n^+): jointly scaling both maximizes Δ_{MTL}. The greyed first row is a reference CLIP model trained on a **real CC3M 500k subset**. Small figures are absolute deltas vs. the SynthCLIP baseline (n^+=1, m=1). Retrieval is R@5; zero-shot is ImageNet top-5.

4 Experiments and Results

Training Configuration. Unless stated otherwise, all experiments use a fixed budget of 500k images; the number of captions follows from the setting, $500k/n^+$ for n^+ positives. We train ViT-B/16 CLIP from scratch for 40 epochs with AdamW, learning rate 5×10^{-4}, cosine decay, 1 warmup epoch and batch size 1536. The hard negative ratio p follows the curriculum scheduler, increasing linearly from 0 to 0.5 over training.

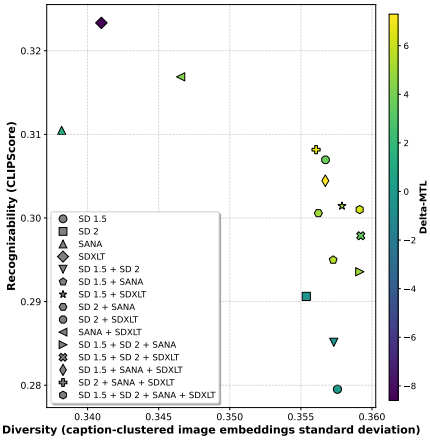


Figure 2: Recognizability vs. Diversity for all 15 subsets of the four-generator pool, colored by Δ_{MTL}. Heterogeneous ensembles shift toward the upper-right quadrant.

Downstream Tasks. We evaluate on vision-only tasks (linear probing and few-shot classification, assessing image-encoder feature quality) and vision-language tasks (zero-shot classification and retrieval), aggregating with the Δ_{MTL} metric (Vandenhende et al., 2021) (Appendix H): the average relative gain or drop against a baseline across all tasks. For Linear Probing and Few-Shot we use CIFAR-10/100 (Krizhevsky et al., 2009), FGVC Aircraft (Maji et al., 2013), DTD (Cimpoi et al., 2014), Oxford Flowers (Nilsback & Zisserman, 2008), Oxford-IIIT Pets (Parkhi et al., 2012), SUN397 (Xiao et al., 2010), Caltech-101 (Fei-Fei et al., 2006), and Food-101 (Bossard et al., 2014). For zero-shot classification we use ImageNet (Deng et al., 2009); for retrieval, Flickr8k, Flickr30k (Young et al., 2014), and MS COCO (Lin et al., 2014). Compositional reasoning uses SugarCrepe (Hsieh et al., 2023) and SugarCrepe++ (Dumpala et al., 2024); per-dataset results are in Appendix I.

4.1 Recognizability vs. Diversity Implications

The first testable implication is that, at equal image budget, heterogeneous ensembles should outperform single-source scaling. We test this by training CLIP on fixed sets of 500k pairs ($n^+ = 1$), each caption rendered by a sampled subset of generators, over all 15 subsets of the four-generator pool, so that any gain reflects inter-source diversity rather than sample count (Figure 2). The measured (R, D) positions show a clear dichotomy: SD v1.5 occupies the high-diversity, lower-recognizability region, while SDXL-Turbo and SANA-1.6B trade intra-concept Diversity for Recognizability, the pattern expected of mode collapse onto a limited set of high-probability interpretations. PolyGen

ensembles shift the training distribution toward the upper-right quadrant, generalizing better than either generator type alone.

4.2 The Heterogeneity-Density Trade-off

The first implication also asks whether heterogeneity (m) contributes independently of positive density (n^+): Table 1 crosses $m = 1$ (SD 1.5 only) to $m = 4$ with n^+ from 1 to 4.

Separating the Two Axes. Because the pivotal comparisons are between nearby configurations, we re-trained the four corner cells of the $m \times n^+$ square with three seeds each; the full seeded comparison is reported in Appendix M. At fixed $n^+ = 4$, heterogeneity is worth $+1.70$ image R@5 (roughly $7\times$ the largest per-cell standard deviation) and $+0.42$ zero-shot, with text retrieval consistent in sign but too noisy at three seeds to be independently decisive. Seeding both levels of n^+ also shows the cost side to be n^+-dependent: at $n^+ = 4$ heterogeneity trades -1.12 linear-probe and -0.81 few-shot points, whereas at $n^+ = 1$ it costs nothing measurable (-0.02 and $+0.08$, inside seed noise). The cost of heterogeneity is therefore visible only at high density rather than being an intrinsic property of m: the two axes compete for the same capacity.

Comparison with Real Data. The reference row of Table 1 is a baseline trained on the 500k real CC3M (Sharma et al., 2018) pairs under the identical protocol. On retrieval both PolyGen operating points exceed it by a wide margin ($+4.60/+6.70$ R@5 without $\mathcal{L}_{I2I}$); single-source synthetic data already matches real CC3M on image retrieval, but PolyGen extends the margin substantially. On zero-shot classification real data retains the advantage: of the 5.08-point SynthCLIP–real gap the ensemble closes 33.6% without $\mathcal{L}_{I2I}$ and 62.7% with it, along a monotone ladder across the framework's axes. On vision-only probes the ensemble without $\mathcal{L}_{I2I}$ is at parity ($-0.50/-0.51$, within seed spread). We read this as task-dependent parity, not evidence that synthetic data is preferable to real: CC3M captions are web alt-text whereas ours are dense and descriptive, hence closer to the Flickr/COCO distribution these benchmarks are built on, so caption style is a named confounder: LLM rewriting of alt-text alone is known to produce large retrieval gains (Fan et al., 2023; Lai et al., 2024), and our single-generator baseline already matches real CC3M on image retrieval (11.6 vs. 11.5) and exceeds it on text retrieval (15.9 vs. 14.7), before any heterogeneity.

Invariance as a Regularizer. The second implication concerns the directional trade-off induced by $\mathcal{L}_{I2I}$, which is an optional operating point rather than a component of the framework: every heterogeneity gain above holds without it. A consistent pattern emerges across Table 1: adding $\mathcal{L}_{I2I}$ maximizes zero-shot accuracy while inducing a regression in linear probing, and the trade-off is attributable to the loss specifically, not to multi-generator training alone. The per-dataset breakdown identifies the mechanism. $\mathcal{L}_{I2I}$ trains the encoder on which images in a batch derive from the same prompt, favouring cross-rendering semantic constants and suppressing intra-class fine detail, so fine-grained and texture-sensitive probes should suffer most and coarse-grained evaluation least. That is the observed pattern: in the $(2, 2)$ block the loss costs most on Oxford Flowers (-13.7), Caltech (-10.8) and DTD (-8.6, a texture dataset), least on CIFAR-10/100 ($-3.2/-1.4$), while ImageNet zero-shot improves ($+1.65$). We measure the mechanism directly with a 4-way generator-identification probe fit on frozen features of held-out captions rendered through all four generators (chance 25%; protocol and full table in Appendix L). Decodability does not fall with heterogeneity, it rises, from 81.3% at m=1 to 88.5% at m=4; a LAION-2B CLIP that never saw these generators scores 87.96%, so linear decodability of generator identity tracks general feature quality rather than the composition of the training source. The clean contrast is instead between two encoders trained on identical data and differing only in whether $\mathcal{L}_{I2I}$ is active: enabling it drops the probe to 68.3% and its margin to the random-initialization level, and images from the same generator rendered from *different* captions become essentially orthogonal ($\cos$: $0.124 \rightarrow 0.0004$) while cross-generator agreement for the same caption is preserved ($0.579 \rightarrow 0.599$). Heterogeneity supplies the cross-source variation, and $\mathcal{L}_{I2I}$ is what converts it into invariance.

4.3 Hard Negatives and Curriculum

The third axis concerns the timing of hard-negative introduction: aggressive early injection should disrupt coarse semantic alignment, while a gradual curriculum should recover both general and compositional performance. Table 2 provides a granular analysis of training dynamics within the $m = 2, n^+ = 2$ configuration block, isolating the contribution of each component (hard negative

n^+	m	$\mathcal{L}_{I2I}$	*HN samples*	$\mathcal{L}_{HN}$	*sched.*	*L.P.*	*Few shot*	*Img Ret.*	*Text Ret.*	*0 shot*	Δ_{MTL}	Sugar Crepe	Sugar Crepe++	Δ^{HN}_{MTL}
1	1	✗	✗	✗	✗	**49.1**	**61.6**	11.6	15.9	8.15	-	55.4	36.7	-
1	2†	✗	✓	✓	✗	47.2 −1.9	60.6 −1.0	13.1 +1.5	17.6 +1.7	9.62 +1.47	+7.3%	55.8 +0.4	36.2 −0.5	-0.3%
2	2	✓	✓	✗	✗	41.5 −7.6	55.7 −5.9	13.2 +1.6	17.9 +2.0	9.95 +1.80	+4.7%	56.9 +1.5	40.0 +3.3	+5.9%
2	2	✓	✓	✗	✓	40.8 −8.3	55.8 −5.8	13.6 +2.0	18.8 +2.9	10.23 +2.08	+7.0%	55.0 −0.4	38.8 +2.1	+2.6%
2	2	✓	✓	✓	✗	42.3 −6.8	56.8 −4.8	14.0 +2.4	19.0 +3.1	10.27 +2.12	+9.0%	55.8 +0.4	39.2 +2.5	+3.7%
2	2	✓	✓	✓	✓	41.4 −7.7	55.8 −5.8	14.6 +3.0	**20.5** +4.6	10.8 +2.65	+12.5%	56.4 +1.0	40.5 +3.8	+6.1%
2	2	✓	✓	✓	✓‡	41.8 −7.3	55.9 −5.7	14.1 +2.5	19.5 +3.6	**11.24** +3.09	+11.6%	**59.1** +3.7	**42.5** +5.8	**+11.3%**
3	3	✓	✓	✓	✓	40.8 −8.3	55.5 −6.1	15.7 +4.1	18.9 +3.0	10.72 +2.57	+11.8%	58.1 +2.7	41.6 +4.9	+9.1%
4	4	✓	✓	✓	✓	42.7 −6.4	56.5 −5.1	15.7 +4.1	20.4 +4.5	11.1 +2.95	**+15.5%**	56.8 +1.4	40.7 +4.0	+6.8%

Table 2: Hard negative ablation results. From the third block onward models use $\mathcal{L}_{I2I}$; when active, $\mathcal{L}_{HN}$ replaces $\mathcal{L}_{MP}$. †Controlled synthetic TripletCLIP-style baseline: SD 1.5 base samples, SDXL-Turbo hard negatives, TripletCLIP loss, no multi-positive training, no curriculum. ‡Conservative schedule (p ramped to 0.25; family in Table 13). Small figures are absolute deltas vs. the first row.

samples, $\mathcal{L}_{HN}$, scheduler) and tracking both general transfer (Δ_{MTL}) and fine-grained compositional reasoning (SugarCrepe, SugarCrepe++).

Comparison with a Synthetic TripletCLIP-style Baseline. To isolate our framework from the loss function alone, we include a controlled fully synthetic TripletCLIP-style baseline (row †, configuration in the caption). PolyGen ($m = 2$, full configuration) reaches $\Delta_{\text{MTL}} = +12.5\%$ against $+7.3\%$, and SugarCrepe++ $= 40.5$ against 36.2: the multi-positive formulation, generator-invariant training and curriculum each contribute beyond what a TripletCLIP-style objective achieves in the fully synthetic regime.

The Limits of Passive and Delayed Hard Negative Injection. Injecting hard negatives naively yields strong compositional performance (SugarCrepe++ $= 40.0$) from constant exposure to fine-grained variations, but this early complexity distracts general semantic alignment ($\Delta_{\text{MTL}} = +4.7\%$). The scheduler alone shields early training and lets standard tasks improve ($\Delta_{\text{MTL}} = +7.0\%$), at the cost of weaker compositional reasoning. Timing matters, and neither extreme is optimal in isolation.

Impact of the Hard Negative Loss. $\mathcal{L}_{HN}$ in isolation proves effective, yielding higher Δ_{MTL} than the equivalent $\mathcal{L}_{MP}$ configuration. The TripletCLIP formulation drops $T^+ \to I^-$ and $T^- \to I^+$ comparisons; since generated images often fail to render subtle caption variations faithfully (Appendix C.3), these introduce noise, and removing them lets the model exploit hard negatives via $I^+ \to T^-$ and $I^- \to T^+$ (derivation in Appendix B.2). Combining scheduler and $\mathcal{L}_{HN}$ gives the full configuration ($\Delta_{\text{MTL}} = +12.5\%$): the scheduler protects early feature formation from noisy gradients, while $\mathcal{L}_{HN}$ offsets the reduced exposure with a stricter margin.

The Schedule Family. Since p saturates at 0.5 (Section 3.3), the schedule has two free parameters, endpoint and ramp speed; four points span the feasible range, from a conservative ramp to p=0.25 up to constant p=0.5. Ordered by cumulative exposure (Table 13, Appendix N), they separate two behaviors. General transfer is a dose effect: zero-shot falls monotonically with exposure, then saturates ($11.24 \to 10.80 \to 10.23 \to 10.27$). Compositional reasoning is not governed by exposure at all: the least-exposed schedule leads on both benchmarks ($+2.70$ SugarCrepe, $+2.04$ SugarCrepe++ over the linear ramp), while the other three lie within 1.6 points of each other with no ordering by dose. Naive injection is therefore not the compositional ceiling maximal exposure would suggest: timing governs the outcome across a family of schedules, not at a single setting. The endpoint is a genuine trade-off rather than a free improvement: the linear ramp stays best on retrieval and the no-scheduler run on vision-only probes.

Are Rendered Image Negatives Necessary? Prior work obtains attribute binding from text-side negatives alone, so we ablate the rendered image negatives I^- at m=2, n^+=2 without scheduler, where $\mathcal{L}_{HN}$ reduces to NegCLIP (Yuksekgonul et al., 2022). Since dropping I^- halves images per epoch, two runs are needed to separate negative type from budget (Appendix O). At matched budget, removing rendered image negatives costs 3.00 zero-shot, 1.36/4.71 retrieval R@5 and 9.43 SugarCrepe++, and doubling the image budget recovers most of the retrieval deficit ($+3.76$ image R@5) but almost none of the compositional one ($+0.66$). Text-only negatives are not inert, improving text-text discrimination ($+3.68$ SC++ TTT) while degrading image-text (-6.90): they train attribute binding on the text side, but the signal does not reach image-text alignment without a rendered image negative.

Impact on Compositional Reasoning. Scaling m and n^+ yields monotonic gains for general open-world tasks, with $m = 4$ reaching the Δ_{MTL} peak at $+15.5\%$. On compositional reasoning the conservative schedule at $m = 2, n^+ = 2$ is strongest overall ($\Delta^{HN}_{\text{MTL}} = +11.3\%$); among m-scaled configurations $m = 3$ leads on both benchmarks ($+9.1\%$), with $m = 4$ close behind ($+6.8\%$). Per-axis decomposition (Table 9) shows $m = 3$ leading on Replace, Swap and the average, with $m = 4$ ahead only on attribute addition: $m = 4$ is Pareto-optimal for joint general and compositional transfer, $m = 3$ preferable when fine-grained syntactic discrimination is the priority.

4.4 Preliminary Scaling Analysis

To assess whether the multi-generator effect persists beyond the controlled 500k budget, we transfer the best PolyGen configuration ($m = 4$) to 3M and 12M images against a matched single-generator baseline (SD 1.5). PolyGen retains a positive margin at both scales, with the comparison caption-asymmetric in the baseline's favour: it uses 12M unique captions ($n^+ = 1$) whereas PolyGen matches it with $\sim 4\times$ fewer. The relative benefit per image is nonetheless markedly smaller than at 500k, which we attribute to generator-specific saturation under uniform allocation; full results and the resulting case for scale-adaptive allocation are in Appendix J.

5 Discussion and Limitations

The Synergy of Diversity and Recognizability. PolyGen integrates complementary manifolds: high-fidelity generators anchor the text concept, high-diversity generators expand its visual support and prevent overfitting to generator artifacts. The signal shared across the ensemble in aggregate is the invariant semantic content, and Appendix L shows $\mathcal{L}_{I2I}$ to be what converts that cross-source variation into invariance, collapsing the between-generator decision margin to the random-feature floor rather than erasing generator information outright.

Limitations and Future Work. First, a heterogeneous ensemble introduces higher pipeline complexity than single-model baselines, though not higher generation cost: fast Recognizability Experts offset slower Diversity Experts, so only the m=4 ensemble is costlier per image than SD 1.5 alone (Appendix C.4), and all generation is offline. Second, at fixed image budget the multi-positive strategy prioritizes semantic robustness over unique concept coverage; Section 4.4 sharpens this into a concrete recalibration toward scale-adaptive allocation. Third, hard negatives come from a single LLM and may carry its biases. We mitigate this by selecting the negative generator through a controlled axis-recovery ablation (Appendix C.1) and removing axes it cannot modify in isolation (Appendix D), but residual bias remains possible; mixing several negative-generating LLMs is the natural extension. Fourth, $\mathcal{L}_{I2I}$ trades texture sensitivity for semantic abstraction (Section 4.2), so it is an operating point rather than a default. Finally, the evaluation follows the standard contrastive pre-training protocol, covering classification, retrieval and compositional reasoning; dense prediction (segmentation, detection) requires architectural adaptations beyond this protocol (Radford et al., 2021; Hammoud et al., 2024; Tian et al., 2023; Patel et al., 2024), and the concept bank and semantic axes, while broad, remain a simplification of real-world visual complexity.

Future work will expand the generator pool with architectures occupying complementary regions of the Recognizability/Diversity space, jointly increasing m and n^+; scale the image budget to characterize the breadth/depth trade-off; and extend programmatic hard negatives to axes such as multi-concept relations, counting and causality.

6 Conclusions

We introduced PolyGen, a fully synthetic vision-language pre-training framework grounded in the Generator-Invariance Hypothesis: an ensemble of architecturally distinct generators biases representations toward the signal shared across sources, the underlying semantic content. We treated it as a training principle with two testable implications, ensemble composition and invariance regularization, and paired it with hard-negative timing, each validated by seeded ablations. PolyGen achieves $+19.0\%$ in Δ_{MTL} over the single-source baseline and $+11.3\%$ on compositional benchmarks, and exceeds a real-data CC3M-subset baseline on retrieval while closing two thirds of the zero-shot gap. Preliminary larger-scale results indicate the effect persists beyond this regime.

ETHICS STATEMENT

A primary motivation for PolyGen is to reduce reliance on large-scale web scraping, which inherently risks including private, non-consensual, or harmful content. Fully synthetic pipelines offer precise control over data composition, enabling privacy-preserving and bias-aware dataset construction. Our pipeline builds on generative models with safety alignment and content filtering (Llama 3, Stable Diffusion, SANA) and grounds caption generation in the MetaCLIP concept bank, a curated distribution designed to reduce the frequency imbalances and harmful associations present in raw web-crawled text. We acknowledge that synthetic generators may still reflect biases from their pre-training data, and that the same techniques could in principle be repurposed to amplify specific biases at scale.

REPRODUCIBILITY STATEMENT

The full training configuration (ViT-B/16 architecture, optimizer, learning rate, batch size, schedule, and curriculum parameters) is reported in Section 4. The caption and hard-negative generation prompts are given in Appendix A; generator selection criteria and inference settings in Appendices C.1, C.3, and C.4; the balanced sampling procedure used to build the caption pool in Appendix G; and the formal loss derivations in Appendix B. Code and the generated synthetic datasets are withheld at submission time to preserve anonymity and will be released upon acceptance.

AI USE STATEMENT

Generative models are the object of study of this work: LLMs and text-to-image models are used as components of the proposed data generation pipeline, as described in Section 3. Beyond this, generative AI tools were used only for writing assistance (language polishing). We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including text, claims, and artifacts.

REFERENCES

Krzysztof Adamkiewicz, Brian Moser, Stanislav Frolov, Tobias Christian Nauen, Federico Raue, and Andreas Dengel. When pretty isn't useful: Investigating why modern text-to-image models fail as reliable training data generators. *arXiv preprint arXiv:2602.19946*, 2026.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *CoRR*, 2024.

Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Shama Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. Sugarcrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations. *Advances in Neural Information Processing Systems*, 37:17972–18018, 2024.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36:35544–35575, 2023.

Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7382–7392, 2024.

Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. Synthclip: Are we ready for a fully synthetic clip training? *arXiv preprint arXiv:2402.01832*, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

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. *Advances in neural information processing systems*, 30, 2017.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20406–20417, 2023.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL `https://arxiv.org/abs/2310.06825`.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images., 2009.

Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, et al. Veclip: Improving clip training via visual-enriched captions. In *European Conference on Computer Vision*, pp. 111–127. Springer, 2024.

Andrea Lampis, Eugenio Lomurno, and Matteo Matteucci. Bridging the gap: Enhancing the utility of synthetic data via post-processing techniques. 2023.

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

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Maitreya Patel, Naga Sai Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, et al. Tripletclip: Improving compositional reasoning of clip via synthetic vision-language negatives. *Advances in neural information processing systems*, 37:32731–32760, 2024.

Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision-language pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6967–6977, 2023.

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

Nicolò Francesco Resmini, Eugenio Lomurno, Cristian Sbrolli, and Matteo Matteucci. Your image generator is your new private dataset. *Image and Vision Computing*, pp. 105727, 2025.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

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*, pp. 10684–10695, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

TII Team. The falcon 3 family of open models, December 2024.

Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36:48382–48402, 2023.

Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15887–15898, 2024.

Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3614–3633, 2021.

Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*, 1:1, 2022.

Haonan Wang, Minbin Huang, Runhui Huang, Lanqing Hong, Hang Xu, Tianyang Hu, Xiaodan Liang, Zhenguo Li, Hong Cheng, and Kenji Kawaguchi. Getting more juice out of your data: Hard pair refinement enhances visual-language models without extra data. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7854–7873, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

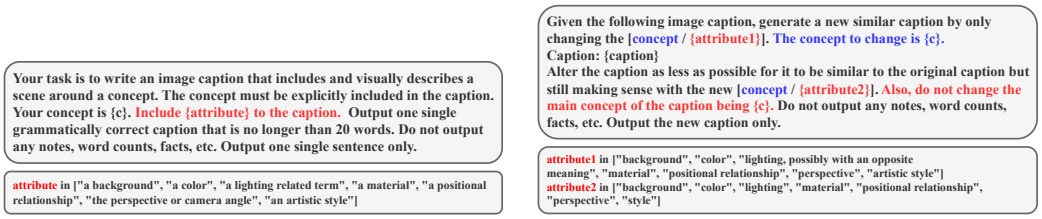
Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformer, 2024. URL https://arxiv.org/abs/2410.10629.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.



(a) Prompt used by Mistral for the generation of the baseline caption set. Red text is only included when an attribute indication is added.

(b) Prompt used by Llama for the generation of the hard negative caption set. When the original caption is generated without using any attribute, c is used as the semantic axis to change.

Figure 3: Zero-shot prompts for the generation of captions.

A Caption Generation Prompts

Figure 3 shows the zero-shot prompts used to generate captions. The baseline set of captions is generated with the prompt shown in Figure 3a, asking the model to generate a caption given the starting concept $c \in \mathcal{C}$, and optionally asking it to include an attribute a from a predefined list of semantic modification axes $\mathcal{A}$. Hard negative captions are then generated with the prompt in Figure 3b, which instructs the LLM to modify the base caption on its selected semantic axis only, while maintaining the rest of the sentence structure.

B Loss Functions: Derivations

This appendix provides the formal derivations of the two loss components used in PolyGen: the image-to-image invariance term $\mathcal{L}_{I2I}$ (Section B.1) and the relationship between the standard CLIP loss and the TripletCLIP loss adopted as the basis of $\mathcal{L}_{HN}$ (Section B.2).

B.1 Derivation of $\mathcal{L}_{I2I}$

The image-to-image invariance term from StableRep (Tian et al., 2023) is designed to align the embeddings of multiple images rendered from the same caption, suppressing generator-specific idiosyncrasies. Let $I = \{i_1, \ldots, i_N\}$ denote a batch of N images and let $\mathcal{P}(i)$ denote the set of positive companions of anchor i, that is, the other images generated from the same caption as i. Under the multi-positive setting with n^+ images per caption, $|\mathcal{P}(i)| = n^+ - 1$.

Ground-truth distribution. For any image anchor i, the ground-truth distribution $p_{i,\cdot}$ over the remaining batch elements $j \in I \setminus \{i\}$ is uniform over the positive companions and zero elsewhere:

$$p_{i,j} = \begin{cases} \dfrac{1}{n^+ - 1} & \text{if } j \in \mathcal{P}(i), \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

By construction, $\sum_{j \in I \setminus \{i\}} p_{i,j} = 1$ for every anchor i, so $p_{i,\cdot}$ is a valid categorical distribution.

Predicted distribution. The predicted distribution $q_{i,\cdot}$ is the softmax over image-to-image cosine similarities:

$$q_{i,j} = \frac{e^{s(i,j)/\tau}}{\displaystyle\sum_{k \in I \setminus \{i\}} e^{s(i,k)/\tau}}, \tag{5}$$

where $s(i, j)$ denotes cosine similarity between image embeddings $f_{\text{img}}(i)$ and $f_{\text{img}}(j)$, and τ is the temperature parameter shared with the standard contrastive objective.

Cross-entropy formulation. $\mathcal{L}_{I2I}$ is the cross-entropy between $p_{i,\cdot}$ and $q_{i,\cdot}$ averaged over anchors (Equation 1); substituting Equation 4 yields the equivalent compact form

$$\mathcal{L}_{I2I} = -\frac{1}{N(n^+ - 1)} \sum_{i \in I} \sum_{j \in \mathcal{P}(i)} \log q_{i,j}, \tag{6}$$

which makes explicit that $\mathcal{L}_{I2I}$ is the negative log-likelihood of the positive companions under the softmax over the remaining batch, averaged over anchors and positives.

Relation to single-positive InfoNCE. Equation 6 generalizes the standard InfoNCE objective. For $n^+ = 2$, the ground-truth distribution collapses to a one-hot vector on the single positive companion j_i^+, and $\mathcal{L}_{I2I}$ reduces to the symmetric pairwise InfoNCE between the two renderings of the same caption, $-\frac{1}{N}\sum_{i\in I}\log q_{i,j_i^+}$. For $n^+ > 2$, $\mathcal{L}_{I2I}$ averages this contribution over all positive companions, which is equivalent to maximizing the geometric mean of the per-pair softmax probabilities. This formulation matches the soft-target multi-positive formulation of Tian et al. (2023), applied here to image-image rather than text-image pairs.

Role as a structural regularizer. $\mathcal{L}_{I2I}$ does not introduce additional supervision: the multi-positive structure of the batch is already exploited by the text-to-image alignment objective $\mathcal{L}_{MP}$. Its purpose is to impose an explicit consistency constraint on the visual encoder, requiring that images of the same caption produce mutually similar embeddings irrespective of the rendering generator. Empirically (Table 1, Section 4.2), this constraint sharpens generator invariance at the measurable cost of low-level texture-sensitive transfer, motivating its presentation as an optional component of the framework.

B.2 Relation between CLIP and TripletCLIP Loss

Let $T^+ = \{t_k^+\}_{k=1}^N$ and $I^+ = \{i_k^+\}_{k=1}^N$ be the features of the base captions and their images, $t_k^+ = f_{txt}(\text{caption}_k)$ and $i_k^+ = f_{img}(\text{image}_k)$, and let T^- and I^- be the features of the hard negatives, obtained from $\text{caption}_k^- = \text{LLM}(\text{caption}_k, \text{hard negative prompt})$ and $\text{image}_k^- = \text{T2I}(\text{caption}_k^-)$. With $\tilde{T} = T^+ \cup T^-$ and $\tilde{I} = I^+ \cup I^-$ the concatenated batches of size $2N$, the standard CLIP loss is $\mathcal{L}_{CLIP}(\tilde{I},\tilde{T}) = \frac{1}{2}(\mathcal{L}_{NCE}^{\tilde{I}\to\tilde{T}} + \mathcal{L}_{NCE}^{\tilde{T}\to\tilde{I}})$. Splitting the sum over the $2N$ anchors of each direction into positive anchors and negative anchors gives

$$\mathcal{L}_{CLIP}(\tilde{I},\tilde{T}) = \frac{1}{4}\left(\mathcal{L}_{NCE}^{I^+\to T^+;T^-} + \mathcal{L}_{NCE}^{I^-\to T^-;T^+} + \mathcal{L}_{NCE}^{T^+\to I^+;I^-} + \mathcal{L}_{NCE}^{T^-\to I^-;I^+}\right).$$

The two text-anchored terms decompose as $\mathcal{L}_{NCE}^{T^+\to I^+;I^-} = \mathcal{L}_{NCE}^{T^+\to I^+} + C_{T^+\to I^-}$ and $\mathcal{L}_{NCE}^{T^-\to I^-;I^+} = \mathcal{L}_{NCE}^{T^-\to I^-} + C_{T^-\to I^+}$, where $C_{T^+\to I^-}, C_{T^-\to I^+} \geq 0$ are log-sum-exp terms arising from the additional hard negative image pairings. The TripletCLIP loss removes exactly these terms,

$$\mathcal{L}_{TripletCLIP}(\tilde{I},\tilde{T}) = \mathcal{L}_{NCE}^{I^+\to T^+;T^-} + \mathcal{L}_{NCE}^{T^+\to I^+} + \mathcal{L}_{NCE}^{I^-\to T^-;T^+} + \mathcal{L}_{NCE}^{T^-\to I^-},$$

so that

$$\mathcal{L}_{CLIP}(\tilde{I},\tilde{T}) = \frac{1}{4}\left(\mathcal{L}_{TripletCLIP}(\tilde{I},\tilde{T}) + C_{T^+\to I^-} + C_{T^-\to I^+}\right),$$

showing that the standard CLIP loss on the full batch is lower-bounded by $\frac{1}{4}$ of the TripletCLIP loss, the difference being the $T^+ \to I^-$ and $T^- \to I^+$ pairings that TripletCLIP discards.

C Generative Models Selection

C.1 LLM Selection for Hard Negative Captions

The role of the LLM in PolyGen is to generate axis-conditioned counterfactual captions: given a base caption $t \in T^+$, a target semantic axis $a \in \mathcal{A}$, and a concept c, the LLM must modify only the dimension specified by a while preserving the rest of the caption structure. The quality of the resulting hard negatives depends on how reliably the LLM follows this instruction. We selected Llama-3.1-8B as the hard negative generator for PolyGen based on the controlled axis-recovery ablation described below.

Task setup. We evaluate a pool M of similar-sized open-weight LLMs: Mistral-V0.2-7B and Mistral-V0.3-7B (Jiang et al., 2023), Llama-3.1-8B (Dubey et al., 2024), Falcon-3-7B (Team, 2024), and Qwen-2.5-7B (Yang et al., 2024). The size range (7B/8B) was chosen to control for capacity. For each candidate model $m \in M$, we generate 4.8k caption pairs covering five semantic axes $\mathcal{A} = [\textit{background}, \textit{color}, \textit{position}, \textit{size}, \textit{concept}]$, with 960 pairs per axis. To detect cases where the LLM fails to introduce any meaningful change, 160 pairs per axis are also generated as an *unaltered* control class using the prompt in Figure 4 (top-left).

Rephrase the provided caption by rearranging its words while keeping the meaning intact.
Caption: {caption}
Do not output any notes, word counts, facts, etc.
Output the new caption only.

Classify what mostly differs between the two images into one of the following categories:

concept: The scene's subject differs.
background: The scene's background differs.
position: The positional relationship of one or more elements differ.
color: The colors of one or more elements differ.
size: The size of one or more elements differ.
unaltered: Both images show the same scene.
[other labels]

It is mandatory to choose exactly one label. You cannot avoid selecting one of the options, even if the differences seem ambiguous.

Output the label and nothing else. Do not include explanations, notes, reasoning, etc.

Classify the semantic difference between two captions into one of the following categories:

concept: The scene's concept differs.
background: The scene's background differs.
position: The described positional relationship differs.
color: The described colors differ.
size: Size adjectives differ.
unaltered: Both captions describe the same scene.
[other labels]

It is mandatory to choose exactly one label. You cannot avoid selecting one of the options, even if the differences seem ambiguous.

The concept of the first caption is {c}.
First caption: {base_caption}
Second caption: {hn_caption}

Output the label and nothing else. Do not include explanations, notes, reasoning, etc.

Figure 4: The prompt used to generate the *unaltered* caption (top-left). The prompt used to classify pairs of images (bottom-left). The prompt used to classify pairs of captions (right).

Model	Accuracy	Avg. F1	Avg. Prec.	Avg. Rec.
Mistral V0.2	61%	60%	76%	61%
Mistral V0.3	62%	62%	77%	62%
Qwen 2.5	67%	67%	78%	67%
Falcon 7B	63%	65%	79%	63%
Llama 3.1	**74%**	**75%**	**80%**	**74%**

Table 3: Aggregated results of the axis-recovery ensemble classification. Llama-3.1-8B exhibits superior adherence to semantic constraints, minimizing false negatives in the counterfactual generation process.

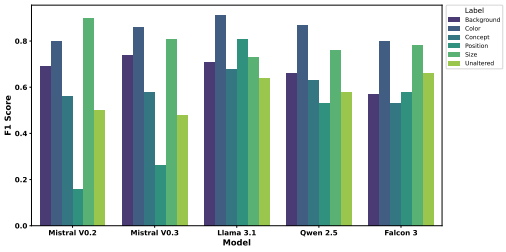


Figure 5: F1-scores per semantic axis for each LLM over the final axis set. *Position* and *concept* discriminate the candidates most.

Recovery task. Given a positive-negative caption pair (c_1, c_2) produced by candidate model m via modification axis a, we evaluate whether a can be recovered from the pair. Recovery is performed by an ensemble of LLM judges, each independently classifying the pair into one of the five axis categories or the *unaltered* class using the classification prompt in Figure 4 (right). The final classification is determined by majority voting. High recovery accuracy indicates that the candidate model produces semantically isolated modifications: a strong axis-aligned signal in the pair is the necessary precondition for the corresponding hard negative to be useful in contrastive training.

Results. Table 3 reports aggregated metrics across the five axes. Llama-3.1-8B achieves the highest accuracy (74%) and F1 score (75%), with substantial margins over the next best alternative (Qwen-2.5, 67% / 67%). The Mistral and Falcon variants cluster between 61% and 63% accuracy, performing closer to the noise floor expected from partially axis-aligned generations. Average precision is comparable across models (76–80%) but recall varies significantly (61–74%): models with low recall produce frequent false negatives where a real modification is missed by the judges, indicating that the original perturbation was either too weak or too entangled with other axes. Llama-3.1-8B exhibits the strongest balance, minimizing both false negatives and ambiguous outputs.

Per-axis behavior. Figure 5 reports F1 scores per semantic axis for each candidate. Some axes (notably *texture*, evaluated in the broader set discussed in Appendix D) are inherently harder to perturb in isolation: modifying the texture of a subject often forces co-modifications of material or lighting in the surrounding scene, producing entangled changes that the judge cannot resolve to a single axis. Llama-3.1-8B leads on most axes, with the largest absolute gain on *position* and *concept*, the two axes that most directly drive compositional reasoning evaluations on SugarCrepe and SugarCrepe++.

C.2 LLM Selection for Baseline Captions

The choice of Mistral-V0.2-7B for generating the base caption set has two motivations. First, it ensures a fairer comparison with the SynthCLIP pipeline, which uses the same model, shifting the focus toward our prompting strategy and framework contributions. Second, we conduct an ablation generating a 500k image-caption dataset for both Mistral-V0.2-7B and Llama-3.1-8B with images

Model	Res.	Steps	FID↓	CLIPScore↑	TIFA↑
Real	512^2	–	–	30.319	88.8
SD 1.5	512^2	50	23.014	29.962	78.7
SDXLT	512^2	2	32.593	**31.748**	87.9
SD 2	512^2	20	53.091	27.310	69.0
		50	55.348	27.244	69.0
	768^2	20	23.613	30.065	80.6
		50	**21.095**	30.162	80.4
SD 2.1	512^2	20	61.443	27.035	68.5
	768^2	50	23.481	30.068	82.5
SD 3.5 LT	512^2	4	35.448	31.586	88.1
	1024^2	4	36.732	31.484	89.4
SANA	512^2	20	27.316	31.163	88.1
	1024^2	20	27.573	31.171	88.0
FLUX	512^2	4	31.583	31.634	90.5
	1024^2	4	28.581	31.403	**90.9**

Table 4: Model comparison across resolutions and inference steps. Models used or usable as *Diversity* or *Recognizability* experts are highlighted accordingly.

Model	Res.	Steps	Accuracy	Avg F1	Avg Prec.	Avg Rec.
SD 1.5	512^2	50	25%	23%	28%	25%
SDXLT	512^2	2	37%	33%	40%	37%
SD 3.5 LT	512^2	4	34%	30%	36%	36%
	1024^2	4	35%	31%	35%	35%
FLUX	512^2	4	36%	33%	42%	36%
	1024^2	4	**38%**	**35%**	**43%**	**39%**

Table 5: Semantic axis classification accuracy for T2I models. High-fidelity Recognizability Experts show stronger attribute binding, essential for rendering fine-grained hard negative distinctions.

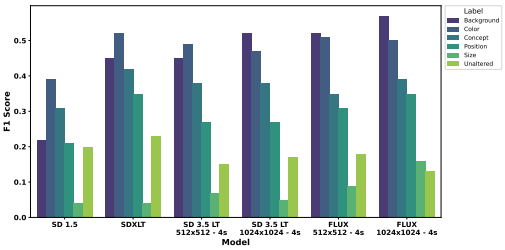


Figure 6: F1-scores per semantic axis for each T2I model. SDXL-Turbo and FLUX demonstrate superior rendering of spatial and object-level changes.

from SDXL-Turbo. We report a slight advantage for Mistral (1.8% top-1, 5.97% top-5) over Llama (1.66% top-1, 5.61% top-5) on ImageNet zero-shot, further validating our choice.

C.3 Diffusion Model Selection

Image quality metrics. The selection of generators (SD 1.5, SD 2, SANA, SDXL-Turbo) results from a comparison against a broader pool also comprising SD 2.1, SD 3.5 Large Turbo, and FLUX.1 Schnell. We randomly sample one caption for each of the 5k samples of MS COCO 2017 (Lin et al., 2014) validation set and generate the corresponding image with each model, comparing based on FID (Heusel et al., 2017), CLIPScore (Hessel et al., 2021), and TIFA (Hu et al., 2023).

Semantic axis classification. We evaluate diffusion models on their ability to render hard negative caption variations using Qwen 2.5 VL (Wang et al., 2024) as the visual classifier. We use 2.4k image pairs (480 per axis, 80 per unaltered class). Results are shown in Table 5: higher image-caption alignment correlates with better hard negative rendering, justifying the use of the TripletCLIP loss to compensate for the noisier visual signal.

C.4 Per-Generator Inference Cost

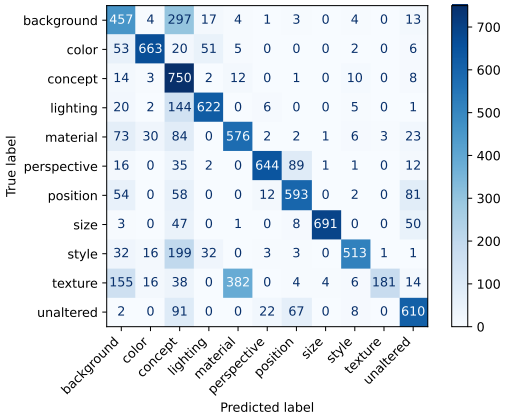
Table 6 reports the per-image inference time for each generator, measured on a single A100 GPU. All generation is performed offline prior to CLIP training and incurs no cost at training or inference time.

D Semantic Axes Selection

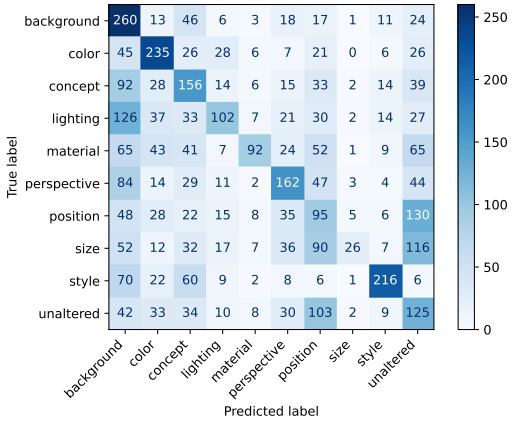
Given the semantic axis classification results for both LLMs and T2I models (Figures 5 and 6), the ability to generate effective hard negatives varies significantly by axis. The final set is obtained by a second classification round from a broader set $\mathcal{A} = [\textit{background}, \textit{color}, \textit{lighting}, \textit{material}, \textit{perspective}, \textit{position}, \textit{size}, \textit{style}, \textit{texture}, \textit{concept}]$, on caption pairs generated by Llama and image pairs generated by SDXLT. Figure 7 shows the confusion matrices. Llama struggles with *texture* modifications without altering global context; SDXLT-generated images fail to represent *size* differences reliably. Both axes are removed from the final set.

Generator	Resolution	Steps	Time/image
SD 1.5	512^2	50	2.50s
SD 2	768^2	50	7.29s
SANA 1.6B	512^2	20	3.75s
SDXL-Turbo	512^2	2	0.22s
Ensemble weighted averages (equal image allocation)			
SD 1.5 only (baseline)	–	–	2.50s
SD 1.5 + SDXL-Turbo ($m = 2$)	–	–	1.36s
SD 1.5 + SANA + SDXL-T ($m = 3$)	–	–	2.16s
Full ensemble ($m = 4$)	–	–	3.44s

Table 6: Per-image generation time (single A100 GPU) and weighted-average cost for representative ensemble configurations under equal image allocation. The $m = 2$ and $m = 3$ ensembles are faster than the SD 1.5 baseline due to SDXL-Turbo's 2-step distillation. Ensemble overhead is not proportional to m: it depends on composition and can be minimized by pairing fast Recognizability Experts with slower Diversity Experts.



(a) Llama results. *Texture* is removed from the final set.



(b) SDXLT results. *Size* is removed from the final set.

Figure 7: Confusion matrices for the hard negative classification task on an extended attribute set. High off-diagonal mass for *texture* (Llama) and *size* (SDXLT) motivates their removal from $\mathcal{A}$.

E T2I Models Comparison in Frequency Domain

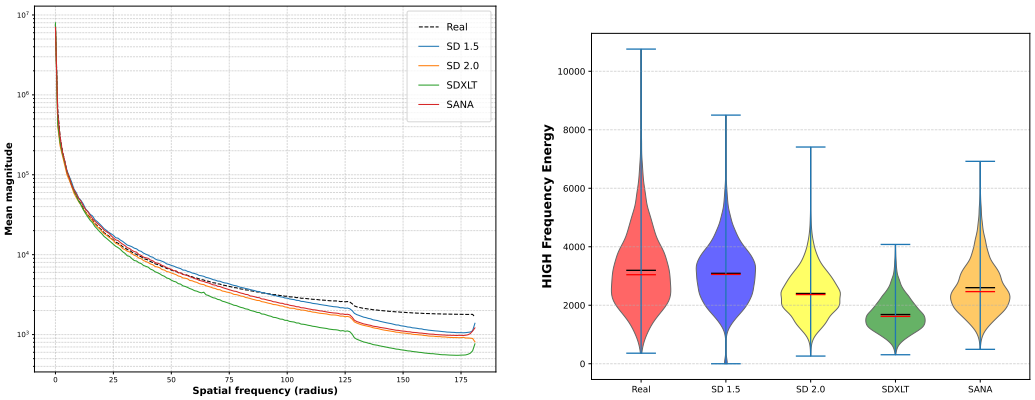
To characterize why different diffusion models lead to different downstream performance, we analyze generated images in the frequency domain. The 2D Fast Fourier Transform (FFT) is applied to each standardized image; from the 2D magnitude spectrum we extract two metrics:

- **Radial Frequency Profile:** The 2D spectrum is reduced to a 1D vector via radial averaging, representing mean magnitude as a function of spatial frequency. Averaged over all 5k images in the set.
- **Frequency Energy:** Total energy in the high-frequency band $[0.5, 1]$ (normalized units), quantified per image.

Figure 8 shows that diffusion models act as low-pass filters with distinct frequency responses. SDXLT exhibits the steepest high-frequency roll-off, consistent with its distillation objective. These generator-specific spectral signatures are the artifacts that $\mathcal{L}_{I2I}$ is designed to suppress.

F Hard-Negative-Aware Loss Functions

Figure 9 presents an ablation comparing $\mathcal{L}_{CLIP}$ with and without $\mathcal{L}_{margin}$ (Wang et al., 2025), $\mathcal{L}_{HN\text{-}NCE}$ (Radenovic et al., 2023), and $\mathcal{L}_{TripletCLIP}$ (Patel et al., 2024). TripletCLIP overcomes



(a) Radial profiles of the employed models.

(b) Frequency energy distribution of the 50% highest frequencies.

Figure 8: Frequency analysis of 5k generated images from MS COCO captions, compared with real MS COCO images. SD 1.5 most closely resembles real images in spectral characteristics; SDXLT shows the most significant deviation, with the steepest high-frequency attenuation.

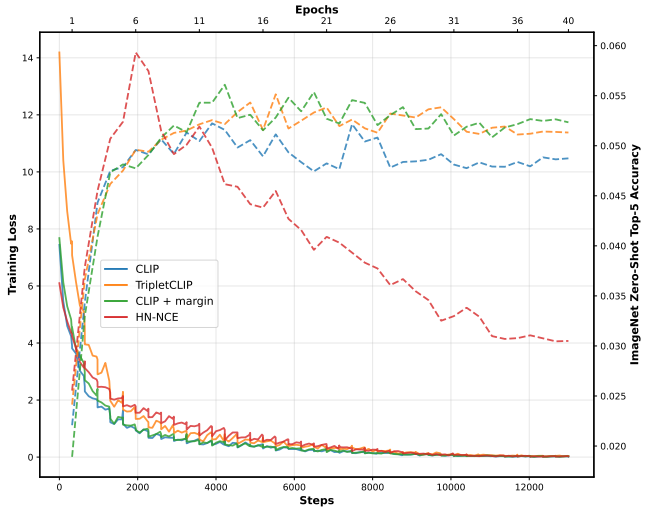


Figure 9: Loss function ablation on 500k Mistral/Llama + SDXLT dataset. $\mathcal{L}_{HN\text{-}NCE}$ shows a strong initial boost followed by rapid degradation. $\mathcal{L}_{TripletCLIP}$ and $\mathcal{L}_{CLIP} + \mathcal{L}_{margin}$ consistently outperform standard $\mathcal{L}_{CLIP}$.

the excessively noisy signal introduced by hard negative images while preserving the image-to-text direction. The margin loss also improves performance by enforcing a relative similarity hierarchy, though it is most effective in a 1-to-many anchor-to-negative setting not present in our current design. $\mathcal{L}_{HN\text{-}NCE}$ peaks early but degrades rapidly, likely due to an excessively punishing weighting scheme in this setting.

G Balanced Sampling

Balanced sampling (Xu et al., 2023) addresses the rich-get-richer problem in web-crawled data, where common concepts appear millions of times while rare concepts are underrepresented. Given a threshold t, each concept $c \in \mathcal{C}$ receives sampling probability $t/\text{count}(c)$ if $\text{count}(c) > t$, and probability 1 otherwise. We generate an initial pool of 100M captions, de-duplicate, and apply the algorithm with $t = 30$, yielding 13.1M captions. We then uniformly downsample to the target size.

The procedure flattens the concept-frequency distribution, reducing the dominance of high-frequency concepts while retaining coverage of the long tail.

H Delta Multi-Task Learning Metric

To aggregate metrics that are not directly comparable across tasks, Vandenhende et al. (2021) defined the Δ_{MTL} metric, which computes the average relative performance gain or drop of a model m with respect to a baseline b over a set of tasks T:

$$\Delta_{MTL} = 100 \times \frac{1}{T} \sum_{i=1}^{T} (-1)^{l_i} \frac{M_{m,i} - M_{b,i}}{M_{b,i}}, \tag{7}$$

with $l_i = 1$ if a lower value indicates better performance for metric i, and 0 otherwise. We write Δ_{MTL}^{HN} for the same quantity computed over the compositional benchmarks, that is, the average relative gain over the baseline across SugarCrepe and SugarCrepe++.

I Complete PolyGen Results

Tables 7 and 8 report per-dataset results for all ensemble (m) and density (n^+) configurations, with and without $\mathcal{L}_{I2I}$, and, in a separate block, for models incorporating hard negative samples. Compositional reasoning results are reported in Table 9.

J Preliminary Scaling Results

This appendix reports the preliminary scaling experiments referenced in Section 4.4. The 500k regime, used as the primary evaluation in the main text, supports full ablation over m, n^+, $\mathcal{L}_{I2I}$, hard negatives, and curriculum under a controlled budget. To assess whether the multi-generator effect persists at larger image budgets, we transfer the best PolyGen configuration ($m = 4$, multi-positive) to two additional regimes and compare against a matched single-generator baseline (SD 1.5).

Setup. At each scale we train ViT-B/16 CLIP from scratch following the protocol described in Section 4.3, with two modifications. First, the total image budget B is increased to 3M and 12M images, respectively. Second, the hyperparameter configuration of both the PolyGen run and the matched single-generator baseline is held fixed at the value identified at 500k; we did not perform a scale-specific sweep over ensemble composition, learning rate schedule, or batch size. In the 12M setting, the PolyGen run renders approximately 3M unique captions through four generators ($n^+ = 4$), while the SD 1.5 baseline uses 12M unique captions ($n^+ = 1$); the total number of training images is matched.

Results. Table 10 reports the ImageNet top-5 zero-shot accuracy at each scale. PolyGen retains a positive margin over the SD 1.5 baseline at both 3M and 12M images. The margin is consistent in sign across both regimes ($+0.25$ at 3M, $+0.64$ at 12M, both on top-5 accuracy in the 25–45 range), indicating that the multi-generator effect observed at 500k does not vanish with increased data volume. Both margins should be read against the measurement noise: at 500k the per-cell zero-shot standard deviation over three seeds reaches ± 0.23 (Table 12), so the $+0.25$ at 3M lies within single-seed variance while the $+0.64$ at 12M sits outside it. These runs are single-seed, and we treat the agreement in sign across two independent scales, rather than either margin alone, as the persistence evidence. The comparison is also caption-asymmetric: the baseline draws on 12M unique captions (n^+=1), whereas PolyGen matches it with $\sim$4$\times$ fewer captions, reallocating the image budget toward caption-structural diversity (n^+=4).

Caveats and scope. We label these experiments as preliminary for two reasons: no scale-specific tuning was performed, so the relative contributions of m and n^+ and the curriculum schedule of p may require recalibration as the budget grows; and the evaluation is limited to ImageNet zero-shot, whereas a scaling-law characterization would require retrieval and compositional evaluation at each budget. We therefore read these results as persistence evidence outside the controlled 500k regime, not as a characterization of how the effect scales. The relative benefit per image is nonetheless markedly smaller than at 500k. A mechanism consistent with this, though not one we have verified,

			HN Components			Vision Datasets									
n^+	m	$\mathcal{L}_{I2I}$	Smpl	Loss	Sched	C10	C100	Air	DTD	Flwr	Pets	SUN	Cal	Food	**Avg**
Linear Probing															
1	1	✗	✗	✗	✗	72.5	50.8	24.7	**47.2**	69.7	38.9	39.6	51.9	46.5	49.1
1	2	✗	✗	✗	✗	72.9	51.0	20.4	44.1	67.6	36.7	38.0	51.7	45.4	47.5
1	3	✗	✗	✗	✗	73.0	51.6	23.5	44.3	70.9	37.4	38.1	51.9	45.8	48.5
1	4	✗	✗	✗	✗	73.1	51.4	23.8	44.7	71.1	36.8	39.1	53.5	46.8	48.9
2	1	✗	✗	✗	✗	72.2	50.3	23.4	47.1	**71.6**	39.2	40.3	54.5	47.4	49.5
2	2	✗	✗	✗	✗	73.0	49.0	22.2	45.9	68.3	36.2	39.4	51.0	46.2	47.9
3	1	✗	✗	✗	✗	72.8	51.3	24.6	46.9	70.6	**39.8**	**41.1**	50.9	46.1	49.3
3	3	✗	✗	✗	✗	73.6	51.6	22.8	45.2	68.5	35.4	39.0	51.1	46.8	48.2
4	1	✗	✗	✗	✗	72.8	51.3	**25.3**	46.9	70.2	38.9	40.9	**54.9**	47.3	**49.8**
4	4	✗	✗	✗	✗	**73.9**	**51.8**	23.9	45.1	69.0	37.7	40.1	50.5	**47.6**	48.9
2	1	✓	✗	✗	✗	70.5	47.2	20.7	41.5	57.7	31.9	34.4	42.8	43.6	43.4
2	2	✓	✗	✗	✗	69.8	47.6	19.4	37.3	54.6	31.0	32.6	40.2	40.7	41.5
3	1	✓	✗	✗	✗	71.7	49.2	20.9	43.5	58.4	34.9	36.0	42.8	42.8	44.4
3	3	✓	✗	✗	✗	70.5	48.3	19.5	37.6	52.8	31.5	31.8	39.1	40.5	41.3
4	1	✓	✗	✗	✗	72.4	50.7	21.4	44.5	57.8	34.9	36.2	43.7	45.0	45.2
4	4	✓	✗	✗	✗	71.6	49.7	19.9	41.1	56.0	31.5	33.2	39.8	43.0	42.9
2	2	✓	✓	✗	✗	70.4	46.9	19.8	38.4	53.4	30.6	31.9	41.6	40.4	41.5
2	2	✓	✓	✗	✓	69.6	45.7	19.2	37.7	54.4	29.8	31.9	39.4	39.3	40.8
2	2	✓	✓	✓	✗	71.0	48.9	18.9	38.7	55.8	31.4	33.2	42.1	41.2	42.3
2	2	✓	✓	✓	✓	70.2	47.4	20.0	38.0	54.6	29.0	32.3	41.1	39.8	41.4
3	3	✓	✓	✓	✓	71.1	48.6	20.0	34.2	53.4	29.0	31.7	38.1	41.6	40.8
4	4	✓	✓	✓	✓	71.8	49.9	19.5	39.5	55.8	31.8	33.4	39.8	43.2	42.7
Few-shot (5-shot, 5-way)															
1	1	✗	✗	✗	✗	46.3	57.8	**39.8**	63.5	85.6	48.5	84.2	75.4	53.4	61.6
1	2	✗	✗	✗	✗	47.8	59.5	39.1	60.9	85.1	47.0	83.8	75.3	51.2	61.1
1	3	✗	✗	✗	✗	48.2	59.2	39.1	61.7	85.8	47.3	83.8	75.4	51.5	61.3
1	4	✗	✗	✗	✗	47.7	**59.7**	39.3	61.7	85.9	47.5	84.3	75.4	52.4	61.5
2	1	✗	✗	✗	✗	46.2	58.1	38.8	65.1	**87.0**	48.9	86.9	77.4	53.4	62.4
2	2	✗	✗	✗	✗	48.0	59.3	37.7	61.4	84.9	47.1	85.1	74.6	52.1	61.1
3	1	✗	✗	✗	✗	47.0	58.5	38.9	65.4	86.0	**49.2**	87.4	78.1	52.1	62.7
3	3	✗	✗	✗	✗	47.9	59.1	37.4	62.5	84.1	46.3	85.7	76.0	51.3	61.1
4	1	✗	✗	✗	✗	47.0	58.3	38.6	**65.7**	86.0	48.7	**88.0**	**78.7**	**54.5**	**62.8**
4	4	✗	✗	✗	✗	**49.5**	**59.7**	38.1	62.9	84.3	47.7	86.7	77.0	52.2	62.0
2	1	✓	✗	✗	✗	43.9	53.9	36.3	57.7	78.3	42.7	84.5	71.6	47.1	57.3
2	2	✓	✗	✗	✗	43.0	54.0	35.5	54.8	75.9	40.6	82.8	69.6	45.1	55.7
3	1	✓	✗	✗	✗	42.9	53.0	36.0	57.8	78.1	42.7	85.3	72.1	48.2	57.3
3	3	✓	✗	✗	✗	43.1	52.6	34.8	54.2	76.0	39.1	82.4	69.5	45.1	55.2
4	1	✓	✗	✗	✗	43.2	54.2	34.5	58.6	78.3	42.5	86.0	73.2	48.8	57.7
4	4	✓	✗	✗	✗	44.8	53.9	34.8	55.3	77.5	40.3	84.3	70.0	46.8	56.4
2	2	✓	✓	✗	✗	42.4	53.9	34.8	54.6	76.8	41.5	82.7	70.0	44.9	55.7
2	2	✓	✓	✗	✓	43.2	54.5	34.1	54.3	76.7	41.6	83.0	69.7	45.1	55.8
2	2	✓	✓	✓	✗	43.4	55.3	34.9	55.5	78.0	42.5	84.0	70.7	46.4	56.8
2	2	✓	✓	✓	✓	42.2	54.0	34.8	54.5	77.0	41.2	83.1	69.7	45.5	55.8
3	3	✓	✓	✓	✓	44.3	54.5	34.5	53.9	76.4	39.0	82.3	69.3	45.3	55.5
4	4	✓	✓	✓	✓	44.8	55.2	35.8	54.7	77.8	40.1	83.9	70.0	46.4	56.5

Table 7: **Classification performance on downstream vision datasets.** Top-1 accuracy for Linear Probing and 5-shot 5-way classification on the datasets of Section 4. Upper block of each panel: all ensemble (m) and density (n^+) configurations. Lower block: hard negative ablation over the sampling strategy (Smpl), the hard negative loss $\mathcal{L}_{HN}$ (Loss) and the curriculum scheduler (Sched). Bold: best in column; underlined: best within the hard negative block. Multi-generator diversity benefits are most pronounced on SUN397 and Oxford Flowers, which require fine-grained intra-class discrimination; CIFAR-10 is largely insensitive to ensemble size, consistent with its coarse-grained class structure. Frozen vision-only probing slightly favours constant, unscheduled exposure: the unscheduled run leads on linear probing (42.3 vs. 41.4), few-shot (56.8 vs. 55.8) and 8 of the 9 individual benchmarks. The drop from the reference row (49.1) is driven by $\mathcal{L}_{I2I}$, not by the scheduler (n^+=m=2 without and with $\mathcal{L}_{I2I}$: 47.9 $\rightarrow$ 41.5); the scheduler buys retrieval, zero-shot transfer and training stability (Table 8).

is generator-specific saturation under uniform allocation: low-diversity Recognizability Experts draw from a narrow manifold whose marginal visual support decays quickly with sample count, so a fixed uniform mixture becomes increasingly wasteful as the budget grows. The decisive test is per-generator saturation curves and allocation weighted by marginal diversity contribution (Section 5).

			HN Comp.			Image Retrieval				Text Retrieval				0-shot
n^+	m	$\mathcal{L}_{I2I}$	Smpl	Loss	Sched	COCO	F8K	F30K	**Avg**	COCO	F8K	F30K	**Avg**	ImgNet
1	1	✗	✗	✗	✗	6.4	15.0	13.3	11.6	8.7	21.5	17.5	15.9	8.15
1	2	✗	✗	✗	✗	7.1	15.6	13.8	12.1	9.8	22.1	20.6	17.5	8.53
1	3	✗	✗	✗	✗	7.4	17.1	15.3	13.3	9.4	24.1	20.2	17.9	8.56
1	4	✗	✗	✗	✗	7.2	17.0	14.3	12.8	9.4	23.3	20.2	17.6	8.92
2	1	✗	✗	✗	✗	7.5	17.2	16.8	13.8	9.5	23.8	22.9	18.7	9.41
2	2	✗	✗	✗	✗	7.6	17.7	17.2	14.2	10.4	25.8	24.3	20.2	9.54
3	1	✗	✗	✗	✗	8.9	17.6	17.5	14.7	**12.1**	26.6	24.7	21.1	9.66
3	3	✗	✗	✗	✗	8.9	**20.6**	19.1	16.2	**12.1**	**29.7**	25.4	**22.4**	9.38
4	1	✗	✗	✗	✗	8.1	17.9	18.4	14.8	10.5	23.0	24.7	19.4	9.56
4	4	✗	✗	✗	✗	**9.0**	19.9	19.4	16.1	11.4	27.1	25.8	21.4	9.86
2	1	✓	✗	✗	✗	7.3	16.2	16.4	13.3	9.1	21.6	22.2	17.6	10.14
2	2	✓	✗	✗	✗	8.2	17.4	17.7	14.4	11.1	24.5	23.7	19.8	11.19
3	1	✓	✗	✗	✗	7.9	17.5	17.3	14.2	10.0	22.0	24.7	18.9	10.88
3	3	✓	✗	✗	✗	**9.0**	20.2	**20.3**	**16.5**	11.5	26.4	**27.0**	21.7	10.98
4	1	✓	✗	✗	✗	8.0	18.6	18.3	15.0	10.4	24.4	24.8	19.9	11.24
4	4	✓	✗	✗	✗	8.8	18.9	19.4	15.7	11.1	24.5	26.2	20.6	**11.34**
2	2	✓	✓	✗	✗	6.7	16.7	16.1	13.2	8.9	23.3	21.5	17.9	9.95
2	2	✓	✓	✗	✓	7.5	16.4	16.9	13.6	10.4	23.0	22.9	18.8	10.23
2	2	✓	✓	✓	✗	7.8	16.9	17.4	14.0	9.8	22.8	24.4	19.0	10.27
2	2	✓	✓	✓	✓	7.7	18.3	17.8	14.6	10.5	26.2	24.9	20.5	10.80
3	3	✓	✓	✓	✓	8.7	19.5	18.8	15.7	10.1	22.9	23.6	18.9	10.72
4	4	✓	✓	✓	✓	8.8	18.9	19.4	15.7	11.5	24.9	24.7	20.4	11.01

Table 8: **Cross-modal retrieval and zero-shot classification.** Recall@5 for Image/Text Retrieval on MS COCO and Flickr; Top-5 accuracy for ImageNet zero-shot classification. Upper block: all ensemble and density configurations; lower block: hard negative ablation, columns as in Table 7. Bold: best in column; underlined: best within the hard negative block. Flickr30k benefits more consistently from ensemble diversity than MS COCO, suggesting that visual diversity interacts more strongly with retrieval benchmarks spanning a wider range of scene types. Among hard negative configurations, the combination of $\mathcal{L}_{HN}$ and curriculum yields the largest gains on Flickr8k text retrieval, the benchmark most sensitive to fine-grained semantic discrimination between visually similar scenes.

		Settings			SugarCrepe								SugarCrepe++ (ITT)					
					Replace			Swap		Add			Replace			Swap		
n^+	m	HN	$\mathcal{L}_{HN}$	S	Att	Obj	Rel	Att	Obj	Att	Obj	**Avg**	Att	Obj	Rel	Att	Obj	**Avg**
1	1	✗	✗	✗	60.0	64.2	53.2	51.2	48.2	59.0	52.0	55.4	42.1	46.7	33.2	31.5	29.8	36.7
2	2	✓	✗	✗	64.6	60.6	53.7	52.9	**57.1**	58.7	50.9	56.9	49.0	47.6	35.9	33.0	34.3	40.0
2	2	✓	✗	✓	64.9	61.7	**54.6**	49.1	46.9	56.5	51.5	55.0	48.9	49.2	36.7	31.2	28.2	38.8
2	2	✓	✓	✗	63.1	61.4	51.4	52.6	49.8	61.3	50.8	55.8	49.8	47.3	33.9	32.1	32.7	39.2
2	2	✓	✓	✓	64.6	61.4	52.4	49.0	53.5	**61.9**	51.8	56.4	**50.6**	47.6	35.5	31.1	**37.6**	40.5
3	3	✓	✓	✓	**65.0**	**64.9**	53.6	**54.7**	**57.1**	58.2	**52.9**	**58.1**	50.3	**52.9**	36.0	**36.0**	32.7	**41.6**
4	4	✓	✓	✓	63.8	64.2	52.4	52.1	51.0	**61.9**	51.9	56.8	**50.6**	51.7	**37.0**	33.3	31.0	40.7

Table 9: **Fine-grained compositionality on SugarCrepe and SugarCrepe++.** Accuracy across Replace, Swap, and Add error types. $m = 3$ outperforms $m = 4$ on Replace and Swap subtasks, which require precise attribute-level discrimination, and on the SugarCrepe average (58.1 vs. 56.8); $m = 4$ recovers only on *attribute* addition (61.9 vs. 58.2), not *object* addition (51.9 vs. 52.9). This supports the variance saturation interpretation: architectural diversity beyond $m = 3$ introduces stylistic noise that interferes with fine-grained syntactic binding, leaving a residual benefit only at the attribute level.

K Qualitative Analysis

This section provides visual evidence for the Generator-Invariance Hypothesis through three complementary perspectives: the conceptual structure of the generator ensemble, cross-generator rendering of controlled semantic perturbations, and failure modes that motivate the ensemble design.

Generator-Invariance Hypothesis: conceptual illustration. Figure 10 depicts the distributional relationship between individual generators and the real data manifold. Diversity Experts (SD 1.5, SD 2) occupy high-variance, lower-fidelity regions, while Recognizability Experts (SDXL-Turbo, SANA) occupy high-fidelity, lower-variance regions. PolyGen trains on the union of these disjoint

Setting	0-shot @ 3M	0-shot @ 12M
Single-generator (SD 1.5)	26.22	41.00
PolyGen ($m = 4$)	26.47	41.64
Δ	+0.25	+0.64

Table 10: Preliminary scaling results: ImageNet top-5 zero-shot accuracy at 3M and 12M training images. The best PolyGen configuration identified at 500k ($m = 4$, multi-positive) is transferred without scale-specific tuning, and compared against a matched single-generator baseline (SD 1.5). PolyGen retains a positive margin at both scales. These experiments are exploratory; a definitive scaling-law characterization requires per-scale tuning and broader downstream evaluation, and is left to future work.

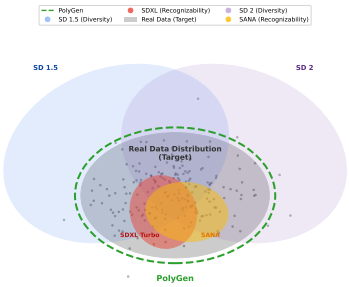


Figure 10: Conceptual illustration of the Generator-Invariance Hypothesis. Individual generators act as domain-specific experts with complementary biases. PolyGen (green dashed) trains on the union of their manifolds, approximating real-data coverage (grey); no single generator's artifacts are reinforced by the union.



Figure 11: Hard negative pairs rendered across the four-generator ensemble. *Top*: concept-level substitution ("oreo" $\rightarrow$ "chocolate chip cookie", $a = \textit{Concept}$). *Bottom*: viewpoint transformation ("bird's eye view" $\rightarrow$ "sidewalk view", $a = \textit{Perspective}$). The only consistent signal across generators is the semantic modification, illustrating the training objective of generator-invariant representation learning.

manifolds, approximating real-data coverage because the generator-specific artifacts occupy distinct spectral regions (Appendix E) and no one of them is reinforced by the union; Appendix L shows that making the encoder invariant to them is the separate job of $\mathcal{L}_{I2I}$.

Controlled semantic perturbations across generators. Figure 11 shows two hard negative pairs rendered through all four generators. Each row pairs a base caption with a counterfactual modifying a single semantic axis (*Concept* and *Perspective*, respectively). Diversity Experts produce high intra-concept variance at the cost of photorealism; Recognizability Experts produce photorealistic outputs at the cost of stylistic variation. The architectural heterogeneity ensures that the only stable signal across columns is the semantic modification, which is the condition for generator-invariant representation learning.

Successful and unsuccessful hard negative renderings. Figures 12 and 13, placed on the final two pages, show cases where hard negatives are correctly and incorrectly rendered across the ensemble. Correct renderings exhibit a clear and consistent semantic shift across all four generators,

providing unambiguous contrastive supervision. In failure cases, individual generators produce ambiguous or semantically inconsistent outputs, for example a color modification that is not visually salient, or a perspective change absorbed into stylistic variation. These failures motivate two design choices: the TripletCLIP objective (Appendix B.2), which excludes noisy image-to-hard-negative comparisons from the loss, and the ensemble approach, which marginalizes individual-generator errors by distributing the learning signal across multiple sources.

L GENERATOR-IDENTIFICATION PROBE

The Generator-Invariance Hypothesis asserts that generator-specific artifacts vary across sources in aggregate while semantic content is consistent by construction. This appendix tests directly how much information about *which* generator produced an image survives in a trained encoder's features, and isolates which part of the framework is responsible for removing it.

Protocol. We render 10k held-out captions, disjoint from every training caption set, through all four generators. The design is therefore caption-balanced: each caption appears exactly once per generator, so semantic content carries zero label information by construction. We freeze the trained image encoder, embed all 40k images, and fit a 4-way linear classifier predicting the source generator from the embedding, with caption-level train/test splits so that no caption appears on both sides. Chance accuracy is 25%. We report the mean and standard deviation over 3 splits, the probe margin (the mean logit gap between the correct generator and the strongest competitor), and the mean cosine similarity between images produced by the same generator from *different* captions, which measures how much a generator's style alone pulls embeddings together.

Encoder	Probe acc. (%)	Probe margin	cos(same gen., diff. caption)
Random-init ViT-B/16	62.56 $\pm$0.37	0.448	0.791
PolyGen m=1, n^+=1	81.26 $\pm$0.35	1.923	0.095
PolyGen m=4, n^+=4	88.47 $\pm$0.14	3.687	0.124
PolyGen m=4, n^+=4, $+\mathcal{L}_{I2I}$	68.33 $\pm$0.77	0.456	0.0004
LAION-2B CLIP (real-data control)	87.96 $\pm$0.19	3.298	0.323

Table 11: **Generator-identification probe.** A frozen encoder's features are used to predict which of the four generators produced each image (chance 25%). Decodability of generator identity tracks general feature quality rather than training source; only $\mathcal{L}_{I2I}$ suppresses it.

Decodability is not by itself a test of the hypothesis. Probe accuracy does not fall from $m = 1$ to $m = 4$; it rises, from 81.26% to 88.47%. The real-data control explains why: a LAION-2B CLIP, which never saw any of these generators during training, reaches 87.96%, statistically indistinguishable from our $m = 4$ model. Linear decodability of generator identity therefore tracks general feature quality, not the composition of the training source, and a naive reading of this probe as a direct test of ensembling would be wrong. We report this explicitly because it corrects an overreaching formulation: the hypothesis concerns which signal a model is *encouraged to encode*, not whether nuisance information is absent from a sufficiently expressive representation.

What the probe does isolate. The clean comparison is between two encoders trained on identical data with identical m, differing only in whether $\mathcal{L}_{I2I}$ is active. Enabling it drops probe accuracy from 88.47% to 68.33% and collapses the probe margin from 3.687 to 0.456, i.e. to the random-initialization level (0.448). The similarity structure shows the mechanism directly: after $\mathcal{L}_{I2I}$, two images from the same generator rendered from different captions are essentially orthogonal ($\cos$: $0.124 \rightarrow 0.0004$), while cross-generator agreement for the same caption is preserved and slightly strengthened ($0.579 \rightarrow 0.599$). Cross-family confusions that were negligible ($\sim$0.7% between SANA and SDXL-Turbo) rise to 17–20%. Generator style is thus no longer a direction the encoder separates confidently, though it remains decodable well above chance, while semantic agreement across sources is retained; this is also why enabling the loss costs linear probing ($48.9 \rightarrow 42.9$) and buys zero-shot accuracy ($9.86 \rightarrow 11.34$).

The invariance mechanism the framework claims is therefore measurable, but it is attributable to the invariance objective rather than to ensembling per se: heterogeneity supplies the cross-source variation that $\mathcal{L}_{I2I}$ converts into invariance, and the framework exposes both operating points.

M SEEDED COMPARISON OF THE HETEROGENEITY AND DENSITY AXES

The pivotal comparisons in Section 4.2 are between nearby configurations, so we re-trained the four corner cells of the $m \times n^+$ square with three seeds each, leaving the protocol otherwise unchanged; the runs reported in Table 1 count as seed 1. Seeding all four cells rather than a single contrast measures each axis at both levels of the other, which turns out to matter.

Configuration	*0-shot*	*Img Ret.*	*Text Ret.*	*L.P.*	*Few shot*
m=1, n^+=1 (SynthCLIP)	7.98 ±0.23	11.49 ±0.08	15.71 ±0.42	48.77 ±0.31	61.63 ±0.07
m=4, n^+=1 (heterogeneity only)	8.90 ±0.04	12.76 ±0.27	17.74 ±0.52	48.75 ±0.21	61.71 ±0.17
m=1, n^+=4 (density only)	9.55 ±0.01	14.49 ±0.25	19.60 ±0.28	**50.04** ±0.19	**62.79** ±0.04
m=4, n^+=4 (both)	9.97 ±0.16	**16.19** ±0.17	**22.06** ±0.84	48.92 ±0.13	61.98 ±0.12
m=4, n^+=4, $+\mathcal{L}_{I2I}$	11.34	15.69	20.61	42.88	56.40

Table 12: **Seeded corner comparison.** The four $m \times n^+$ corner configurations, three seeds each (mean ± std); the runs of Table 1 count as seed 1. The $\mathcal{L}_{I2I}$ row is a single run, shown for reference. Retrieval is R@5; zero-shot is ImageNet top-5.

The heterogeneity effect. At fixed $n^+ = 4$, heterogeneity is worth $+1.70$ image R@5, $+2.46$ text R@5 and $+0.42$ top-5 zero-shot. Taking the largest per-cell standard deviation within each metric as the noise estimate, the image-retrieval effect is roughly $7\times$ that noise and the zero-shot effect lies outside it. Text retrieval is the noisiest measurement in the table (± 0.84, with one seed of the $(m{=}4, n^+{=}4)$ cell sitting 1.8 points above the other two), so we rest the claim on image R@5 and report text retrieval as consistent in sign and magnitude but not independently decisive at three seeds.

The cost is a high-density effect. Measuring both levels of n^+ changes what the vision-only cost of heterogeneity is. At $n^+ = 4$ it trades -1.12 linear-probe and -0.81 few-shot points for the gains above, both more tightly resolved (roughly $8\times$ and $11\times$ the per-cell std) than the zero-shot gain. At $n^+ = 1$ the same axis costs nothing measurable: -0.02 linear probe and $+0.08$ few-shot, against per-cell standard deviations of ± 0.21–0.31 and ± 0.07–0.17, i.e. indistinguishable from zero in either direction. The vision-only cost is therefore a property of the high-density regime rather than of m, and we would have mischaracterized it had we seeded only the $n^+ = 4$ contrast.

The reading we consider correct is that heterogeneity and density compete for the same capacity. At $n^+ = 1$ heterogeneity is close to free ($+0.92$ zero-shot, $+1.27/+2.03$ R@5, no probe cost), while on top of $n^+ = 4$ it buys less ($+0.42$ zero-shot) and starts charging. This is consistent across the full grid: zero-shot and retrieval improve monotonically with m within every n^+ block of Table 1, and all 15 generator subsets of Figure 2 order consistently. It also confirms that the combined configuration is linear-probe-neutral against the baseline ($+0.16$, not significant at three seeds) with few-shot barely moving ($+0.35$); its real gains are in retrieval and zero-shot.

N CURRICULUM SCHEDULE FAMILY

As defined in Section 3.3, p is bounded by 0.5 by construction, so the free parameters of a schedule are the endpoint below saturation and the ramp speed. Table 13 reports four configurations spanning the feasible range, all at $m{=}2, n^+{=}2$ with the full recipe and otherwise identical settings, ordered by cumulative hard-negative exposure (the sum of p over training, in epoch-units).

Schedule	Exposure	*0-shot*	*Img Ret.*	*Text Ret.*	Sugar Crepe	Sugar Crepe++
$0 \to 0.25$ linear (conservative)	4.9	**11.24**	14.13	19.47	**59.10**	**42.54**
$0 \to 0.5$ linear (Table 2)	9.8	10.80	**14.6**	**20.5**	56.40	40.50
$0 \to 0.5$ by epoch 20, held	14.8	10.23	13.81	18.88	56.63	38.93
Constant 0.5 (no scheduler)	20.0	10.27	14.0	19.0	55.80	39.20

Table 13: **The four-point schedule family**, at $m{=}2, n^+{=}2$ with the full recipe, ordered by cumulative hard-negative exposure. General transfer is a dose effect; compositional reasoning is not.

General transfer is a dose effect. Zero-shot accuracy falls monotonically with cumulative exposure and then saturates ($11.24 \to 10.80 \to 10.23 \to 10.27$), the last two being effectively tied, as expected since both spend most of training at $p = 0.5$.

Compositional reasoning is not. The conservative schedule leads on both compositional benchmarks, by $+2.70$ SugarCrepe and $+2.04$ SugarCrepe++ over the linear ramp, while the remaining three points lie within 0.83 and 1.57 of each other with no ordering by dose. The least-exposed schedule is the best on both benchmarks, so compositional performance is not governed by cumulative exposure and naive injection is not the compositional ceiling that maximal exposure would suggest. This strengthens the curriculum claim: the timing of hard-negative introduction governs the outcome across a family of schedules rather than at a single setting.

The endpoint is a trade-off, not a free improvement. The conservative schedule is not uniformly better. On retrieval the linear ramp of Table 2 remains the best of the four ($14.6/20.5$ against $14.13/19.47$), and on vision-only probes the no-scheduler run leads (linear probe 42.3, few-shot 56.8, against $41.84/55.88$ conservative and $41.4/55.8$ linear). The endpoint trades retrieval and probe accuracy against zero-shot and compositional reasoning.

The fast-ramp run peaks at epoch 27 and declines by roughly 0.6 points by epoch 40, converging onto the constant-p run: prolonged saturated exposure mildly degrades the model, the failure mode the curriculum is designed to avoid (Appendix P).

O TEXT-ONLY HARD NEGATIVES

Prior work obtains attribute-object binding from text-side negatives alone, which raises the question of whether the rendered image negatives I^- are necessary at all. We ablate them at $m{=}2, n^+{=}2$ with the caption negatives T^- retained and no scheduler, so that the comparison is entirely scheduler-free. Once image negatives are removed, the TripletCLIP objective reduces exactly to the NegCLIP objective of Yuksekgonul et al. (2022).

A single ablation cannot separate negative type from budget, because dropping I^- halves the number of images per epoch. We therefore run two configurations. *Run 1 (matched text, halved images)* drops the I^- subset and is otherwise identical to the naive hard-negative row: 125k captions T^+, 250k images I^+, 125k T^-. *Run 2 (matched images, doubled text)* doubles T^+ to 250k, giving 500k I^+ and 250k T^-. Run 2 matches the reference row on images per epoch (499,200), steps per epoch (325), global batch size and epoch count; its caption pool is twice as large, which favours Run 2, so the deficits below are a lower bound.

Run	*0-shot*	*Img Ret.*	*Text Ret.*	Sugar Crepe	Sugar Crepe++
$T^- + I^-$, no scheduler (Table 2)	**10.27**	**14.0**	**19.0**	**55.80**	**39.20**
T^- only, half image budget (Run 1)	6.00	8.88	10.06	52.27	29.11
T^- only, matched image budget (Run 2)	7.27	12.64	14.29	54.64	29.77

Table 14: **Removing rendered image negatives.** Run 1 and Run 2 differ only in budget, which separates the cost of the negative type from the cost of the halved image count.

The gap between the reference row and Run 1 is 4.27 zero-shot points, and the pair splits it cleanly: 1.27 points are image budget (Run 1 $\rightarrow$ Run 2) and 3.00 points are negative type (Run 2 $\rightarrow$ reference row). The half-budget run should therefore not be quoted on its own. At matched budget, removing the rendered image negatives costs 3.00 zero-shot, $1.36/4.71$ retrieval R@5, 1.16 SugarCrepe and 9.43 SugarCrepe++. The asymmetry is the result: doubling the image budget buys back most of the retrieval and zero-shot deficit ($+3.76$ image R@5) but almost none of the compositional one ($+0.66$ SugarCrepe++).

Two further observations. First, text-only negatives are not inert. Against the synthetic no-hard-negative baseline they improve text-text discrimination (SugarCrepe++ TTT $+3.68$) while degrading image-text (-6.90). They do train attribute binding on the text side, exactly as prior work predicts; the signal simply does not cross into image-text alignment without a rendered image negative to anchor it. Second, the representation-quality cost is small but real. Run 2 reaches linear probe 41.77 and few-shot 56.05; the matched control is the unscheduled row of Table 7 ($42.3/56.8$), since this ablation is scheduler-free throughout, and Run 2 is below it on both. Comparing instead against the *scheduled* row ($41.37/55.79$) would reverse the sign, which is why we quote the matched one. Rendered image negatives therefore buy multimodal alignment above all, but not at the expense of visual features: omitting them also costs a little frozen-feature quality.

P Training Dynamics and Model Selection

All configurations in this paper are evaluated at a fixed final checkpoint (epoch 40) under an identical protocol, with no per-configuration checkpoint selection anywhere. We audit this choice against the full per-epoch ImageNet zero-shot logs of all 26 configurations in Tables 1 and 2. Configurations do peak at different epochs, but by at most 0.16 top-5 points relative to their own reported last-epoch result, and the *ordering* of configurations, on which every comparison in this paper depends, is settled well before training ends: ranking the 16 configurations of Table 1 by zero-shot accuracy at each epoch, the Spearman correlation with the final ranking is already 0.990 at epoch 30, with only 2 of the 120 configuration pairs discordant, both among the three weakest and near-tied configurations.

The pivotal comparisons are likewise stable epoch by epoch. The heterogeneity contrast at $n^+ = 4$ without $\mathcal{L}_{I2I}$ holds the same sign over the last 9 consecutive epochs ($+0.239 \pm 0.139$ across the final 10, $+0.300$ at epoch 40), and the full recipe stays above naive hard-negative injection at every one of the 40 epochs ($+0.924 \pm 0.087$ across the final 10). The within-run variability floor, the standard deviation of zero-shot accuracy over the final 5 epochs, is 0.030 median and 0.104 maximum, an order of magnitude below the pivotal effects; it is a lower bound on run-to-run variance, which the seeded runs of Table 12 measure directly.

Two runs are genuine exceptions: the fast-ramp curriculum of Appendix N peaks at epoch 27 and is 0.57 points lower by epoch 40, and the matched-budget text-only run of Appendix O peaks at epoch 15 and is 1.78 points lower by epoch 40. In both cases the late decline is itself the finding, caused by prolonged saturated hard-negative exposure and by text-only-negative saturation respectively, rather than an artifact of when we evaluated.

Q Placement of Modern Generators

The four generators instantiating PolyGen were released between 2022 and 2024, which raises the question of whether the Recognizability/Diversity distinction survives in more recent model families: if modern generators were uniformly recognizable, the best strategy would be to pick the single most diverse modern model. The available evidence points the other way. Adamkiewicz et al. (2026) benchmark thirteen open-weight text-to-image models released between 2022 and 2025, including FLUX, SD 3.5, SANA and Qwen-Image, specifically as training-data generators, and find that synthetic-to-real transfer is inversely correlated with release date, with SD 1.5 the strongest generator in their pool and the deficit traced to texture impoverishment, high-frequency mismatch, and density–coverage collapse. Newer models are therefore moving further into the high-density, low-coverage corner, the low-diversity regime in our terms, and a single modern generator is precisely the configuration this evidence penalizes. Detailed captions of the kind our pipeline produces partially mitigate the regression, but the texture and high-frequency deficits persist regardless of prompt detail; these are exactly the generator-specific artifacts that the multi-generator design spreads across sources instead of reinforcing (Appendix E).

Our own measurements place FLUX.1-Schnell and SD 3.5 Large Turbo on the Recognizability axis directly, through their prompt adherence (Table 4, where FLUX attains the highest TIFA score in the pool) and their hard-negative rendering ability (Table 5). We do not infer their Diversity from FID, which conflates fidelity with mode coverage, and rely instead on the density–coverage measurements of Adamkiewicz et al. (2026) over these exact models. FLUX- and SD3-class models would therefore enter the framework as additional Recognizability Experts rather than as replacements for the Diversity Experts; substituting them for the latter would yield a recognizability-only ensemble, which the subset analysis of Figure 2 shows to underperform mixed compositions.

Two limits apply. First, our Diversity metric has not been computed for these models: it relies on cluster statistics at the 500k scale ($K = 1000$ clusters, roughly 500 images each), and generating datasets of that size with these generators is beyond our compute budget; a small-scale proxy on the 5k caption set of Appendix C.3 would give only a noisier relative placement, not comparable to Figure 2. Second, the framework assigns roles by measured (R, D) position rather than by model identity or release date, so its validity does not rest on the generators used to instantiate it: a future generator scoring high on both axes would be an ideal ensemble member, and the measurement is exactly the admission test it would have to pass.



Figure 12: Representative examples in which the hard negative semantic modification is correctly rendered across the generator ensemble. Each row shows a base caption (top) and its counterfactual (bottom) rendered by SD 1.5, SD 2, SANA, and SDXL-Turbo. The semantic shift is visually unambiguous across all four generators, providing high-quality contrastive supervision.



Figure 13: Representative failure cases in which one or more generators produce ambiguous or semantically inconsistent hard negative renderings. Failures are concentrated in axes requiring precise low-level attribute rendering (e.g., *Color*, *Size*), consistent with the per-axis analysis in Appendix D (Figure 6). The multi-generator ensemble mitigates individual failures by distributing the contrastive signal across sources; the TripletCLIP objective further reduces the impact of visually ambiguous images by excluding $T^+ \to I^-$ and $T^- \to I^+$ comparisons from the loss.