Narrowing the Performance Gap in Synthetic VLM Pre-training via Multi-Generator Ensembles

Leonardo Brusini[1], Cristian Sbrolli[1], Eugenio Lomurno[1], Toshihiko Yamasaki[2], Matteo Matteucci[1]

[1]Politecnico di Milano, AIRLab
Via Ponzio 34/5, 20133 Milan, Italy
name.surname@polimi.it

[2]The University of Tokyo, Computer Vision and Media Lab
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
yamasaki@cvm.t.u-tokyo.ac.jp

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. In this work, we introduce ***PolyGen****, a framework that redefines synthetic data construction by prioritizing manifold coverage and compositional rigor over simple dataset size. PolyGen employs a Polylithic approach to train on the intersection of architecturally distinct generators, effectively marginalizing out model-specific artifacts. Additionally, we introduce a Programmatic Hard Negative curriculum that enforces fine-grained syntactic understanding. By structurally reallocating the same data budget from unique captions to multi-source variations, PolyGen achieves a more robust feature space, outperforming the leading single-source baseline (SynthCLIP) by +19.0% on aggregate multi-task benchmarks and on the SugarCrepe++ compositionality benchmark (+9.1%). These results demonstrate that structural diversity is a more data-efficient scaling law than simply increasing the volume of single-source samples.*

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 [26] and ALIGN [15] 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 precarious: massive, web-scale datasets comprising billions of image-text pairs.

This reliance on web-scraped data presents two existential challenges. First, recent work has documented concerning trends suggesting that the supply of high-quality, human-generated data may become a bottleneck for future scaling [35]. Second, "real-world" datasets are inherently noisy, rife with misalignment, societal biases, and privacy or copyright issues that require expensive curation. A compelling alternative is the transition to fully synthetic data [10, 33]. Given the fidelity of modern text-to-image (T2I) diffusion models, it is theoretically possible to generate infinitely scalable, safe, and controllable datasets on demand [8].

Despite this promise, a persistent "Synthetic Gap" remains: models trained exclusively on synthetic data consistently underperform their real-data counterparts, particularly in generalization tasks. Recent work by Fan et al. [8] has identified a fundamental cause: a trade-off between *diversity* and *recognizability* that is inherent to single-generator training. Moreover, empirical evidence suggests that even ensembles of identical models trained on the same data can outperform single-model approaches, revealing that individual generators are confined to local optima in the manifold space [19].

The manifestation of this trade-off varies by architecture. Older diffusion models (e.g., Stable Diffusion 1.5) produce high semantic variance but suffer from perceptual artifacts and frequency-domain signatures, while modern models (e.g., SDXL-Turbo) achieve higher photorealism at the expense of reduced output diversity. When trained on images from a single generator, VLMs inevitably learn to rely on these low-level, generator-specific cues, such as texture patterns or spectral fingerprints [3], rather than the abstract semantic features necessary for open-world generalization.

To address these challenges, we introduce **PolyGen** (Polylithic Generation Framework), an end-to-end pipeline for training vision-language models on fully synthetic data. Our approach is grounded in the *Generator-Invariance Hypothesis*: by training on images sampled from multiple architecturally distinct generators, we force the model to learn

representations that are invariant to generator-specific artifacts, effectively marginalizing out spurious correlations and isolating the semantic signal as the learning objective.

We adapt our approach to also generate multiple hard negative samples, addressing a separate 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 [43].

Finally, PolyGen operationalizes the Generator-Invariance hypothesis through three complementary mechanisms:

1. **Multi-Positive Ensemble:** We break the single-generator trade-off by generating n^+ diverse visualizations for each caption using a heterogeneous pool (SD 1.5, SD 2, SDXL-Turbo, SANA). By treating these architecturally distinct outputs as positive pairs through multi-positive objectives, we force the model to learn representations invariant to specific spectral artifacts but sensitive to the shared semantic content.
2. **Programmatic Hard Negatives:** We leverage LLMs to construct structured and controlled semantic perturbations (counterfactuals) for the base captions. Training on these hard negatives forces the model to abandon "bag-of-words" shortcuts in favor of fine-grained compositional understanding, distinguishing between different semantic and compositional bindings.
3. **Curriculum Scheduler:** We gradually increase the proportion of hard negatives during training, allowing the model to first establish coarse semantic structure before refining discriminative boundaries.

Empirically, PolyGen demonstrates that balancing synthetic data diversity and quality is crucial. By engineering the training distribution rather than merely scaling it, we achieve a **+19.0%** improvement in Delta Multi-Task Learning over strong single-source baselines, narrowing the gap to real-data performance and suggesting a viable path toward sustainable large-scale vision-language training.

2. Related Work

Hybrid Synthetic Strategies. A dominant line of research mitigates the data scarcity problem by leveraging real-world captions to generate synthetic images. StableRep [32] demonstrated that multi-positive contrastive learning, treating multiple synthetic images generated from the same caption as positive pairs, can outperform models trained on real data. LaCLIP [7] and VeCLIP [18] use instead Large Language Models (LLMs) to rewrite captions, improving alignment and diversity while maintaining the original image-text grounding. However, these "hybrid" approaches remain tethered to the constraints of web-scraped data, inheriting its linguistic/visual biases, privacy risks, and finite scale.

Fully Synthetic Pipelines. Recent works have attempted to close the loop by generating both captions and images [33]. Fan et al. [8] proposed scaling laws for synthetic data, showing that synthetic pre-training can rival real-world baselines. A landmark study, SynthCLIP [10], introduced an end-to-end pipeline using LLM-generated captions and Stable Diffusion images, training a CLIP model from scratch without any human data. Recent work has further demonstrated that synthetic data can achieve parity with real datasets in image classification tasks while providing privacy guarantees [27], though this result remains limited to single-task supervised learning. While these approaches are pioneering, they suffer from *architectural homogeneity*: they typically rely on a single generative model, often relying on dated architectures (e.g., Stable Diffusion v1.5), causing the downstream model to overfit to generator-specific artifacts [3]. We posit that this lack of generator diversity 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.

Hard Negative Mining. A well-known limitation of contrastive vision-language models is their tendency to rely on lexical co-occurrence patterns rather than true compositional reasoning [43]. Benchmarks like ARO and SugarCrepe [13] systematically evaluate this failure mode through tasks requiring attribute binding, spatial reasoning, and object-relation understanding, revealing significant performance gaps even in state-of-the-art models. To enhance discriminative capability, Hard Negative Mining has evolved from importance weighting [25, 28], to hard negative generation with rule-based swapping [43] and LLM-driven rewriting [24]. However, existing methods either depend on external parsers or apply unconstrained LLM generation to unstructured text, risking semantic drift. While intrinsic data mining has been proposed [36], it remains computationally expensive and bound by the static data distribution. We propose a *constructive* alternative: by synthesizing structurally grounded base captions first, we can programmatically force the LLM to perturb specific semantic axes, ensuring safer, high-fidelity counterfactuals.

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 2): *Polylithic Synthesis* combines structured caption generation with multi-generator image rendering, followed by *Curriculum-Based Contrastive Training*. We detail each component below.

3.1. Structured Caption Pairs

The first component of *Polylithic Synthesis* (Stage 1, Figure 2) generates semantically controlled caption pairs. Prior work on synthetic data generation typically relies on "creative" prompting, generating 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 [16] to generate the base captions T^+, conditioning the 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 [40], 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 8) and an explicit *Concept* axis for "change of subject" negatives. The additional attribute indication also serves to increase variability when scaling generation leads to the use of the same concept multiple times. (full generation prompts reported in Appendix 6). T^+ is obtained after downsampling a broader pool of captions using MetaCLIP's balanced sampling algorithm (Appendix 11).

Hard Negative Generation. Thanks to the structured base captions, we can leverage LLMs to generate controlled negatives. We use Llama-3.1-8B [5] (ablated in Sec. 4.5) to generate the counterfactual hard negatives T^-. The model receives $t \in T^+$, the corresponding semantic axis a, and concept c, with strict instructions to only modify the specified axis while preserving syntactic structure. As illustrated in Figure 3, this strategy ensures global semantic coherence. Unlike rigid programmatic substitution, which risks generating logical absurdities by modifying terms in isolation, the LLM dynamically adjusts contextually dependent attributes to align with the new concept.

3.2. Multi-Generator Synthesis Ensemble

The second component of Stage 1 renders each caption pair through multiple diffusion models. A core challenge in synthetic training is to determine the optimal generator $\mathcal{G}$. Relying on a single model 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.

Metric-Driven Selection. To systematically evaluate candidate generators, we establish a quantitative framework for measuring two complementary properties of synthetic datasets in open-vocabulary contrastive settings. While Fan et al. [8] defined *Recognizability* and *Diversity* for supervised classification datasets (with fixed label sets), we adapt these metrics for the CLIP (self-supervised) paradigm:

- *Recognizability (Fidelity).* In the absence of class labels, we measure the alignment between the generated image x and its conditioning caption t. We define Recognizability as the expected CLIPScore [11] over the dataset $\mathcal{D}$:

$$R(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(x,t)\in\mathcal{D}} \text{CLIPScore}(f_{img}(x), f_{txt}(t)) \quad (1)$$

High recognizability indicates the generator reliably respects the semantic constraints of the prompt.

- *Diversity:* Measuring diversity in open-world data is challenging due to the high dimensionality of the visual space and the semantic relativity of what constitutes meaningful variation. We propose a cluster-based approach that captures intra-concept variance. We partition the dataset into $K = 1000$ semantic clusters based on caption embeddings, then define Diversity as the average standard deviation of image embeddings within each cluster k:

$$D(\mathcal{D}) = \frac{1}{K} \sum_{k=1}^{K} \sqrt{\frac{1}{|\mathcal{D}_k|} \sum_{x\in\mathcal{D}_k} ||f_{img}(x) - \mu_k||^2} \quad (2)$$

where μ_k is the centroid of image embeddings in cluster k. This metric specifically captures the ability of the generator to produce visually distinct interpretations of the same semantic concept, or group of concepts.

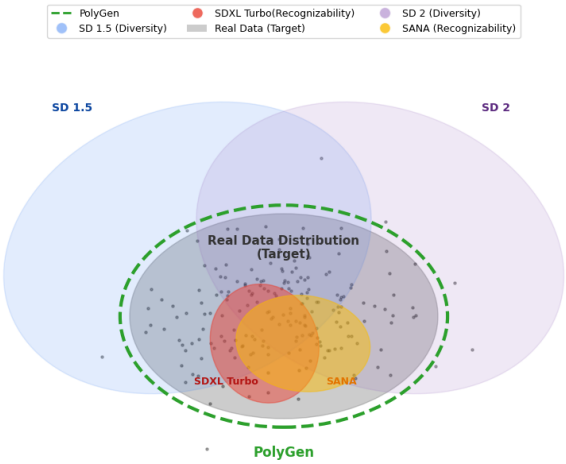


Figure 1. **Conceptual Illustration of the Generator-Invariance Hypothesis.** Individual generators (colored circles) act as "Experts" with specific biases: older models like SD 1.5 (Blue) offer high diversity but low fidelity, while modern distilled models like SDXLT (Red) offer high fidelity but suffer from mode collapse. PolyGen (Green Dashed) trains on the union of these manifolds, approximating the coverage of the Real Data Distribution (Grey) by marginalizing out generator-specific artifacts.

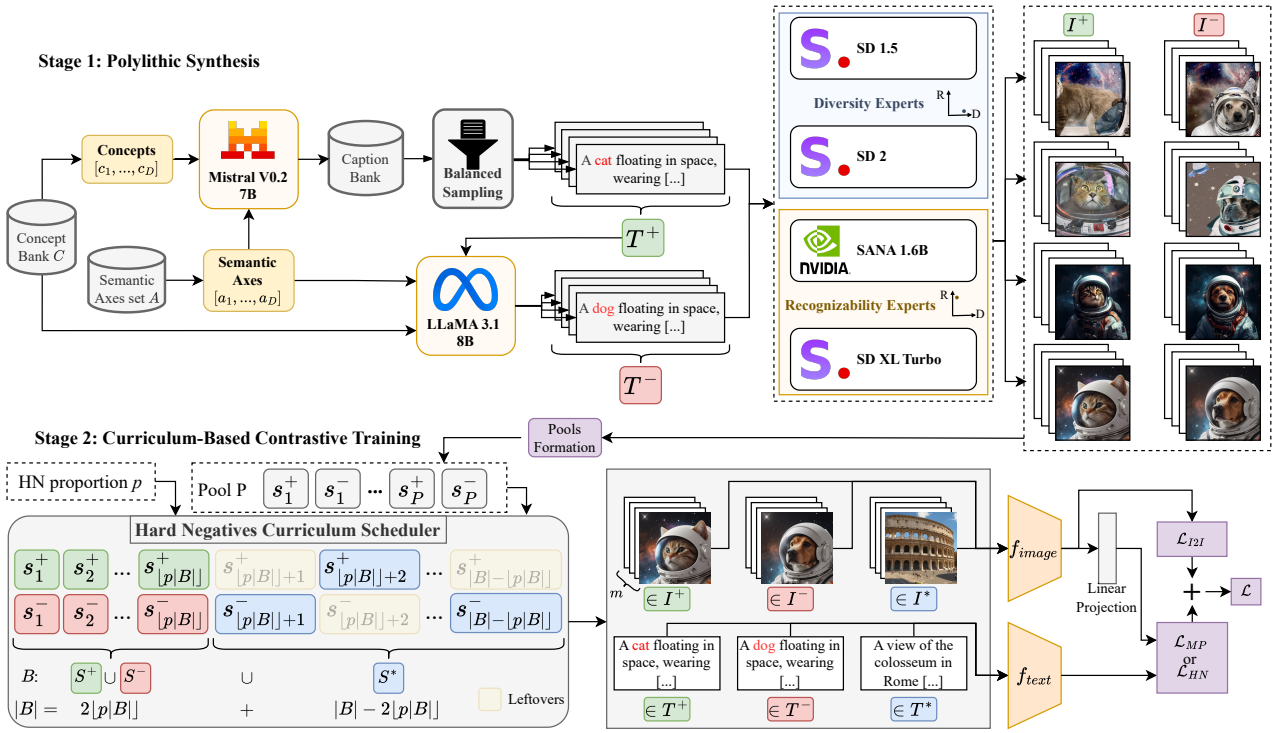


Figure 2. **PolyGen Pipeline Overview.** *Stage 1 (Polylithic Synthesis):* Concepts from MetaCLIP Bank and semantic axes feed into a dual-LLM system. Mistral V0.2 generates base captions T^+ conditioned on concept-attribute pairs, while Llama 3.1 produces hard negatives T^- by modifying specific semantic dimensions. Caption pairs are rendered through four diffusion models: Diversity Experts (SD 1.5, SD 2) ensure broad manifold coverage despite lower fidelity, while Recognizability Experts (SANA, SDXL-Turbo) provide high prompt adherence despite limited variance. *Stage 2 (Curriculum-Based Training):* A scheduler linearly increases hard negative proportion p from 0 to 0.5 during training. Unused samples populate a leftover queue, maintaining data efficiency. Batches combine multi-positive samples (I^+, T^+) and hard negatives (I^-, T^-) to compute $\mathcal{L}_{I2I}$ and $\mathcal{L}_{HN}$.

The Generator-Invariance Strategy. Guided by the *Generator-Invariance Hypothesis* (Fig. 1), we construct an ensemble $\mathcal{G}$ that maximizes coverage of the visual-semantic spectrum. We select four architecturally distinct models:

- *Recognizability Experts:* SDXL-Turbo [30] and SANA-1.6B [39], selected for high prompt adherence and photorealism despite lower output diversity.

- *Diversity Experts:* Stable Diffusion v1.5 and v2 [29], selected for high variance ensuring broad manifold coverage despite lower fidelity.

For every caption pair (t^+, t^-), we generate corresponding images using all four models. This forces the downstream model to learn representations invariant to generator-specific spectral signatures such as the compression artifacts in SD1.5 or the low-frequency bias in SDXL-Turbo (Appendix 9), as the only signal consistent across the ensemble is the semantic concept itself.

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 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 [32]. 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. Beyond text-to-image alignment, we evaluate the impact of explicit Image-to-Image regularization. We incorporate the contrastive term $\mathcal{L}_{I2I}$ to penalize the model for learning generator-specific idiosyncrasies. For any image anchor i, we define a ground-truth distribution p_i that is uniform $(1/(n^+ - 1))$ over the other images generated from the same caption. Given the predicted categorical distribution:

$$q_{i,j} = \frac{e^{(s(i,j)/\tau)}}{\sum_{k \in I \setminus \{i\}} e^{s(i,k)/\tau)}}, \tag{3}$$

the loss is computed as the cross-entropy:

$$\mathcal{L}_{I2I} = -\frac{1}{N} \sum_{i \in I} \left(\sum_{j \in I \setminus \{i\}} p_{i,j} \log q_{i,j} \right). \tag{4}$$

Figure 3. **Hard Negative Generation Across Multi-Generator Ensemble.** Two examples of controlled semantic perturbations rendered through four diffusion models. *Top:* Concept-level modification ("oreo" $\rightarrow$ "chocolate chip cookie", a = Concept). *Bottom:* Viewpoint transformation ("bird's eye view" $\rightarrow$ "sidewalk view", a = Perspective). Diversity Experts (SD 1.5, SD 2) produce high intra-concept variance with lower photorealism, while Recognizability Experts (SANA, SDXL-Turbo) generate photorealistic outputs with reduced stylistic variation. The architectural heterogeneity forces models to learn generator-invariant semantic representations, as the only consistent signal across columns is the underlying concept modification.

This term acts as a structural regularizer, forcing the visual encoder to discard "spurious" features such as spectral artifacts or stylistic biases, and move the representation toward the invariant semantic core defined by the prompt.

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 [24].

We define the Hard-Negative NCE loss for image-to-text as:

$$\mathcal{L}_{NCE}^{I^+ \rightarrow T^+;T^-} = -\frac{1}{N}\sum_{k=1}^{N} \log \frac{e^{s(i_k,t_i)/\tau}}{\sum_{t \in T^+ \cup T^-} e^{s(i_k,t)/\tau}}, \quad (5)$$

analogous for text-to-image. The NegCLIP loss is then:

$$\mathcal{L}_{NegCLIP}^{I^+,T^+,T^-} = \mathcal{L}_{NCE}^{I^+ \rightarrow T^+;T^-} + \mathcal{L}_{NCE}^{T^+ \rightarrow I^+}. \quad (6)$$

Finally, the total TripletCLIP loss is defined as:

$$\mathcal{L}_{TripletCLIP} = \mathcal{L}_{NegCLIP}^{I^+,T^+,T^-} + \mathcal{L}_{NegCLIP}^{I^-,T^-,T^+}. \quad (7)$$

This formulation forces the model to explicitly distinguish the base concept from its semantic counterfactual.

Hard Negatives Curriculum Scheduler. Naive injection of hard negatives causes early training instability because fine-grained verification tasks disrupt coarse-grained semantic alignment, preventing stable conceptual clusters. We introduce a curriculum scheduler linearly increasing the in-batch hard negative ratio p from $0 \rightarrow 0.5$ over the training epochs. To ensure full data utilization throughout the epoch, we implement a *Leftover Queue*: samples that are not paired with their corresponding hard negatives are queued and batched in subsequent steps. This scheduler, when used in combination with a decaying learning rate, allows the model to learn a more robust knowledge baseline first, and exploits hard negatives for learning fine-grained discriminations later in training. Since batch composition varies, we adapt $\mathcal{L}_{HN}$ to balance subset contributions:

$$\mathcal{L}_{HN} = \frac{1}{|B|}(|T|\mathcal{L}_{NegCLIP}^{I,T,T^-} + |T^-|\mathcal{L}_{NegCLIP}^{I^-,T^-,T}), \quad (8)$$

with $T = T^* \cup T^+$, $I = I^* \cup I^+$, and $|B| = |T| + |T^-|$.

4. Experiments and Results

4.1. Experimental Setup

Training Configuration. To ensure a rigorous comparison with prior state-of-the-art SynthCLIP [10], we conduct our

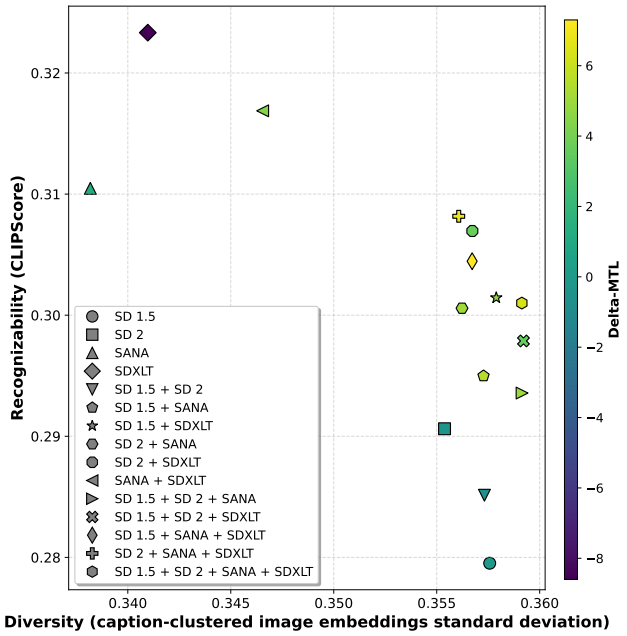


Figure 4. Recognizability vs. Diversity, with color indicating downstream performance. While individual models trade off diversity or recognizability, PolyGen ensembles break this Pareto frontier, maximizing both axes to achieve superior Δ_{MTL}.

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	1	✗	49.1	61.6	11.6	15.9	8.15	-	-
1	2	✗	47.5	61.1	12.1	17.5	8.53	+3.1%	+3.1%
1	3	✗	48.5	61.3	13.3	17.9	8.56	+6.1%	+6.1%
1	4	✗	48.9	61.5	12.8	17.6	8.92	+6.2%	+6.2%
2	1	✗	49.5	62.4	13.8	18.7	9.41	+11.0%	-
2	2	✗	47.9	61.1	14.2	20.2	9.54	+12.7%	+1.3%
3	1	✗	49.3	62.7	14.7	21.1	9.66	+16.1%	-
3	3	✗	48.2	61.1	16.2	**22.4**	9.38	+18.7%	+1.7%
4	1	✗	**49.8**	**62.8**	14.8	19.4	9.56	+14.1%	-
4	4	✗	48.9	62.0	16.1	21.4	9.86	**+19.0%**	+3.8%
2	1	✓	43.4	57.3	13.3	17.6	10.14	+6.3%	-
2	2	✓	41.5	55.7	14.4	19.8	11.19	+12.3%	+4.8%
3	1	✓	44.4	57.3	14.2	18.9	10.88	+11.8%	-
3	3	✓	41.3	55.2	**16.5**	21.7	10.98	+17.4%	+4.1%
4	1	✓	45.2	57.7	15.0	19.9	11.24	+15.6%	-
4	4	✓	42.9	56.4	15.7	20.6	**11.34**	+16.6%	+0.4 %

Table 1. Heterogeneity (m) vs. Density (n^+) averaged results. Jointly scaling m and n^+ maximizes Δ_{MTL}. Notably, strong invariance ($m = 4, \mathcal{L}_{I2I}$) improves abstract reasoning (Zero-shot) at the cost of texture-based classification (L.P. : Linear Probing).

primary evaluation on a controlled subset of 500k captions sampled from the CC3M dataset. We train a ViT-B/16 CLIP architecture from scratch for 40 epochs. We use the AdamW optimizer with a learning rate of 5×10^{-4} and a cosine decay schedule with 1 warmup epoch. For PolyGen, we use a batch size of 1536. To handle the Multi-Positive input fairly, we fix the number of training images and scale the number of captions accordingly: this corresponds to $500k/n^+$ captions in case of n^+ positives.
The Hard Negative ratio p follows our curriculum scheduler, linearly increasing from 0 to 0.5 over the training course.

Downstream Tasks We adopt SynthCLIP evaluation setting, evaluating our models in both vision-only tasks, meaning the quality of the features extracted by the image encoder (i.e., linear probing, few-shot classification), and vision-language tasks, with zero-shot retrieval and classification. Also following SynthCLIP, results are summarized into a unique Δ_{MTL} metric [34] (Appendix 12), aggregating top-1 accuracy for vision tasks, top-5 accuracy for zero-shot classification, and recall@5 for retrieval, representing the relative performance drop or improvement w.r.t. a baseline model, while we report full per-task results in Appendix 13.

Datasets. The models are evaluated on a comprehensive suite of downstream benchmarks to assess both vision-only and vision-language capabilities: for Linear Probing & Few-Shot we use: CIFAR10 and CIFAR100 [17], FGVC Aircraft [21], DTD (Textures) [2], Oxford Flowers [22], Oxford-IIIT Pets [23], SUN397 [38], Caltech-101 [9], Food-101 [1]. We also use ImageNet [4] for zero-shot classification (accuracy), Flickr8k and Flickr30k [42], and MS COCO [20] for text-to-image and image-to-text retrieval. Lastly, for hard negative tasks, we use SugarCrepe [13] and SugarCrepe++ [6]. Results in this section are reported per-task, as averages over the datasets. Per-dataset results are reported in Appendix 13.

4.2. Recognizability vs. Diversity implications

Figure 4 illustrates the relationship between the distributional properties of the training data (Recognizability vs. Diversity) and the downstream performance (indicated by color). To isolate the impact of manifold coverage, we train a series of CLIP models on fixed sets of 500k image-text pairs ($n^+ = 1$). For the ensemble runs ($m > 1$), the image for each caption is uniformly sampled from the pool of selected generators. This ensures that any performance gain is attributed to inter-source diversity rather than an increase in the number of training samples. The plot reveals a clear dichotomy in generative behaviors. Older architectures, specifically Stable Diffusion (SD) v1.5, occupy the high-diversity, lower-recognizability region. In contrast, modern distilled models like SDXL-Turbo and SANA-1.6B exhibit significantly higher Recognizability but at a noticeable cost to intra-concept Diversity. This phenomenon is representa-

tive of "mode collapse" in high-fidelity generators, where the model converges on a limited set of high-probability visual interpretations to ensure aesthetic appeal. Our strategy effectively moves the training distribution toward the upper-right quadrant, effectively improving performance on downstream tasks. This empirically validates our Generator-Invariance Hypothesis: while individual generators are limited by their specific spectral biases, the ensemble marginalizes these artifacts. By forcing the model to learn features consistent across disjoint manifolds, we achieve robust generalization (Δ_{MTL}) that neither type of generator could achieve in isolation.

4.3. The Heterogeneity-Density Trade-off

We isolate the impact of our two core design variables: the ensemble size (m) and the positive sample density (n^+). Table 1 presents the results, starting from $m = 1$ (SD 1.5), $m = 2$ including SDXLT, $m = 3$ also including SANA, and with $m = 4$ indicating the full PolyGen ensemble.

The Heterogeneity Gain ($m \uparrow$). Increasing the ensemble size from single-source ($m = 1$) to the full ensemble ($m = 4$) yields a universal improvement in Δ_{MTL} across all configurations. Critically, this gain is driven by Vision-Language tasks. For example, in the $n^+ = 1$ block, 0-shot accuracy rises from 8.15 ($m = 1$) to 8.92 ($m = 4$). This confirms the Generator-Invariance Hypothesis: seeing the same concept rendered by diverse engines forces the model to decouple semantics from generator-specific artifacts, resulting in more robust open-world representations.

The Density Gain ($n^+ \uparrow$). While diversity optimizes generalization, density optimizes robustness. Increasing the number of positive pairs (n^+) improves performance on both axes (Vision-Only and Vision-Language). By clustering multiple images around a single caption, the model learns a stable "centroid" representation. The combination of both factors is multiplicative: the full PolyGen configuration ($n^+ = 4, m = 4$) reaches a peak Δ_{MTL} of +19.0%, significantly outperforming methods that rely on density alone ($n^+ = 4, m = 1$, +14.1%).

Invariance as a Regularizer. A revealing pattern emerges when analyzing the impact of explicit invariance mechanisms: the Image-to-Image loss ($\mathcal{L}_{I\to I}$) and the Ensemble size (m). While increasing m or adding $\mathcal{L}_{I\to I}$ maximizes Zero-Shot Accuracy (with the latter achieving the table's peak), it induces a regression in Linear Probing. For instance, single-source models often retain slightly higher probing scores than their ensemble counterparts ($m = 4$).

We attribute this to the "Semantic Purist" nature of our framework. Simple linear classifiers often exploit low-level texture shortcuts or generator-specific spectral artifacts to discriminate classes. By enforcing invariance across diverse manifolds (via multi-positives) or explicitly penalizing feature divergence (via $\mathcal{L}_{I\to I}$), PolyGen strips away these non-semantic features. While this removes the textural cues used by linear probes, it forces the model to learn robust, abstract representations that transfer better to open-world tasks. Ultimately, even with this trade-off, the full PolyGen configuration yields the highest Δ_{MTL} improvement, offering a pareto-optimal solution that prioritizes semantic alignment, the primary goal of Vision-Language pre-training, while maintaining competitive visual representation.

n^+	m	*HN samples*	$\mathcal{L}_{HN}$	*scheduler*	*L.P.*	*Few-shot*	*Img Ret.*	*Text Ret.*	*0-shot*	Δ_{MTL}	SugarCrepe	SugarCrepe++	Δ^{HN}_{MTL}
1	1	✗	✗	✗	**49.1**	**61.6**	11.6	15.9	8.15	-	55.4	36.7	-
2	2	✓	✗	✗	41.5	55.7	13.2	17.9	9.95	+4.7%	56.9	40.0	+5.9%
2	2	✓	✗	✓	40.8	55.8	13.6	18.8	10.23	+7.0%	55.0	38.8	+2.6%
2	2	✓	✓	✗	42.3	56.8	14.0	19.0	10.27	+9.0%	55.8	39.2	+3.7%
2	2	✓	✓	✓	41.4	55.8	14.6	**20.5**	10.8	+12.5%	56.4	40.5	+6.1%
3	3	✓	✓	✓	40.8	55.5	**15.7**	18.9	10.72	+11.8%	**58.1**	**41.6**	**+9.1%**
4	4	✓	✓	✓	42.7	56.5	**15.7**	20.4	**11.1**	**+15.5%**	56.8	40.7	+6.8%

Table 2. Average per-task results of models using hard negatives. From the second to the last row, models use $\mathcal{L}_{I2I}$. When used, $\mathcal{L}_{HN}$ replaces $\mathcal{L}_{MP}$.

4.4. Hard Negatives and Curriculum

Table 2 provides a granular dissection of the training dynamics, particularly within the $m = 2, n^+ = 2$ block.

The Limits of Passive and Delayed Learning. We first establish the baseline behavior using "Naive" Hard Negatives, simply including counterfactual captions in the batch without targeted objectives. This approach yields high compositional performance (SugarCrepe++ 40.0%) because the model is exposed to fine-grained variations throughout training. However, this early complexity acts as a distracter for general semantic alignment, resulting in a modest improvement on standard tasks (Δ_{MTL} +4.7%) compared to the cleaner runs in the previous section. Attempting to solve this via the Curriculum Scheduler alone introduces the inverse problem. By delaying the introduction of hard negatives, we shield the early training phase, allowing standard tasks to improve (Δ_{MTL} +7.0%). Yet, this "delayed viewing" comes at the cost of compositional reasoning: the model simply does not see the difficult counterfactuals often enough to learn the subtle syntax of attribute binding.

Impact of the Hard Negative loss. The behavior of $\mathcal{L}_{HN}$ in isolation proves effectively, with a higher Δ_{MTL} compared to the same configuration using $\mathcal{L}_{MP}$. We attribute this to a key change of TripletCLIP loss, to which $\mathcal{L}_{HN}$ is based, and standard CLIP. As the generated images often

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. **LLM Selection for Hard Negatives.** Semantic Axes aggregated results of the LLM negative generation evaluation.

fail to correctly represent slight variations in captions (Appendix 7.3), TripletCLIP removes from the loss $T^+ \rightarrow I^-$ and $T^- \rightarrow I^+$ comparisons. This allows the model to keep exploiting hard negatives ($I^+ \rightarrow T^-$ and $I^- \rightarrow T^+$), while not being misdirected by the noisier signal given by hard negative images.

On the other hand, removing such components of the loss leads to sub-optimal performance gains in compositional benchmarks.

Optimization Synergy. The full PolyGen configuration empirically resolves the stability-plasticity dilemma. By combining the Scheduler and Triplet Loss, we achieve a Δ_{MTL} of +12.5%, surpassing both the Scheduler-only and Loss-only configurations. This indicates a corrective mechanism: the Scheduler protects early feature formation from noisy gradients, while the Triplet Loss compensates for the reduced exposure time by enforcing a stricter margin on the hard negatives once they are introduced. This recovery allows us to match the high compositional score of the Naive approach without sacrificing general performance.

Scaling and the Pareto Frontier. Consistent with our main results, scaling the ensemble size (m) and positive density (n^+) yields monotonic gains for general open-world tasks, with the $m = 4$ configuration reaching the Δ_{MTL} peak. However, for compositional reasoning, we observe a distinct peak at $m = 3$, outperforming $m = 4$. We attribute this to a *variance saturation point*: while maximal diversity ($m = 4$) benefits broad semantic coverage, it introduces excessive visual noise for fine-grained syntactic tasks, making it harder for the model to isolate subtle attribute changes amidst drastic stylistic shifts. Ultimately, the configuration $m = 4$ represents the *Pareto-optimal* choice: it maximizes general transferability while significantly improving compositional reasoning, offering the most robust balance between semantic breadth and syntactic depth.

4.5. Ablation: LLM Selection for Synthesis

Our pipeline employs two different LLMs. For base captions, we use Mistral-v0.2-7B to maintain parity with the SynthCLIP baseline. Hard negative synthesis is, instead, more critical, as it demands high-fidelity instruction following to modify specific semantic attributes without altering the overall scene structure. To this end, we evaluated a suite of 7B/8B-parameter LLMs for correct negative generation: given a positive-negative caption pair generated by a model from a pair (c, a), we verify if an LLM ensemble can recover the correct a from the caption pair. As shown in Table 3, Llama-3.1-8B significantly outperforms its peers, achieving 74% classification accuracy on its generated negatives.

5. Discussion and Limitations

The Synergy of Diversity and Recognizability. Our results highlight a fundamental trade-off in current generative backbones when used for synthetic training VL pipelines: high-diversity models introduce noise that dilutes the learning signal, while high-fidelity models induce overfitting to narrow visual modes. PolyGen resolves this by integrating these complementary manifolds. The high-fidelity generators provide strong "semantic anchors" that ensure the text concept is correctly instantiated, while the high-diversity generators expand the visual support of that concept, preventing the model from latching onto specific patterns. This synergy forces the model to ignore the distinct spectral fingerprints of each architecture and converge on the only shared signal: the invariant semantic content.

Why Synthetic Training? The transition to synthetic pre-training is motivated by the need for curation over collection. Unlike web-scraping, synthetic pipelines offer precise control, enabling privacy-preserving, bias-aware, and multimodally scalable datasets. However, adoption has been hindered by the lower efficiency of synthetic data. PolyGen contributes to take a step forward: a relevant factor for this "inefficiency" is single-generator homogeneity. By showing that manifold diversity drastically improves per-sample utility, we establish that synthetic pre-training is not just about scale, but it can be optimized for efficiency through structured ensemble sampling.

Limitations and Future Work. Our approach entails specific design trade-offs. First, while not necessarily computationally more expensive, maintaining a heterogeneous ensemble introduces higher pipeline complexity than single-model baselines. Second, to adhere to a fixed budget, our multi-positive strategy prioritizes semantic robustness over unique concept coverage, a balance that requires recalibration for larger-scale regimes. Lastly, while our current concept bank and semantic axes ensure broad coverage, it remains a simplification of real-world visual complexity. Future work will extend the strategy along three dimensions: (1) increasing ensemble size, thus increasing both m and n^+, with architectures showing diverse Recognizability and Diversity properties; (2) scaling the data budget to evaluate the breadth-depth Pareto frontier at larger-scale regimes; (3) expanding programmatic hard negatives to cover complex axes

such as multi-concept relationships, counting, and causality, ultimately obtaining a synthetic curriculum capable of fully grounding open-world reasoning.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 6

[2] 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. 6

[3] 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)*, pages 1–5. IEEE, 2023. 1, 2

[4] 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*, pages 248–255. IEEE, 2009. 6

[5] 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. 3, 11

[6] 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. 6

[7] 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. 2

[8] 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*, pages 7382–7392, 2024. 1, 2, 3

[9] 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. 6

[10] 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. 1, 2, 5

[11] 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*, pages 7514–7528, 2021. 3, 11

[12] 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. 11

[13] 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. 2, 6

[14] 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*, pages 20406–20417, 2023. 11

[15] 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*, pages 4904–4916. PMLR, 2021. 1

[16] 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. 3, 11

[17] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009. 6

[18] 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*, pages 111–127. Springer, 2024. 2

[19] Andrea Lampis, Eugenio Lomurno, Matteo Matteucci, and Via Ponzio. Bridging the gap: Enhancing the utility of synthetic data via post-processing techniques. 2023. 1

[20] 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*, pages 740–755. Springer, 2014. 6, 11

[21] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[22] 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*, 2008. 6

[23] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 6

[24] 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. 2, 5, 13

[25] 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*, pages 6967–6977, 2023. 2, 13

[26] 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*, pages 8748–8763. PmLR, 2021. 1

[27] Nicolò Francesco Resmini, Eugenio Lomurno, Cristian Sbrolli, and Matteo Matteucci. Your image generator is your new private dataset. *Image and Vision Computing*, page 105727, 2025. 2

[28] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020. 2

[29] 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*, pages 10684–10695, 2022. 4

[30] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2024. 4

[31] TII Team. The falcon 3 family of open models, 2024. 11

[32] 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. 2, 4

[33] 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*, pages 15887–15898, 2024. 1, 2

[34] 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. 6, 14

[35] 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. 1

[36] 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)*, pages 7854–7873, 2025. 2, 13

[37] 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. 11

[38] 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*, pages 3485–3492. IEEE, 2010. 6

[39] 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. 4

[40] 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. 3, 14

[41] 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. 11

[42] 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. 6

[43] 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. 2

6. Caption Generation Prompts

Fig. 5 shows the zero-shot prompts we use to generate captions. The baseline set of captions is generated with the prompt shown in Fig. 5a, asking the model to generate a caption given the starting concept $c \in C$, and optionally, asks 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 Fig. 5b, which instructs the LLM to modify the base caption on its selected semantic axis only, while maintaining the rest of the sentence structure.

7. Generative Models Selection

7.1. LLM selection for hard negative captions - attribute classification

The task of generating a hard negative caption requires strict instruction following. To select the best LLM for the task, we evaluate a pool M of similar-sized LLMs, namely Mistral-V0.2-7B and Mistral-V0.3-7B [16], Llama-3.1-8B [5], Falcon-3-7B [31], and Qwen-2.5-7B [41], on a hard-negative attribute classification ensemble. Given the LLM $m \in M$, the concept c, the base caption c_1, and the hard negative caption c_2 generated by m by modifying attribute $a \in A$, we use the prompt defined in Fig. 6 to ask each LLM to classify which attribute of A has been used to generate c_2. The final classification is given by majority voting of the predictions made by each model. For each $m \in M$, we generate $4.8k$ pairs of base and hard negative captions. Specifically, we generate 960 base captions for each of a limited set of axes $\mathcal{A} = [\textit{background}, \textit{color}, \textit{position}, \textit{size}, \textit{concept}]$. Out of those 960 captions, 160 are used for generating an *unaltered* caption, where the model is asked to just rephrase the original caption (Fig. 6). This is used as a control label to assess the model's understanding of whether meaningful information has been changes. The remaining 800 captions are used to generate hard negatives. Tab. 4 shows how Llama greatly overperforms other LLMs in this task, generating captions clearly and strictly modifying the requested axis.

7.2. LLM selection for baseline captions

The choice of Mistral-V0.2-7B for generating the base caption set has two different reasons. First, for a fairer comparison to the SynthCLIP pipeline, which uses the same model to generate its caption set, shifting the focus towards the prompting strategy and our specific contributions. Secondly, we conduct an ablation on the LLM, generating a $500k$ image-caption pair dataset for both Mistral-V0.2-7B and Llama-3.1-8B, with corresponding images generated by SDXLT. We report a slight advantage on ImageNet zero-shot accuracy with captions generated by Mistral (1.8% top-1, 5.97% top-5) over Llama (1.66% top-1, 5.61% top-5), further validating our choice.

7.3. Diffusion models selection

Image quality metrics. The selection of the image generators used in PolyGen (SD 1.5, SD 2, SANA, SDXLT) is the result of a comparison of a broader pool of models, also comprising Stable Diffusion v2.1 (SD 2.1), Stable Diffusion 3.5 Large Turbo (SD 3.5 LT), and FLUX.1 Schnell (FLUX). We randomly sample one caption for each of the $5k$ samples of MS COCO 2017 [20] validation set, and generate the corresponding image with each model. We use these image sets to compare models based on FID [12], CLIPScore [11], and TIFA score [14]. Although FID was originally presented as a metric to measure the fidelity of images from GAN models trained on narrow domains, this measure correlates with diversity, as it measures the distribution shift of a set of image features w.r.t corresponding real data. In many diffusion models employing CFG (cite), a low CFG scale is often correlated with lower (better) FID, but also with lower (worse) alignment with the input caption, measured by the CLIPScore, and consequently with lower perceptual quality. For these reasons, "Diversity expert" models are related to a low FID, while "Recognizability experts" are related to a higher CLIPScore and TIFA. Fig. 8 shows the results of these evaluations on various resolutions and denoising steps. Although we select the final ensemble taking into account computational constraints, models like FLUX and SD 3.5 could be considered to be used as "Recognizability experts" and SD 2.1 as "Diversity expert".

Semantic axis classification. We also test the diffusion models on their ability to correctly and clearly report the variation contained in hard negative captions. To do so, we classify the semantic axis similarly to what we did with the LLMs. This time, we use Qwen 2.5 VL [37] as the classifier, taking as input the image generated from the base caption, the image generated from the hard negative caption, and the image classification prompt shown in Fig. 6. This time, we use a total of 2.4k pairs of images, 480 for each of the 5 same axes used for testing the LLMs, 80 of which are used for the unaltered class. Tab. 5 shows how the task of generating useful hard negative images is easier for models with higher image-caption alignment and perceptual quality. Overall, the lower scores compared to the similar experiment on LLMs, due to the noisier signal given by images, justify the use and the performance improvement of the TripletCLIP loss in contrastive models using hard negative samples.

8. Semantic Axes Selection

Given the results for the semantic axis classification, both to evaluate LLMs and T2I models, we notice how the ability of a model to generate effective hard negatives varies significantly depending on the specific axis being modified (Fig. 7, 9). The final set of axes is then obtained by a second round of classification from a broader set of axes A=[*background*,

Your task is to write an image caption that includes and visually describes a scene around a concept. The concept must be explicitly included in the caption. Your concept is {c}. Include {attribute} to the caption. Output one single grammatically correct caption that is no longer than 20 words. Do not output any notes, word counts, facts, etc. Output one single sentence only.

attribute in ["a background", "a color", "a lighting related term", "a material", "a positional relationship", "the perspective or camera angle", "an artistic style"]

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

Given the following image caption, generate a new similar caption by only changing the [concept / {attribute1}]. The concept to change is {c}.
Caption: {caption}
Alter the caption as less as possible for it to be similar to the original caption but still making sense with the new [concept / {attribute2}]. Also, do not change the main concept of the caption being {c}. Do not output any notes, word counts, facts, etc. Output the new caption only.

attribute1 in ["background", "color", "lighting, possibly with an opposite meaning", "material", "positional relationship", "perspective", "artistic style"]
attribute2 in ["background", "color", "lighting", "material", "positional relationship", "perspective", "style"]

(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 5. **Zero-shot programmatic conditioning.** (a) Baseline prompt enforcing structured concept-attribute tuples (c, a) to ensure semantic orthogonality. (b) Hard Negative prompt imposing strict counterfactual constraints to generate valid T^- samples without syntactic drift.

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 6. **Verification protocols for synthetic data.** 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 (bottom-left).

color, *lighting*, *material*, *perspective*, *position*, *size*, *style*, *texture*, *concept*], both on pairs of captions generated with Llama, and pairs of images generated by SDXLT. Fig. 10 shows the results of these experiments. We can notice how Llama struggles with generating hard negatives when *texture* is the attribute to be changed. At the same time, SDXLT-generated images struggle to represent the *size* differences between captions. For these reasons, these two attributes are removed from the final set of semantic axes, reported in Fig. 5.

9. T2I Models Comparison in Frequency Domain

To get additional insights into the characteristics of the generated images and to try to understand why different diffusion models lead to different training performance, we analyze the images in the frequency domain. This analysis compares the spectral characteristics of images generated by each model against the baseline dataset of real images. To do so, the 2D Fast Fourier Transform (FFT) is applied to each standardized image to convert it from the spatial domain to the frequency domain. From the 2D magnitude spectrum of each image, we extract two key sets of metrics:

- **Radial Frequency Profile:** The 2D spectrum is reduced to a 1D vector using a radial averaging function. This represents the average magnitude as a function of spatial frequency (i.e., the radius from the center). This process is repeated for all $5k$ images in the set, and the mean radial profile for that set is computed.
- **Frequency Energy:** The energy distribution is quantified for each individual image. We calculate the total energy on the high-frequency components $[0.5, 1]$ (normalized units).

Fig. 11 shows this analysis for each of the 4 diffusion models compared to the MS COCO real images. It is evident how the diffusion process acts similarly to a low-pass filter, attenuating high-frequency components. It is nevertheless clear how different models have different frequency responses, with SDXLT having the steepest decay of high-frequency components. As this might be describe as a consequence of

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 4. **Aggregated Results of ensemble attribute classification.** Llama 3.1 exhibits superior adherence to semantic constraints, minimizing the risk of false negatives (logical contradictions) in the counterfactual caption generation process.

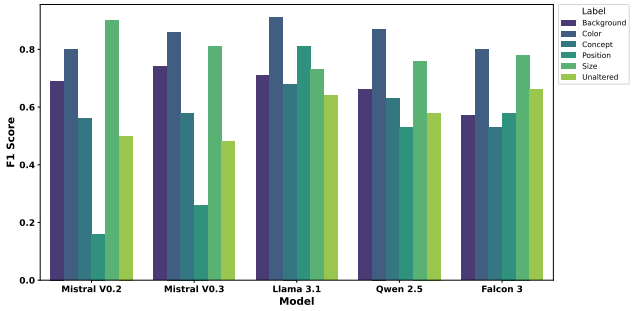


Figure 7. **F1-scores of classification for each LLM and semantic axis (label).** Classification performance across modification types. Some axes (e.g., texture) prove inherently more difficult for LLMs to modify without altering the global context.

the deeply different training procedures and objective of each model (e.g., the strong compression achieved via distillation on SDXLT for low-steps inference), we observe this might also correlate with diversity and distribution shift, measured via FID score, with the only exception of SANA, which has a less steep decay but the lowest recorded diversity. This analysis suggests that the frequency characteristics of generated images could be a contributing factor to the recognizability-diversity tradeoff observed earlier, and consequently to the downstream performance of models trained on such data. Further investigation is although needed to fully understand these relationships.

10. Hard-negative-aware loss functions

Figure 12 presents an ablation on a set of hard-negative-aware loss funcions comparing $\mathcal{L}_{CLIP}$, with and without $\mathcal{L}_{margin}$ [36], $\mathcal{L}_{HN-NCE}$ [25], and $\mathcal{L}_{TripletCLIP}$ [24]. From these results, it seems that TripletCLIP overcomes the excessively noisy signal introduced by hard negative images while still leveraging hard negatives, preserving the image-to-text direction. Differently, the margin loss also seems to help performance, trying to enforce a relative hierarchy of similarity between the anchor, its hard negatives, and other "easier", less similar negatives. This loss can, however, be exploited more effectively when there is a 1-to-many relationship between anchors and hard negatives, which is not the case in our current dataset design. A very different behavior is shown by HN-NCE, whose importance-weighting strategy seems peak at very early training stages, but then quickly degrades, possibly due to an excessively punishing

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
		50	60.213	27.128	68.8
	768^2	20	23.047	29.991	81.3
		50	23.481	30.068	82.5
SD 3.5 LT	512^2	4	35.448	31.586	88.1
		8	36.008	31.109	87.7
	1024^2	4	36.732	31.484	89.4
		8	36.337	31.358	89.1
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**

Figure 8. **Model comparison on different image resolutions and number of inference steps.** Best results per metric are highlighted in bold. Models which are or can be used 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%
		8	32%	33%	32%	32%
	1024^2	4	35%	31%	35%	35%
		8	35%	30%	36%	35%
FLUX	512^2	4	36%	33%	42%	36%
	1024^2	4	**38%**	**35%**	**43%**	**39%**

Table 5. **Visual grounding of semantic perturbations.** High-fidelity architectures (Recognizability Experts) show significantly stronger attribute binding capabilities, essential for correctly rendering the fine-grained distinctions required by the HardNegative curriculum.

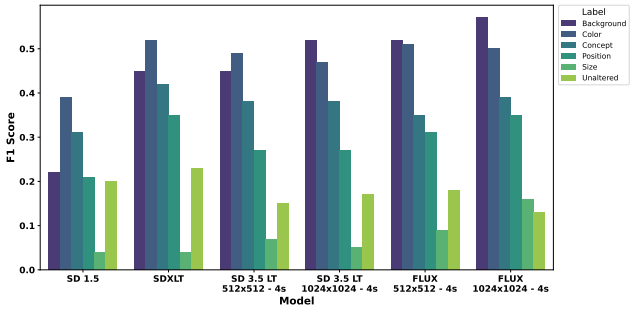
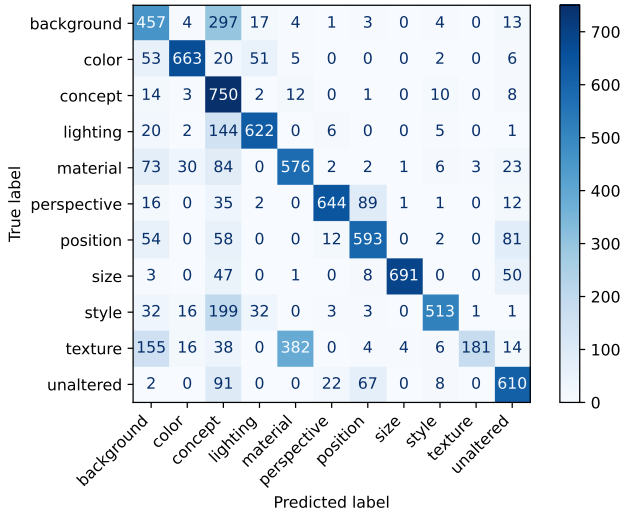
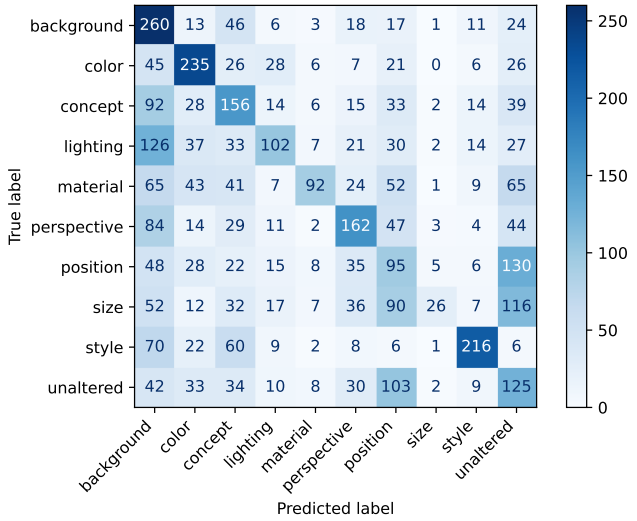


Figure 9. **F1-scores of classification for each T2I and semantic axis (label).** SDXL-Turbo and FLUX demonstrate superior ability in rendering spatial and object-level changes compared to older architectures.



(a) Llama results. Texture attribute is removed from the final set of semantic axes.



(b) SDXLT results. Size is removed from the final set of semantic axes.

Figure 10. **Confusion matrices for semantic axis verification.** This analysis drove the pruning of the modification space $\mathcal{A}$, eliminating axes with high confusion rates (e.g., "Texture", "Size") to ensure the training signal remains structurally consistent.

weighting scheme in these settings.

11. Balanced sampling

Balanced sampling is a sampling algorithm proposed by Xu et al. [40] as a data curation strategy designed to overcome the "rich-get-richer" problem found in raw web-crawled datasets, where a small number of common concepts appear millions of times, while rare concepts are underrepresented, leading to a highly skewed distribution that limits a model's ability to generalize. Given a thresholding count t, this algorithm assigns a sampling probability to each concept in C equal to $t/count(c)$, where $count(c)$ equals to how many times c appears in the caption set, if $count(c) > t$, and a probability of 1 otherwise. Each caption is sampled based on the probability of each concept contained in it. We generate an initial caption set of $100M$ captions, then de-duplicate the captions and run this sampling algorithm with $t = 30$, leading to a final caption set of $13.1M$ captions. We finally uniformly downsample the caption set to our final target caption set T^+ of the desired size. Fig. 13 shows how this algorithm manages to apply a strong downsampling to high-frequency concepts.

12. Delta Multi-Task Learning Metric

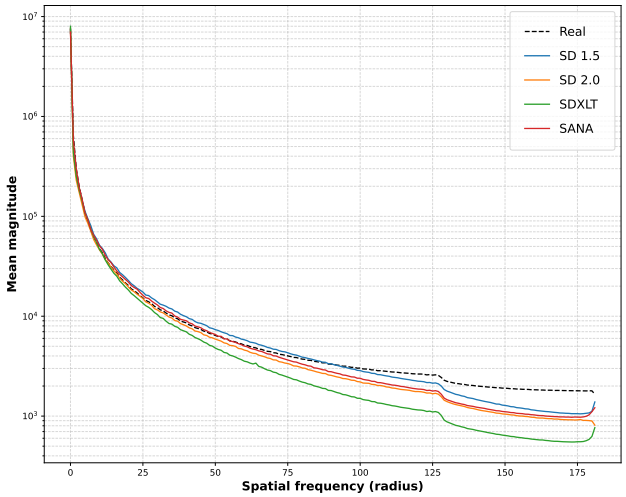
To summarize the overall performance of models capable of performing multiple tasks, aggregating metrics which are not directly comparable, Vandenhende et al. [34] defined the Delta Multi-Task Learning (Δ_{MTL}) metric, which computes the average relative performance gain or drop of a model m w.r.t. a baseline model b on a set of tasks T:

$$\Delta_{MTL} = 100 \times \frac{1}{T} \sum_{i=1}^{T} (-1)^{l_i} \frac{M_{m,i} - M_{b,i}}{M_{b,i}}, \qquad (9)$$

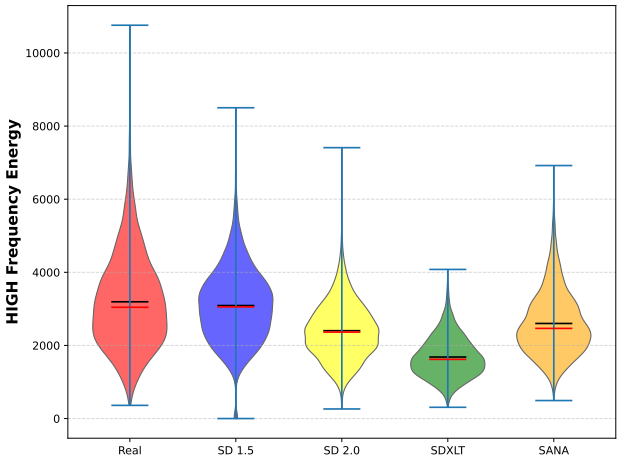
with l_i being 1 if a lower value means better performance for metric i, and 0 otherwise.

13. Complete PolyGen results

Tab. 6 and 7 report the results of evaluation over each dataset for models with different configurations of model ensemble (m) and positive images (n^+), with and without $\mathcal{L}_{I2I}$. Similarly, Tab. 8 and 9 report the results of evaluation over each dataset for models including hard negative samples. Models leveraging hard negatives are also tested with the compositional benchmarks *SugarCrepe* and *SugarCrepe++*, whose results are reported in Tab. 10.



(a) Radial profiles of the employed models.



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

Figure 11. Frequency analysis of $5k$ generated images from MS COCO captions, compared with real MS COCO images. We observe that SD 1.5 is the model that most closely resembles real images in terms of frequency characteristics, while SDXLT shows the most significant deviation.

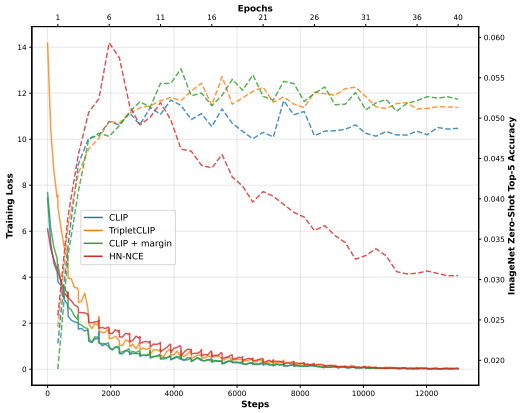


Figure 12. Loss function ablation on $500k$ Mistral/Llama+SDXLT dataset. $\mathcal{L}_{HN-NCE}$ shows a strong initial boost, but then quickly degrades. $\mathcal{L}_{TripletCLIP}$ and $\mathcal{L}_{CLIP} + \mathcal{L}_{margin}$ manage to outperform the standard $\mathcal{L}_{CLIP}$.

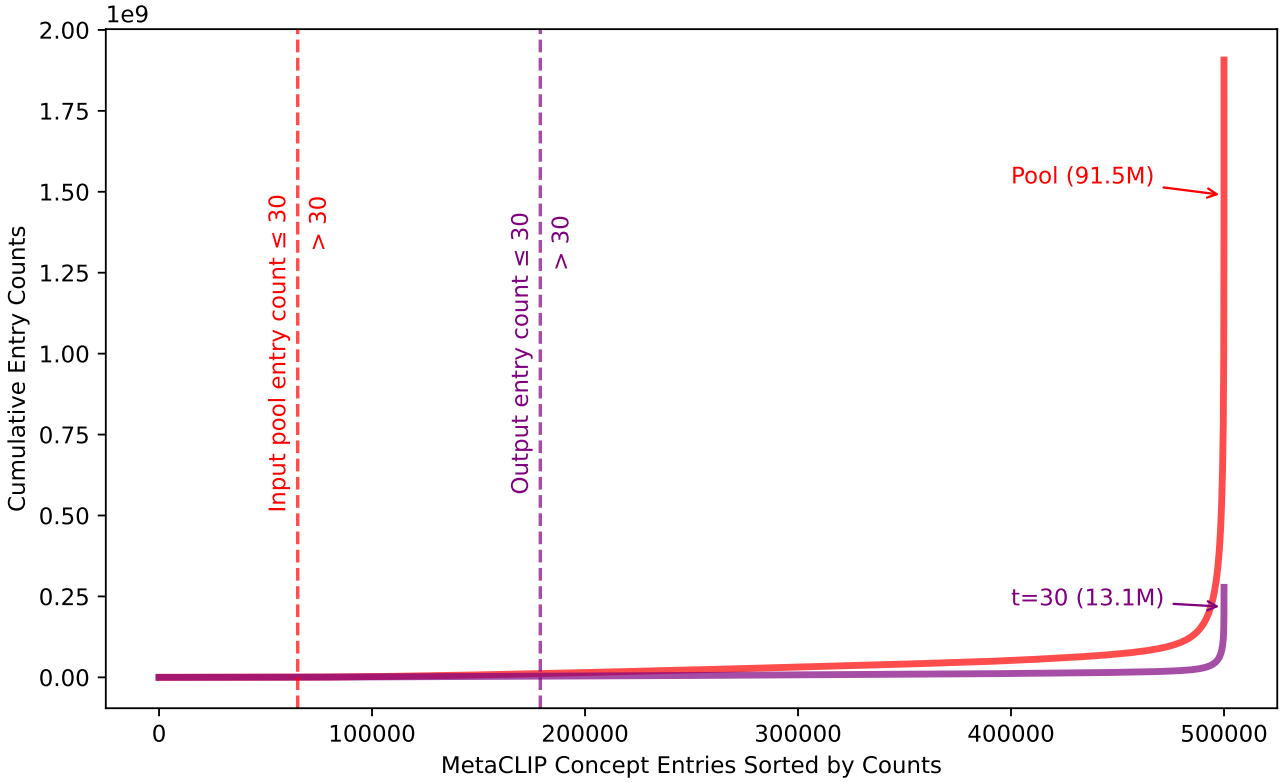


Figure 13. **Distributional rectification via Balanced Sampling.** Cumulative entry count for the deduplicated starting caption set (pool) and the final caption set after balanced sampling. The algorithm flattens the power-law distribution, increasing the effective sampling rate of tail concepts to ensure uniform semantic coverage.

				Vision Datasets									
Method	n^+	m	$\mathcal{L}_{I2I}$	CIFAR10	CIFAR100	Aircraft	DTD	Flowers	Pets	SUN397	Caltech	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
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

Table 6. **Classification performance on downstream vision datasets.** We report Top-1 accuracy for both Linear Probing and 5-shot 5-way classification. The $\mathcal{L}_{I2I}$ column indicates the inclusion of the image-to-image consistency loss.

			Image Retrieval				Text Retrieval				0-shot
n^+	m	$\mathcal{L}_{I2I}$	MS Coco	Flkr 8K	Flkr 30K	**Avg**	MS Coco	Flkr 8K	Flkr 30K	**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**

Table 7. **Cross-modal retrieval and zero-shot classification.** We report Recall@5 for Image/Text Retrieval tasks on MS COCO and Flickr datasets, and Top-5 accuracy for ImageNet zero-shot classification.

		HN Components			Vision Datasets									
n^+	m	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**
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														
1	1	✗	✗	✗	**46.3**	**57.8**	**39.8**	**63.5**	**85.6**	**48.5**	**84.2**	**75.4**	**53.4**	**61.6**
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 8. **Ablation study of Hard Negative (HN) components on vision tasks.** We analyze the impact of sampling strategies (Smpl), HN loss ($\mathcal{L}_{HN}$), and the scheduler (Sched) on classification accuracy.

		HN Comp.			Image Retrieval				Text Retrieval				0-shot
n^+	m	Smpl	Loss	Sched	MS Coco	F8K	F30K	**Avg**	MS 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
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 9. **Impact of Hard Negative strategies on cross-modal retrieval.** We compare different HN configurations on vision-language benchmark datasets (MS COCO, Flickr).

		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 10. **Fine-grained compositionality evaluation on SugarCrepe and SugarCrepe++.** We report accuracy across different error types (Replace, Swap, Add). 'S' denotes the use of the Hard Negative scheduler.

Figure 14. **Visualizing Generator Invariance.** The ensemble renders identical semantic shifts across disjoint visual manifolds, forcing the encoder to align with the shared concept modification rather than generator-specific stylistic cues.

Figure 15. **Failure modes and Ensemble resilience.** Instances of attribute leakage or visual ambiguity in single generators. The multi-generator approach mitigates these local errors by averaging the learning signal across the ensemble.