UNPAIRED MULTIMODAL LEARNING WITH CROSS-MODAL SELF-SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal learning typically relies on paired observations, yet many domains, including biology and medicine, provide abundant unimodal data with only sparse and potentially imperfect pairing. We introduce `unimo-x`, a general learning framework for training multimodal Transformers on highly unpaired corpora. Our approach uses teacher-anchored cross-modal self-supervision to turn unpaired observations into training signals, complementing paired supervision while mitigating steganographic shortcuts that can undermine consistency-based learning. We instantiate `unimo-x` in 4M and evaluate it across several datasets and tasks. Unpaired observations improve image generation and caption semantics over paired-only training. With 25%, 17.5%, or 10% paired samples, `unimo-x` recovers substantial benefits of paired supervision, from competitive image-generation quality with four times fewer pairs to stronger caption representations than paired-only training even at 10% pairing. These findings point to a practical route to multimodal learning in which sparse paired examples guide training on much larger unimodal corpora.

1 INTRODUCTION

Multimodal foundation models learn cross-modal representations and generation from large paired corpora (Radford et al., 2021; Team, 2024), making access to correspondences a central constraint on scaling. Yet many domains provide abundant unimodal data with only sparse, potentially imperfect pairing. In biology and medicine, for example, destructive assays, acquisition costs, and independently collected measurements limit access to matched observations (Cui et al., 2025; Xu et al., 2022). In oncology specifically, imaging, pathology, and molecular profiles provide complementary views of a tumor, but are often collected in different cohorts. Learning their relationships illustrates a broader challenge: how can sparse paired examples support multimodal learning from much larger unpaired corpora?

Answering it calls for multimodal training on predominantly unpaired data. Sparse paired examples supervise cross-modal mappings, while larger unimodal corpora offer additional structure and diversity. Exploiting them requires a training signal that connects modalities without observed counterparts, which unimodal objectives alone do not explicitly provide. Our goal is thus to train a unified multimodal model on predominantly unpaired corpora, using limited paired supervision to establish cross-modal connections.

Concretely, our approach uses *cross-modal self-supervision* to turn observations of a single modality into training signals across modalities, extending learning beyond examples with known correspondences. The model generates the missing modalities and learns through them: reconstructing the observed input provides feedback on its cross-modal predictions (Zhu et al., 2017; Li et al., 2024b; Zhao et al., 2025). The central challenge is ensuring that these generated intermediates express meaningful content. A model can otherwise hide information in its predictions and recover it without learning the intended cross-modal relationships (Chu et al., 2017; Porav et al., 2019). Overcoming this *steganographic* shortcut is essential for turning reconstruction into useful cross-modal supervision.

To this end, we introduce `unimo-x`, a framework for **un**paired **m**ulti**mo**dal learning through cross-modal (**x**) self-supervision (Figure 1). An exponential-moving-average teacher generates intermediate modalities through chained sampling; a student predicts these intermediates and reconstructs

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.

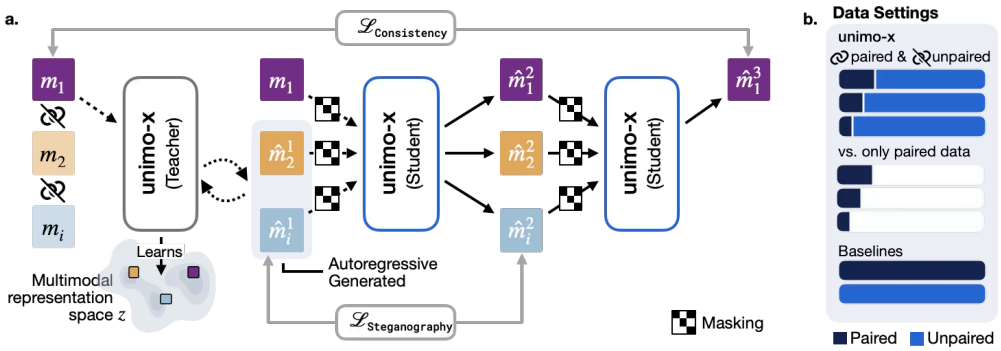


Figure 1: **Learning from unpaired data with `unimo-x`.** **a**, Given modality m_1, the teacher generates intermediate modalities $\hat{m}_i^1$ through chained sampling. The student predicts these intermediates under an anchoring loss that mitigates steganography, then reconstructs the input to close the cycle. Teacher weights track an exponential moving average of student weights; dashed arrows indicate no gradient flow. **b**, Evaluated training settings combine paired and unpaired data in varying proportions and from different corpora.

the source, avoiding backpropagation through the full sampling procedure. An anchoring loss constrains the student's intermediates toward the teacher's outputs to mitigate steganographic shortcuts. Together with sparse paired supervision, this enables a unified multimodal Transformer to learn from predominantly unpaired corpora. We instantiate `unimo-x` in 4M (Mizrahi et al., 2023) and make the following contributions:

- We introduce `unimo-x`, a learning framework that turns predominantly unpaired data into cross-modal supervision for a unified Transformer. Its student-teacher formulation combines sparse paired examples with cross-modal reconstruction without requiring backpropagation through iterative generation.
- We identify and quantify steganographic shortcuts that undermine cross-modal learning, and introduce a teacher-anchoring objective that suppresses these shortcuts while preserving the semantic fidelity of generated intermediates.
- Experiments span four datasets (CC12M and CommonCatalog for training, COCO and ImageNet for evaluation) and four tasks: image generation, captioning, retrieval, and caption-based classification. With 10%, 17.5%, or 25% pairing, `unimo-x` improves caption class information over paired-only training at every fraction and image generation at 17.5% and 25%. At 25%, FID surpasses our own fully paired run (40.90 versus 42.79) and approaches released 4M-B (40.56), and the gain persists when unpaired data come from a different corpus.

2 Related Work

Multimodal foundation models. Multimodal foundation models learn representations and generative capabilities across modalities through contrastive learning, conditional generation, and unified sequence modeling (Radford et al., 2021; Ramesh et al., 2022; Alayrac et al., 2022; Girdhar et al., 2023; Driess et al., 2023; Huang et al., 2023; Li et al., 2023). These approaches increasingly extend to scientific and biomedical data (Cui et al., 2025; Wenckstern et al., 2026; Bai et al., 2025; Han et al., 2025; Yan et al., 2025); Li et al. (2024a) provide a comprehensive review. Their reliance on paired observations, however, makes cross-modal supervision a constraint on training scale. Our work addresses this constraint by enabling a unified multimodal model to learn from predominantly unpaired corpora, using sparse paired examples to establish cross-modal connections.

Learning from unpaired data. One approach to learning without observed correspondences is to infer alignments through distributional matching or optimal transport (Xi et al., 2024; De Bortoli et al., 2024; Samaran et al., 2024; Lee et al., 2025; Chi et al., 2026). Another line of work connects pretrained representations to support cross-modal prediction and generation (Zhai et al., 2022; Merullo et al., 2022; Liu et al., 2023; Koh et al., 2023). Unpaired observations can also improve unimodal representations (Gupta et al., 2026), while theoretical work studies conditions under which

shared structure can be identified across unpaired domains (Sturma et al., 2023; Timilsina et al., 2024). We focus on joint multimodal training: generated intermediates turn unpaired observations into supervision for cross-modal prediction. The multimodal backbone is trained from scratch, while modality-specific tokenizers and the text encoder provide frozen pretrained components.

Cycle consistency. Cycle consistency provides supervision by reconstructing an observation after transforming it through another domain or representation. It has supported correspondence learning, unpaired image translation, depth estimation, and temporal alignment (Zhou et al., 2016; Zhu et al., 2017; Godard et al., 2017; Dwibedi et al., 2019). Related reconstruction and back-translation objectives also exploit unpaired data in language and biomedical applications (Sennrich et al., 2016; Lample et al., 2018; Xu et al., 2022; Zhang et al., 2018), with further extensions to multimodal generation and reasoning (Xu et al., 2023; Diesendruck et al., 2024; Zhang et al., 2026). Closest to our setting, ITIT (Li et al., 2024b) combines paired and unpaired image–text data using a shared encoder and separate image and text decoders, while DoraCycle (Zhao et al., 2025) adapts a unified generative model through multimodal cycles. Our contribution centers on teacher-anchored cross-modal self-supervision within a shared encoder–decoder backbone, together with diagnostics and ablations of the shortcuts that arise in generated intermediates.

Steganography and shortcut learning. Cycle-consistent models can hide source information in generated outputs and recover it on the return path, achieving accurate reconstruction without learning the intended translation (Chu et al., 2017; Porav et al., 2019). This behavior is a form of shortcut learning, in which a model exploits predictive cues that do not support the intended generalization (Geirhos et al., 2020; Hermann et al., 2024; Chi et al., 2025). In our setting, high-capacity intermediate modalities can carry private codes that reduce reconstruction loss while degrading semantic fidelity. We study this failure mode in discrete multimodal representations through semantic-distance and intermediate-substitution diagnostics. An anchoring loss constrains student predictions toward teacher-generated targets, mitigating the measured shortcuts without requiring observed counterparts for unpaired samples.

3 Learning from Highly Unpaired Data Using UNIMO-X

3.1 Problem Setting

Learning across modalities. We consider $M \geq 2$ modalities, with observations $x_m \in \mathcal{X}_m$. To account for missing modalities, we define $\tilde{\mathcal{X}}_m = \mathcal{X}_m \cup \{\varnothing\}$, where $\varnothing$ denotes a missing observation. Our goal is to learn a multimodal encoder

$$f_\theta : \prod_{m=1}^{M} \tilde{\mathcal{X}}_m \to \mathbb{R}^e$$

that maps observations with potentially missing modalities into a shared latent space. Together with a shared decoder and modality-specific output heads, the model predicts across modalities and provides representations for downstream tasks.

We assume access to a *paired dataset* $\{(\mathbf{x}_i^{(1)}, \ldots, \mathbf{x}_i^{(M)})\}_{i=1}^{n_{\text{paired}}}$, in which all modalities are observed, and an *unpaired dataset* $\{(\mathbf{x}_i^{(m)})_{m \in M_i}\}_{i=1}^{n_{\text{unpaired}}}$, where $M_i \subseteq [M]$ denotes the observed modalities for sample i. We focus on the setting in which each unpaired sample contains a single observed modality and the unpaired pool is larger than the paired pool.

Beyond two modalities. The model uses a shared encoder-decoder backbone across all M modalities. It can condition on several modalities and predict additional ones through modality-specific heads, allowing cycles with more than one intermediate modality. This is important for models in biology and medicine, where connecting genomic, transcriptomic, proteomic, and spatial measurements requires learning across several complementary modalities (Cui et al., 2025). The same parameters support the successive cross-modal predictions and the return to the source. Our experiments instantiate this formulation with image tokens, image-derived CLIP representations, and captions (§ 4.1); extending it to independently acquired biomedical modalities remains future work.

3.2 CROSS-MODAL SELF-SUPERVISION

Learning through generated modalities. We represent observations as token sequences using frozen modality-specific tokenizers. Paired observations first establish cross-modal mappings. Unpaired observations then train these mappings by supplying their own reconstruction targets: the model generates missing modalities and uses the resulting intermediates to recover the observed source. The reconstruction error supervises both the prediction of intermediates and the return to the source, connecting modalities without an observed counterpart (Figure 1).

Two requirements determine the training procedure. Reconstruction must transmit a learning signal through discrete intermediate predictions, and those predictions must retain meaningful cross-modal content rather than encode a private reconstruction code. We address these requirements with a student-teacher formulation, a straight-through estimator, and teacher anchoring.

3.2.1 PAIRED PRE-TRAINING

We first train on paired observations using multimodal masked modeling (Mizrahi et al., 2023; Bachmann et al., 2022). For each sample, we draw disjoint input and target token sets, $\mathcal{I}$ and $\mathcal{T}$, distributed across the observed modalities. Let $\mathcal{T}_m$ contain the target positions for modality m, and let $\mathcal{M}$ contain the modalities with nonempty target sets. We minimize cross-entropy, averaged first over target tokens within each modality and then across target modalities

$$\mathcal{L}_{\text{paired}}(\theta) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \frac{1}{|\mathcal{T}_m|} \sum_{i \in \mathcal{T}_m} -\log p_\theta\big(x_{m,i} \mid x_{\mathcal{I}},\, x_{m,<i}\big). \tag{1}$$

Conditioning on preceding target tokens $x_{m,<i}$ applies to autoregressive modalities. Modality-specific tokenization and masking are specified in § 4.1 and Table 4.

3.2.2 JOINT TRAINING ON PAIRED AND UNPAIRED DATA

After K_{pre} paired pre-training steps, we introduce the unpaired pool. Each update combines a paired batch and an unpaired batch under the objective

$$\mathcal{L} = \mathcal{L}_{\text{paired}} + \lambda(t)\, \mathcal{L}_{\text{Cycle}}, \tag{2}$$

where $\lambda(t)$ gradually introduces the unpaired training signal while paired supervision continues.

Student-teacher formulation. Iterative generation requires multiple forward passes over discrete samples. To avoid backpropagating through this sampling procedure, we maintain a student and an exponential-moving-average (EMA) teacher, following common self-supervised learning approaches (Siméoni et al., 2026; Zhou et al., 2022). The teacher is copied from the student when cycle training begins, updated after each optimizer step, and receives no gradients. The teacher supplies sampled intermediates, while the student predicts intermediate-token distributions in a single forward pass, with teacher forcing for autoregressive modalities. Reconstruction uses the student-completed intermediates, allowing its loss to train the intermediate predictions without traversing the teacher's sampling steps.

The teacher also provides targets for an anchoring loss. Anchoring penalizes deviations from its cross-modal predictions, initialized through paired supervision; reconstruction rewards intermediates that retain information needed to recover the source. These complementary objectives discourage steganographic shortcuts while learning from unpaired observations. The teacher's sampled token targets receive no reconstruction gradients, whereas the student's intermediate predictions are optimized for source recovery. Anchoring constrains this reconstruction-driven adaptation toward the teacher's cross-modal predictions, while continued paired supervision maintains an external training signal. The resulting cycle is computed in three passes, summarized in Algorithm 1 in § A.3.

Pass 1: Teacher generation. Let s denote the tokenized observation in source modality m_0, and let $(m_1, \ldots, m_K)$ be a selected chain of intermediate modalities distinct from the source. The teacher generates each intermediate conditioned on the source and all preceding intermediates

$$\hat{z}_{m_k} \sim G_{\bar{\theta}}\big(\cdot \mid s, \hat{z}_{m_1}, \ldots, \hat{z}_{m_{k-1}}\big), \qquad k = 1, \ldots, K, \tag{3}$$

where $G_{\bar{\theta}}$ denotes the teacher's generation procedure with EMA parameters $\bar{\theta}$. For $k = 1$, the teacher conditions only on s. Generation is performed without gradients.

Pass 2: Student prediction and anchoring. For each intermediate modality m, let Ω_m denote the positions to be predicted by the student. The student receives a visible subset s_{vis} of the source tokens together with the visible teacher tokens $\hat{z}_{\setminus\Omega}$, and predicts the hidden tokens in one forward pass. The source can be fully visible or masked, while Ω_m can cover part or all of an intermediate. For autoregressive intermediates, prediction uses teacher forcing on the teacher-generated sequence. These choices specify the conditioning available to the student without changing the three-pass procedure.

Let $\mathcal{C}$ denote the subset of intermediate modalities to which anchoring is applied. For nonempty $\mathcal{C}$, the anchoring loss is the cross-entropy against the teacher's tokens

$$\mathcal{L}_{\text{Steganography}} = \frac{1}{|\mathcal{C}|} \sum_{m \in \mathcal{C}} \frac{1}{|\Omega_m|} \sum_{i \in \Omega_m} -\log q_\theta \Big(\hat{z}_{m,i} \,\Big|\, \text{sg}\big[E_\theta(s_{\text{vis}}, \hat{z}_{\setminus\Omega})\big] \Big), \tag{4}$$

where E_θ is the student encoder, q_θ is the student's predicted distribution, and sg stops gradients at the encoder output; this detached output underlies all Pass 2 predictions, so neither the anchoring nor the reconstruction loss updates the encoder through this pass. Conditioning on preceding tokens for autoregressive predictions is implicit. We set this loss to zero when no intermediate modality is anchored. The intermediates passed to reconstruction retain teacher tokens at visible positions and use student predictions at masked positions

$$\tilde{z}_{m,i} = \begin{cases} \arg\max_v \; q_\theta(v \mid \cdot) & i \in \Omega_m, \\ \hat{z}_{m,i} & \text{otherwise.} \end{cases} \tag{5}$$

Pass 3: Source reconstruction. The student reconstructs the source from the second-pass intermediates. It receives intermediate embeddings after the prescribed masking and, where specified, visible source tokens. Only hidden source positions Ω_s contribute to the reconstruction loss

$$\mathcal{L}_{\text{Consistency}} = \frac{1}{|\Omega_s|} \sum_{i \in \Omega_s} -\log p_\theta\big(s_i \mid \tilde{e},\, s_{\setminus\Omega_s}\big), \tag{6}$$

where $\tilde{e}$ denotes the intermediate embeddings supplied to reconstruction after masking. Autoregressive decoding additionally conditions on preceding source tokens through teacher forcing, omitted from equation 6 for brevity. The visible source tokens in this pass need not match s_{vis} in pass 2. When all source positions are hidden, $s_{\setminus\Omega_s}$ is empty and the encoder receives only intermediate representations.

A straight-through estimator (Bengio et al., 2013) propagates reconstruction gradients through the student's discrete intermediate predictions (§ A.3). The forward pass uses the selected token embeddings; gradients to the prediction logits flow through probability-weighted embeddings at student-predicted positions. Teacher-generated token indices remain fixed. Thus, reconstruction trains both the return prediction and the student-generated intermediates.

The cycle objective combines reconstruction with teacher anchoring

$$\mathcal{L}_{\text{Cycle}} = \mathcal{L}_{\text{Consistency}} + \beta\, \mathcal{L}_{\text{Steganography}},$$

where β weights the anchoring term. This objective uses the observed source as its reconstruction target and teacher predictions as its anchoring targets, so neither term requires a paired counterpart. The cycle configuration sets the intermediate modalities, masking patterns, and anchored subset $\mathcal{C}$.

3.3 Instantiation in 4M

We instantiate `unimo-x` with the shared encoder-decoder backbone of 4M (Mizrahi et al., 2023). Each input token combines a content embedding, a fixed sin-cos positional embedding, and a learned modality embedding. The encoder attends jointly to visible tokens from all input modalities. Decoder blocks combine self-attention, cross-attention to the encoder output, and feed-forward layers; decoder self-attention is restricted to tokens of the same modality. Queries for non-autoregressive prediction combine a learned mask token with positional and modality embeddings. Modality-specific output heads are tied to their decoder embeddings. This shared backbone supports the generation, prediction, and reconstruction passes, with EMA weights used for teacher generation. The chosen tokenizers, decoding rules, cycle directions, and training settings are described together in § 4.1.

4 EMPIRICAL EVALUATION

We evaluate how effectively `unimo-x` learns from predominantly unpaired corpora through image generation, captioning, retrieval, and caption-based classification. Controlled comparisons vary the paired fraction and unpaired data source, while ablations examine the role of teacher anchoring in preserving meaningful intermediate predictions.

4.1 EXPERIMENTAL SETUP

Data and pairing. We train on CC12M (Changpinyo et al., 2021), comprising approximately 12 million image-caption pairs. A fraction $\pi \in \{10\%, 17.5\%, 25\%\}$ retains both modalities; the remainder is divided equally into image-only and caption-only samples by discarding the counterpart. The paired and unpaired pools are disjoint. At $\pi = 25\%$, we replace the unpaired pool with approximately 9 million CC-BY samples from CommonCatalog (Gokaslan et al., 2024), split in the same way, to test learning across corpora. We evaluate generation and retrieval on COCO (Lin et al., 2014) and caption-based classification on ImageNet (Russakovsky et al., 2015). Dataset construction and preprocessing are detailed in § A.6.

Model and training. We instantiate `unimo-x` with a 199M-parameter 4M-B backbone (Mizrahi et al., 2023), trained from scratch at 224×224 resolution with $M = 3$ modality streams: captions, RGB tokens, and image-derived CLIP representations. Frozen 4M tokenizers produce discrete RGB DiVAE tokens and quantized CLIP ViT-B/16 features (Bachmann et al., 2024; Radford et al., 2021). CLIP features are derived from each observed image and require no additional paired measurement, but inherit image-text supervision from CLIP pretraining. Our paired fractions therefore describe the training corpus, not the supervision used to pretrain the frozen components. Raw image patches are also available as encoder inputs. Captions use a frozen Flan-T5-XL encoder (Chung et al., 2024) followed by a trainable LayerNorm-Linear-LayerNorm projection. Decoder self-attention is bidirectional for image-like tokens and causal for captions.

For unpaired captions, the cycle follows caption $\rightarrow$ CLIP $\rightarrow$ RGB $\rightarrow$ caption, anchoring both image-like intermediates with $\beta = 0.5$. The student predicts masked intermediate tokens from the full caption, then reconstructs the caption from masked copies of the completed intermediates without source tokens visible to the encoder. For unpaired images, the cycle follows RGB $\rightarrow$ caption $\rightarrow$ RGB. The student predicts the caption from a masked image using teacher forcing, then reconstructs the image from its argmax caption and a masked source hint; this direction uses reconstruction alone. Default runs introduce cycle training after approximately 118B tokens of paired-only training at the same π. Architecture and training settings are given in § A.1 and Table 4; the three-pass procedure and gradient estimator are detailed in § A.3.

Comparisons and evaluation. We compare against paired-only training at each π, a no-cycle control using unpaired data for unimodal masked modeling, and fully paired training. We also evaluate released 4M-B weights (Mizrahi et al., 2023) and ITIT (Li et al., 2024b), which learns through image-text reconstruction cycles with direction-specific decoders. Our paired-only, no-cycle, and cycle configurations share the same frozen pretrained components, holding inherited supervision fixed across these comparisons. ITIT uses the same tokenizers and text encoder, while released 4M-B shares the RGB and CLIP tokenizers. Released 4M-B was trained on fully paired CC12M for 500B tokens (Mizrahi et al., 2023), about three times our 150B-token budget, which accounts for its gap to our fully paired run in Table 1. Baseline details appear in § A.4 and Table 6. Runs are compared by gradient-bearing training tokens, including student-predicted intermediates. This excludes teacher sampling and does not equate total computational cost.

We measure image generation with Fréchet Inception Distance (FID) and CLIPScore on 30,000 COCO validation samples, and captioning and retrieval with CIDEr and text-to-image Recall@1 on the 5,000-image Karpathy test split. Retrieval compares model-generated CLIP representations. Caption-based classification on a 100-class ImageNet subset measures the object-class information preserved in generated text. Generation follows caption $\rightarrow$ CLIP $\rightarrow$ RGB for images and RGB $\rightarrow$ CLIP $\rightarrow$ caption for text. Sampling settings are provided in § A.2 and Table 5; metric definitions and evaluation protocols are in § A.5.

Table 1: COCO results. For each of our runs, we select the checkpoint with the best FID within 150B training tokens on which gradients were calculated and report all metrics at that checkpoint.

Model	π	Cycle	FID ↓	CLIPScore ↑	CIDEr ↑	T2I R@1 ↑
4M-B (Mizrahi et al., 2023)	100%	✗	**40.56**	**19.92**	0.148	**6.98**
`unimo-x` *only paired*	100%	✗	42.79	19.02	0.144	6.74
ITIT (Li et al., 2024b)	25%	✗	46.95	17.63	0.023	—
ITIT (Li et al., 2024b)	25%	✓	45.20	18.90	0.023	—
`unimo-x` *only paired*	10%	✗	49.15	18.07	0.088	4.84
`unimo-x` *only paired*	17.5%	✗	47.94	18.45	0.113	5.10
`unimo-x` *only paired*	25%	✗	45.15	18.87	**0.171**	6.38
`unimo-x` *no cycle*	10%	✗	49.29	18.24	0.085	4.32
`unimo-x` *no cycle*	17.5%	✗	47.23	18.09	0.111	4.64
`unimo-x` *no cycle*	25%	✗	44.95	18.61	0.155	6.00
`unimo-x` *with cycle*	10%	✓	48.58	18.23	0.103	4.70
`unimo-x` *with cycle*	17.5%	✓	43.72	18.59	0.117	5.00
`unimo-x` *with cycle*	25%	✓	40.90	18.96	0.135	5.86

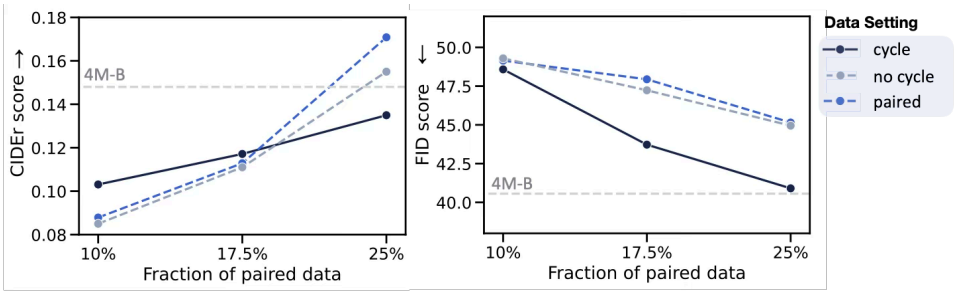


Figure 2: FID and CIDEr across paired fractions π. Cross-modal self-supervision improves FID over paired-only and no-cycle training, with the largest gains at 17.5% and 25% pairing. With 17.5% pairing, FID is lower than for paired-only training at 25%.

4.2 Results

Cross-modal self-supervision makes unpaired data useful. Adding unpaired data through unimodal masked modeling leaves performance largely unchanged, whereas cross-modal self-supervision improves image generation (Figure 2). The *no-cycle* control uses the same unpaired pool at the same sampling rate, comparing unimodal masked modeling with the full anchored cycle objective. For an unpaired image, $\mathcal{I}$ and $\mathcal{T}$ contain only RGB tokens; for an unpaired caption, they contain only caption tokens. FID gains are largest at 17.5% and 25% pairing. At 25% pairing, `unimo-x` achieves FID 40.90, improving on our fully paired baseline (42.79) and approaching released 4M-B (40.56) with four times fewer paired samples; qualitative examples appear in Figure 8. In contrast, text-to-image retrieval in the model's own CLIP space is slightly lower than for paired-only training at every paired fraction (by 0.1 to 0.5 R@1 on CC12M and by 1.5 with CommonCatalog), a decrease we do not yet explain. Captioning results are examined below.

Introducing unpaired data at approximately 27B tokens also improves performance, although the later starting point gives the best final results (§ B.1, Figure 4). Gains over paired-only training persist both with and without an EMA teacher (§ B.2, Figure 5).

Gains extend to unpaired data from a different corpus. Replacing the unpaired CC12M pool with CommonCatalog preserves the image-generation gain. At 25% pairing, cross-modal self-supervision improves FID to 42.7, compared with 45.7 for the no-cycle control and 45.2 for paired-only training (Table 2). The 3.0 FID gain over the control follows the in-distribution CC12M trend (45.0 → 40.9). As with CC12M, CIDEr ends below paired-only training (0.128 versus 0.171) while the ImageNet probes in Table 3 improve, so the lower score reflects reduced agreement with COCO references rather than lost object-class information.

Table 2: **Unpaired training across corpora.** With 25% of CC12M paired, adding unpaired CommonCatalog data through cross-modal self-supervision improves FID over paired-only and no-cycle baselines.

Model	π	Cycle	FID ↓	CLIPScore ↑	CIDEr ↑	T2I R@1 ↑
`unimo-x` *only paired*	25%	✗	45.15	18.87	**0.171**	**6.38**
`unimo-x` *no cycle (CommonCatalog)*	25%	✗	45.73	18.50	0.147	**6.38**
`unimo-x` *with cycle (CommonCatalog)*	25%	✓	**42.72**	**18.61**	0.128	4.88

Teacher anchoring constrains intermediate predictions. Consecutive high-capacity intermediates with matching spatial dimensions can carry private reconstruction codes at the expense of semantic content. Their discrete VQ-VAE representation prevents direct application of continuous pixel-space perturbation diagnostics. We evaluate teacher anchoring with three complementary diagnostics: teacher-student semantic drift, the reconstruction penalty from substituting teacher intermediates, and semantic fidelity to a reference representation. Token replacement and embedding noise, low-pass filtering, and distance-preserving losses did not eliminate the shortcuts in our experiments. In particular, smoothing removed fine stripes but broader artifacts emerged, motivating direct supervision of intermediate content through teacher anchoring (§ A.3).

Measuring steganography. We study the cycle `caption` → (`clip` → `rgb`) → `caption`. Let x denote the source caption, c_1 and c_2 the pass-1 teacher and pass-2 student CLIP intermediates, r_2 the pass-2 RGB intermediate, and c_{GT} the CLIP representation of the corresponding ground-truth image. Teacher-student semantic drift is measured after decoding CLIP tokens into feature maps, using cosine distances averaged over corresponding spatial positions (§ A.5)

$$d_{\mathrm{sem}}(a,b) = \frac{1}{196}\sum_{p=1}^{196}\left[1-\cos(D_{\mathrm{CLIP}}(a)_p, D_{\mathrm{CLIP}}(b)_p)\right], \qquad \delta_{\mathrm{sem}} = \frac{d_{\mathrm{sem}}(c_1,c_2)}{d_{\mathrm{rnd}}},$$

where D_{CLIP} is the frozen 4M CLIP VQ-VAE decoder, p indexes spatial positions, and d_{rnd} is the average of this distance over random sample pairs. A value of $\delta_{\mathrm{sem}} \approx 1$ indicates a distance comparable to the random-pair baseline; smaller values indicate closer agreement between teacher and student intermediates. Semantic drift alone, however, does not establish steganography.

We next test whether reconstruction depends specifically on the student-generated CLIP intermediate. Keeping r_2 fixed, we replace c_2 with c_1 and measure the change in caption reconstruction loss

$$\begin{aligned}\Delta\mathrm{CE} = \mathrm{CE}(x, p_\theta(\cdot \mid E_\theta(r_2, c_1))) \\ - \mathrm{CE}(x, p_\theta(\cdot \mid E_\theta(r_2, c_2)))\,.\end{aligned}$$

Here, CE denotes caption reconstruction cross-entropy. A positive gap indicates that reconstruction benefits from the student intermediate, whereas $\Delta\mathrm{CE} \approx 0$ indicates little effect of the substitution.

Because a substitution gap alone need not imply a private code, we also measure the semantic fidelity of c_2 to the reference CLIP representation

$$1 - \epsilon^{(2)}_{\mathrm{sem}} = 1 - \frac{d_{\mathrm{sem}}(c_2, c_{\mathrm{GT}})}{d_{\mathrm{rnd}}}.$$

A value of 1 indicates maximal cosine similarity, while 0 corresponds to the average random-pair distance. Since one caption can describe multiple valid images, we interpret fidelity relative to the paired-only model. We evaluate the anchored and anchor-free models on the same 512 validation examples at four checkpoints.

Figure 3 tracks these diagnostics in the early-fork setting, where cycle training begins at approximately 27B tokens. Without anchoring, teacher-student semantic drift reaches $\delta_{\mathrm{sem}} = 0.92$ at 32B tokens, while student intermediates yield increasingly lower reconstruction losses than teacher intermediates and semantic fidelity declines. Together, these observations indicate that reconstruction increasingly relies on student-specific information that does not preserve the intended semantics. Teacher anchoring suppresses this measured shortcut behavior, keeping the reconstruction gap near zero and fidelity close to the paired-only reference. Further analysis in § B.3 documents the associated downstream degradation (Figure 6) and visualizes the deterioration of unanchored CLIP intermediates during training (Figure 7).

Table 3: **Caption-based classification on ImageNet.** Top-1 accuracy (%) on 100 classes. Classifiers are trained on generated captions or class-name prompts and tested on generated captions.

Paired fraction	Training setting	Caption-trained probe (CLIP)		Prompt-trained transfer	
		10 per class	64 per class	CLIP	MiniLM
10%	`unimo-x` *only paired*	13.37	18.04	12.74	9.54
	`unimo-x` *with cycle*	**17.75**	**24.70**	**17.00**	**13.72**
17.5%	`unimo-x` *only paired*	18.47	23.52	13.74	11.46
	`unimo-x` *with cycle*	**20.51**	**28.60**	**20.30**	**16.08**
25%	`unimo-x` *only paired*	20.08	27.92	17.40	13.14
	`unimo-x` *with cycle*	**26.07**	**33.26**	**24.24**	**17.82**
	`unimo-x` *with cycle (CommonCatalog)*	25.33	32.52	23.72	17.62

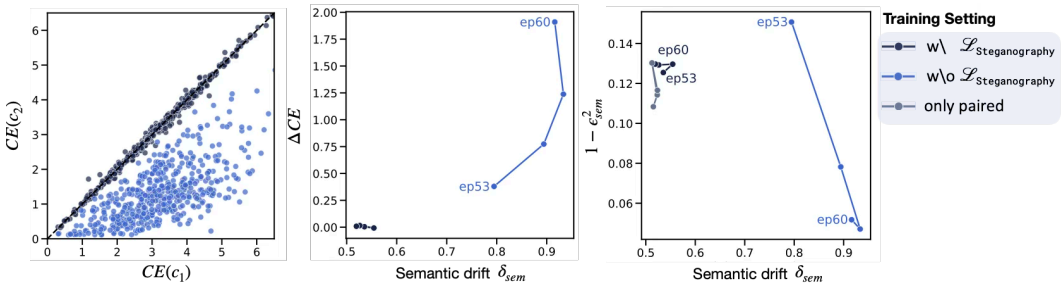


Figure 3: **Teacher anchoring mitigates steganographic shortcuts.** Left: caption reconstruction losses using teacher (c_1) or student (c_2) CLIP intermediates, with RGB fixed; the dashed line marks equal losses. Middle: without anchoring, semantic drift grows alongside the reconstruction advantage ΔCE of student intermediates. Right: their semantic fidelity declines, while anchoring maintains fidelity near the paired-only reference.

Generated captions preserve more object-class information. We assess whether generated captions retain information useful for object classification, complementing reference-based caption evaluation. We use a random 100-class ImageNet-1K subset, with the data split and probe protocols specified in § A.5. Caption-trained linear probes use frozen CLIP text embeddings with 10 or 64 examples per class (Table 3). Accuracy improves at both training-set sizes and every paired fraction, including with CommonCatalog. At 10% pairing, accuracy with 64 examples per class rises from 18.04% to 24.70%.

The last two columns test classifiers trained on fixed class-name prompts and evaluated on generated captions, preventing them from learning model-specific phrasing during training. Accuracy improves with both CLIP and MiniLM embeddings (Wang et al., 2020; Reimers & Gurevych, 2019), showing that gains extend beyond the CLIP space used during cycle training. CIDEr improves at 10% and 17.5% pairing but decreases at 25%, where every classification protocol improves. Lower reference-caption agreement thus coexists with better preservation of object-class information.

5 Conclusion

We introduced `unimo-x`, a cross-modal self-supervision framework for training multimodal Transformers on predominantly unpaired corpora. Combining sparse paired supervision with teacher-anchored reconstruction improves image generation and preserves more object-class information in captions, including when unpaired data come from a different corpus. With four times fewer paired samples, `unimo-x` matches the image-generation quality of fully paired training. Teacher anchoring mitigates steganographic shortcuts and may more broadly help cycle-consistent models preserve meaningful intermediate representations. Since the framework requires only modality-specific tokenizers, a shared backbone, and a small paired set, it opens multimodal learning to domains where scarce pairing has so far ruled it out. Further work should examine scaling with model size and modality diversity. A compelling direction is learning across heterogeneous biomedical corpora, connecting molecular, tissue, and clinical evidence to model relationships across biological scales.

AI use statement. We used generative AI to assist with coding, but not for synthetic data generation, hypothesis generation, or method development. Theoretical modeling and proofs are not applicable. Additionally, we used AI to proofread the manuscript. We reviewed all AI-assisted text and code and take responsibility for the final content.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a Visual Language Model for Few-Shot Learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. MultiMAE: Multi-modal Multi-task Masked Autoencoders. In *European Conference on Computer Vision*, pp. 348–367. Springer, 2022.

Roman Bachmann, Oğuzhan F Kar, David Mizrahi, Ali Garjani, Mingfei Gao, David Griffiths, Jiaming Hu, Afshin Dehghan, and Amir Zamir. 4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities. *Advances in Neural Information Processing Systems*, 37:61872–61911, 2024.

Lei Bai, Zhongrui Cai, Yuhang Cao, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, et al. Intern-S1: A Scientific Multimodal Foundation Model. *arXiv preprint arXiv:2508.15763*, 2025.

Dina Bashkirova, Ben Usman, and Kate Saenko. Adversarial Self-Defense for Cycle-Consistent GANs. In *Advances in Neural Information Processing Systems*, 2019.

Sagie Benaim and Lior Wolf. One-Sided Unsupervised Domain Mapping. In *Advances in Neural Information Processing Systems*, 2017.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. *arXiv preprint arXiv:1308.3432*, 2013.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3557–3567. IEEE, 2021.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325*, 2015.

Changxi Chi, Yufei Huang, Jun Xia, Jiangbin Zheng, Yunfan Liu, Zelin Zang, and Stan Z Li. Departures: Distributional Transport for Single-Cell Perturbation Prediction with Neural Schrödinger Bridges. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20544–20552, 2026.

Ziheng Chi, Yifan Hou, Chenxi Pang, Shaobo Cui, Mubashara Akhtar, and Mrinmaya Sachan. Chimera: Diagnosing Shortcut Learning in Visual-Language Understanding. *arXiv preprint arXiv:2509.22437*, 2025.

Casey Chu, Andrey Zhmoginov, and Mark Sandler. CycleGAN, a Master of Steganography. *NeurIPS Workshop on Machine Deception*, 2017.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Haotian Cui, Alejandro Tejada-Lapuerta, Maria Brbić, Julio Saez-Rodriguez, Simona Cristea, Hani Goodarzi, Mohammad Lotfollahi, Fabian J Theis, and Bo Wang. Towards multimodal foundation models in molecular cell biology. *Nature*, 640(8059):623–633, 2025.

Valentin De Bortoli, Iryna Korshunova, Andriy Mnih, and Arnaud Doucet. Schrodinger Bridge Flow for Unpaired Data Translation. *Advances in Neural Information Processing Systems*, 37: 103384–103441, 2024.

Maurice Diesendruck, Jianzhe Lin, Shima Imani, Gayathri Mahalingam, Mingyang Xu, and Jie Zhao. Learning How To Ask: Cycle-Consistency Refines Prompts in Multimodal Foundation Models. *arXiv preprint arXiv:2402.08756*, 2024.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. PaLM-E: An Embodied Multimodal Language Model. *arXiv preprint arXiv:2303.03378*, 2023.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal Cycle-Consistency Learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1801–1810. IEEE, 2019.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind One Embedding Space To Bind Them All. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190. IEEE, 2023.

Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised Monocular Depth Estimation with Left-Right Consistency. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6602–6611. IEEE, 2017.

Aaron Gokaslan, A Feder Cooper, Jasmine Collins, Landan Seguin, Austin Jacobson, Mihir Patel, Jonathan Frankle, Cory Stephenson, and Volodymyr Kuleshov. Common Canvas: Open Diffusion Models Trained on Creative-Commons Images. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8250–8260. IEEE, 2024.

Sharut Gupta, Shobhita Sundaram, Chenyu Wang, Stefanie Jegelka, and Phillip Isola. Better Together: Leveraging Unpaired Multimodal Data for Stronger Unimodal Models. In *International Conference on Learning Representations*, volume 2026, pp. 73970–74025, 2026.

Xing Han, Hsing-Huan Chung, Joydeep Ghosh, Paul Pu Liang, and Suchi Saria. Massively Multimodal Foundation Models: A Framework for Capturing Interactions with Specialized Mixture-of-Experts. *arXiv preprint arXiv:2509.25678*, 2025.

Katherine Hermann, Hossein Mobahi, Thomas Fel, and Michael Mozer. On the Foundations of Shortcut Learning. In *International Conference on Learning Representations*, volume 2024, pp. 43832–43868, 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.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language Is Not All You Need: Aligning Perception with Language Models. *Advances in Neural Information Processing Systems*, 36: 72096–72109, 2023.

Andrej Karpathy and Li Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137, 2015.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding Language Models to Images for Multimodal Inputs and Outputs. In *International Conference on Machine Learning*, pp. 17283–17300. PMLR, 2023.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-Based & Neural Unsupervised Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5039–5049, 2018.

Taekyung Lee, Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Unpaired Point Cloud Completion via Unbalanced Optimal Transport. In *Forty-second International Conference on Machine Learning*, 2025.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal Foundation Models: From Specialists to General-Purpose Assistants. *Foundations and Trends in Computer Graphics and Vision*, 16(1-2):1–214, 2024a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning*, pp. 19730–19742. PmLR, 2023.

Tianhong Li, Sangnie Bhardwaj, Yonglong Tian, Han Zhang, Jarred Barber, Dina Katabi, Guillaume Lajoie, Huiwen Chang, and Dilip Krishnan. Leveraging Unpaired Data for Vision-Language Generative Models via Cycle Consistency. In *International Conference on Learning Representations*, volume 2024, pp. 34254–34274, 2024b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly Mapping from Image to Text Space. *arXiv preprint arXiv:2209.15162*, 2022.

David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4M: Massively Multimodal Masked Modeling. *Advances in Neural Information Processing Systems*, 36:58363–58408, 2023.

Horia Porav, Valentina Musat, and Paul Newman. Reducing Steganography In Cycle-consistency GANs. In *CVPR Workshops*, pp. 78–82, 2019.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*, 2022.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Jules Samaran, Gabriel Peyré, and Laura Cantini. scConfluence: single-cell diagonal integration with regularized Inverse Optimal Transport on weakly connected features. *Nature Communications*, 15(1):7762, 2024.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86–96, 2016.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel HAZIZA, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired Multi-Domain Causal Representation Learning. *Advances in Neural Information Processing Systems*, 36:34465–34492, 2023.

Chameleon Team. Chameleon: Mixed-Modal Early-Fusion Foundation Models. *arXiv preprint arXiv:2405.09818*, 2024.

Subash Timilsina, Sagar Shrestha, and Xiao Fu. Identifiable Shared Component Analysis of Unpaired Multimodal Mixtures. *Advances in Neural Information Processing Systems*, 37:127799–127833, 2024.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575, 2015.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788, 2020.

Johann Wenckstern, Eeshaan Jain, Benedikt von Querfurth, Yexiang Cheng, Kiril Vasilev, Matteo Pariset, Phil F. Cheng, Petros Liakopoulos, Olivier Michielin, Andreas Wicki, Gabriele Gut, and Charlotte Bunne. The Virtual Tissues foundation model resolves spatial proteomics across scales. *Nature*, 657:744–754, 2026.

Johnny Xi, Jana Osea, Zuheng Xu, and Jason Hartford. Propensity Score Alignment of Unpaired Multimodal Data. *Advances in Neural Information Processing Systems*, 37:141103–141128, 2024.

Sihan Xu, Ziqiao Ma, Yidong Huang, Honglak Lee, and Joyce Chai. CycleNet: Rethinking Cycle Consistency in Text-Guided Diffusion for Image Manipulation. *Advances in Neural Information Processing Systems*, 36:10359–10384, 2023.

Yang Xu, Edmon Begoli, and Rachel Patton McCord. sciCAN: single-cell chromatin accessibility and gene expression data integration via cycle-consistent adversarial network. *npj Systems Biology and Applications*, 8(1):33, 2022.

Siyuan Yan, Zhen Yu, Clare Primiero, Cristina Vico-Alonso, Zhonghua Wang, Litao Yang, Philipp Tschandl, Ming Hu, Lie Ju, Gin Tan, et al. A multimodal vision foundation model for clinical dermatology. *Nature Medicine*, 31(8):2691–2702, 2025.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-Shot Transfer with Locked-image text Tuning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18102–18112. IEEE, 2022.

Zirui Zhang, Haoyu Dong, Kexin Pei, and Chengzhi Mao. R-C2: Cycle-Consistent Reinforcement Learning Improves Multimodal Reasoning. *arXiv preprint arXiv:2603.25720*, 2026.

Zizhao Zhang, Lin Yang, and Yefeng Zheng. Translating and Segmenting Multimodal Medical Volumes with Cycle-and Shape-Consistency Generative Adversarial Network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9242–9251. IEEE, 2018.

Rui Zhao, Weijia Mao, and Mike Zheng Shou. DoraCycle: Domain-Oriented Adaptation of Unified Generative Model in Multimodal Cycles. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2835–2846. IEEE, 2025.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT Pre-training with Online Tokenizer. In *International Conference on Learning Representations*, 2022.

Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. Learning Dense Correspondence via 3D-guided Cycle Consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 117–126, 2016.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251. IEEE, 2017.

APPENDIX
UNPAIRED MULTIMODAL LEARNING WITH CROSS-MODAL SELF-SUPERVISION

A EXPERIMENTAL DETAILS

A.1 ARCHITECTURE AND TRAINING

The architecture follows 4M-B (§ 3.3). Unless stated otherwise, runs share the architecture and training settings in Table 4, with the paired fraction and unpaired data source varied as specified. In the main experiments, cycle and no-cycle runs resume from the paired-only checkpoint at approximately 118B tokens and continue its learning-rate schedule. Training uses four nodes with four NVIDIA GH200 GPUs each, for a total of 16 GPUs per run. During joint training, each GPU processes 128 paired and 96 unpaired samples per update. Reaching the paired-only checkpoint takes approximately 24 hours; subsequent cycle training takes another 45 hours, giving roughly three days per full run. Early-fork ablations instead resume from approximately 27B tokens, reached after six hours, and train for about one additional day.

Table 4: Default architecture and training hyperparameters; ablation-specific changes are stated in the text.

Hyperparameter	**Setting**
Model	
Backbone	4M-B: 12 encoder and 12 decoder blocks, width 768, 12 heads, SwiGLU, 199M parameters
RGB	224^2 pixels (input), 196 DiVAE tokens, 16k codebook
CLIP	196 VQ-VAE tokens of CLIP ViT-B/16 features, 8k codebook
Caption	T5 vocabulary, at most 64 tokens; frozen Flan-T5-XL encoder
Optimization	
Optimizer	AdamW, $\beta = (0.9,\ 0.95)$, weight decay 0.05
Learning rate	Peak 2×10^{-4}; final 5×10^{-6}
Schedule	Linear warm-up over 4B tokens; cosine decay over 300B tokens
Training length	$\sim$150B tokens
Batch per update	$2{,}048$ paired + $1{,}536$ unpaired samples
Gradient clipping	Norm 3
Precision	`bfloat16`
Paired masking (as in 4M)	
Input / target tokens	128 / 128
Input Dirichlet α	0.5 per modality
Target Dirichlet α	0.5 for RGB and CLIP tokens; 0.05 for captions
Cycle training	
Unpaired pool	Equal proportions of image-only and caption-only samples
Source per update	Image or caption, sampled with equal probability
Cycle weight $\lambda(t)$	$0 \to 1$ over the first $1{,}000$ cycle updates
Image mask ratio	$\mathcal{N}(0.75,\ 0.25)$, clipped to $[0.5,\ 1]$; capped at 0.75 when the image conditions a caption
Label smoothing	0.1
Teacher EMA decay	0.999; initialized from the student at the fork
Teacher sampling (pass 1)	
Caption $\to$ CLIP	49 ROAR steps; temperature 6 (decaying); CFG 3; top-p 0.8
Caption, CLIP $\to$ RGB	25 ROAR steps; temperature 3 (decaying); CFG 3; top-p 0.8
Image $\to$ caption	Temperature 0.7; top-p 0.9; no CFG

A.2 GENERATION

Both generation directions use CLIP as an intermediate representation. Text-to-image generation follows the released 4M configuration (Mizrahi et al., 2023). Image-to-text generation uses 4M's RGB-to-CLIP settings, followed by caption decoding with beam search, as in ITIT (Li et al., 2024b). Table 5 lists the settings used for `unimo-x` and the released 4M-B.

Table 5: Generation settings used for evaluation.

Generation stage	Settings
Text-to-image	
Caption → CLIP	49 ROAR steps; temperature 6 (decaying); CFG 3; top-p 0.9
Caption, CLIP → RGB	25 ROAR steps; temperature 0.5; CFG 3; top-p 0.9
RGB decoding	DiVAE diffusion decoder; 50 steps
Image-to-text	
RGB → CLIP	1 step; temperature 0.01; CFG 2
RGB, CLIP → caption	Beam search; width 4

A.3 TRAINING ALGORITHM

The three-pass cycle procedure is summarized in Algorithm 1. Below, we describe the straight-through estimator and alternative defenses against steganographic shortcuts.

Straight-through estimator. The second-pass intermediates $\tilde{z}$ combine student argmax predictions with teacher tokens retained at visible positions. The discrete student predictions are not differentiable. Therefore, we pass them to the reconstruction step through a straight-through estimator (Bengio et al., 2013; Li et al., 2024b). For an intermediate modality m with input embedding table E_m and second-pass logits $\ell_{m,i}$ at position i, the reconstruction step receives

$$\tilde{e}_{m,i} = E_m[\tilde{z}_{m,i}] + \alpha_m \, \mathbb{1}[i \in \Omega_m] \left(\bar{e}_{m,i} - \mathrm{sg}[\bar{e}_{m,i}]\right), \qquad \bar{e}_{m,i} = \sum_v \mathrm{softmax}(\ell_{m,i})_v \, E_m[v]. \quad (7)$$

Here, sg denotes stop-gradient. Since $\bar{e}_{m,i} - \mathrm{sg}[\bar{e}_{m,i}]$ is zero in value, the forward pass uses exactly the embedding of the selected token. The surrogate backward path differentiates the probability-weighted average embedding. The indicator restricts this path to student-predicted positions; teacher token indices retained at visible positions remain fixed (Equation 5). For a caption intermediate, Ω_m covers all positions. For image tokens, E_m is the 4M input embedding table (Mizrahi et al., 2023), and $\alpha_m = 0.5$ scales the gradient flowing back to the intermediate predictions. For captions, E_m is the frozen T5 input embedding table, $\alpha_m = 1$, and $\tilde{e}$ is passed through the frozen T5 encoder.

Alternative defenses against steganography. Prior defenses against steganography in cycle models target continuous pixel outputs, where hidden signals can take the form of small, high-frequency perturbations (Chu et al., 2017; Bashkirova et al., 2019; Porav et al., 2019). These interventions do not directly address shortcuts encoded through discrete token choices. DoraCycle (Zhao et al., 2025) reports a related collapse into self-consistent but irrelevant captions and addresses it with gradient surgery. We tested three alternatives to teacher anchoring, which are outlined in the following paragraphs. None prevented the intermediates from developing private codes for RGB reconstruction. Some delayed the failure by a few thousand updates, but reconstruction loss continued to decrease as intermediate predictions deteriorated.

Noise. Following Bashkirova et al. (2019), we randomly replaced a fraction of the intermediate tokens and added Gaussian noise to their embeddings before the third pass. Noise can disrupt a small pixel-space signal, but in our case the code is carried by the choice of discrete codebook indices and produces visible stripes in the decoded intermediate. Perturbing a subset of tokens did not eliminate the shortcut, consistent with the finding that private codes are distributed across multiple tokens; stronger perturbations risk corrupting the intermediate content itself.

Algorithm 1 Three-pass anchored cycle step (unpaired sample).

Input: Student θ consisting of the context encoder E_θ and decoder D_θ, EMA teacher $\bar{\theta}$; a single-modality source sample s tokenized as $t^{(s)}$; chain $\mathcal{T}(s) = (m_1, \ldots, m_K)$.
Output: Cycle loss l_{Cycle}

M_x^p denotes the positions of x hidden in pass p, and $\bar{M}_x^p$ the visible ones. For a caption intermediate, $\bar{M}_m^2 = \emptyset$ and D_θ is teacher-forced on $\hat{t}^{(m)}$.

Pass 1: Teacher generation.

$\mathcal{C} \leftarrow \{t^{(s)}\}$ ▷ source fully visible
for $m \in \mathcal{T}(s)$ **do** ▷ chained generation
  $\hat{t}^{(m)} \leftarrow \text{Generate}_{\bar{\theta}}(m \mid \mathcal{C})$ ▷ 4M sampler
  $\mathcal{C} \leftarrow \mathcal{C} \cup \{\hat{t}^{(m)}\}$
end for

Pass 2: Student prediction and anchoring.

$X_2 \leftarrow t^{(s)}_{\bar{M}_s^2} \cup \{\hat{t}^{(m)}_{\bar{M}_m^2}\}_{m \in \mathcal{T}(s)}$ ▷ encoder input
$\{z^{(m)}\}_{m \in \mathcal{T}(s)} \leftarrow D_\theta\big(\text{sg}[E_\theta(X_2)]\big)$ ▷ one pass for all m; no gradients for encoder
$l_{\text{Steganography}} \leftarrow 0$
for $m \in \mathcal{T}(s)$ **do**
  $\tilde{t}^{(m)} \leftarrow \hat{t}^{(m)};\ \tilde{t}^{(m)}[M_m^2] \leftarrow \arg\max z^{(m)}[M_m^2]$ ▷ Equation 5
  if `is_spatial`(m) **then**
    $l_{\text{Steganography}} \leftarrow l_{\text{Steganography}} + \mathcal{L}_{\text{Steganography}\, M_m^2}\big(z^{(m)}, \hat{t}^{(m)}\big)$ ▷ Equation 4
  end if
end for

Pass 3: Source reconstruction.

$e^{(m)} \leftarrow \text{ST}_m\big(\tilde{t}^{(m)}, z^{(m)}, M_m^2\big)$ for all $m \in \mathcal{T}(s)$ ▷ Equation 7
$X_3 \leftarrow \{e^{(m)}_{\bar{M}_m^3}\}_{m \in \mathcal{T}(s)} \cup t^{(s)}_{\bar{M}_s^3}$ ▷ a grid source adds a heavily masked hint; for a caption source $\bar{M}_s^3 = \emptyset$
$l_{\text{Consistency}} \leftarrow \mathcal{L}_{\text{Consistency}\, M_s^3}\big(D_\theta(E_\theta(X_3)),\ t^{(s)}\big)$ ▷ Equation 6
$l_{\text{Cycle}} \leftarrow l_{\text{Consistency}} + \beta\, l_{\text{Steganography}}$
return l_{Cycle}

Low-pass filtering. Before the third pass, we average-pooled the 14×14 grid of intermediate embeddings to 7×7 and upsampled it back to 14×14 with bilinear interpolation. This suppresses fine-scale variation in both the embeddings and their gradients. Fine stripes disappeared, but broader patterns emerged and still supported source reconstruction. Smoothing therefore changed the spatial scale of the artifacts without eliminating the shortcut.

Distance preservation. For spatially aligned image modalities, we introduced a loss that penalizes changes in the distances between source regions relative to the corresponding intermediate regions, following Benaim & Wolf (2017). Although distance correlation reached 0.98 by the end of training, stripe artifacts persisted, becoming fewer and wider than without the loss. The model satisfied the spatial constraint while retaining the shortcut.

These observations motivated teacher anchoring, which supervises intermediate predictions with the teacher's outputs for the same input. Deviations from these targets are penalized even when they improve source reconstruction.

A.4 Baselines

ITIT (Li et al., 2024b). ITIT jointly learns image-to-text and text-to-image generation from paired and unpaired data. Its architecture combines a shared image-text encoder with separate autoregressive text and parallel image decoders. Paired examples provide direct cross-modal supervision, while unpaired samples support image-text-image and text-image-text reconstruction cycles. To

Table 6: Comparison of published ITIT, our matched reimplementation, and `unimo-x`. Our reimplementation of ITIT and `unimo-x` use the same tokenizers, data, and training-token budget in this comparison.

	ITIT (Li et al., 2024b)	**ITIT** (our reimplementation)	**`unimo-x`** (our method)
Architecture	Two direction-specific decoders	Two direction-specific decoders	Shared multimodal backbone
Image tokenizer	VQGAN (WebLI)	DiVAE / `tok_rgb`	DiVAE / `tok_rgb`
Tokens / vocab.	256 / 8,192	196 / 16,384	196 / 16,384
Resolution	256×256	224×224	224×224
Text encoder	T5-XXL	Flan-T5-XL	Flan-T5-XL
Training data	CC3M + 55M images + 55M texts	CC12M, 25% paired	CC12M, 25% paired
Optimizer	Adafactor	AdamW	AdamW
Initialization	Random	Random	Random
Cycle training	After paired warm-up	After paired warm-up	After paired warm-up
Training tokens	$\sim$1.5T	$\sim$150B	$\sim$150B

limit computational overhead, iterative generation is performed without gradients, followed by a differentiable prediction step through which the reconstruction loss is backpropagated. As no public implementation was available, we reimplemented ITIT following the architecture and training procedure described in the original paper. All reported ITIT results use this reimplementation.

For the comparison in Table 6, our ITIT reimplementation and `unimo-x` share the data, paired fraction, tokenizers, training-token budget, paired warm-up, and evaluation code. A central architectural difference is the use of direction-specific decoders in ITIT versus a shared multimodal backbone in `unimo-x`. Published ITIT results additionally differ in tokenizer, resolution, text encoder, data, optimizer, and training budget, and are not directly comparable. Our conclusion is restricted to this matched evaluation, in which the ITIT reimplementation does not recover the gain obtained by `unimo-x`.

A.5 Evaluation metrics

We evaluate image generation, captioning, retrieval, and the class information preserved in generated captions. Throughout, π denotes the paired fraction of the training corpus.

Fréchet Inception Distance. FID (Heusel et al., 2017) measures the discrepancy between real and generated image distributions in Inception feature space

$$\mathrm{FID} = \|\mu_{\mathrm{real}} - \mu_{\mathrm{gen}}\|_2^2 + \mathrm{Tr}\left(\Sigma_{\mathrm{real}} + \Sigma_{\mathrm{gen}} - 2\left(\Sigma_{\mathrm{real}}^{1/2}\Sigma_{\mathrm{gen}}\Sigma_{\mathrm{real}}^{1/2}\right)^{1/2}\right), \tag{8}$$

where $\mu_{\mathrm{real}}, \Sigma_{\mathrm{real}}$ and $\mu_{\mathrm{gen}}, \Sigma_{\mathrm{gen}}$ are the empirical means and covariances of the real and generated image features, respectively. Lower values indicate closer distributional agreement. We evaluate FID on 30,000 COCO validation samples, decoding generated RGB tokens into images with the frozen DiVAE diffusion decoder and resizing them from 224^2 to 256^2, as in 4M.

CLIPScore. CLIPScore (Hessel et al., 2021) measures alignment between a generated image and its conditioning caption using their embeddings from a pretrained CLIP ViT-L/14 model (Radford et al., 2021), reported as cosine similarity $\times 100$. Higher scores indicate stronger image-text alignment. We evaluate CLIPScore on the same 30,000 COCO validation samples used for FID. Whereas FID compares image distributions, CLIPScore assesses correspondence between individual images and captions.

CIDEr. CIDEr (Vedantam et al., 2015) measures agreement between generated captions and human-written references through TF-IDF-weighted n-gram similarity. Higher scores indicate

greater agreement with the reference descriptions. We evaluate CIDEr on the 5,000-image Karpathy test split of COCO (Karpathy & Fei-Fei, 2015). Because this metric depends on lexical overlap, we complement it with caption-based classification to assess the object-class information conveyed by generated text.

Retrieval Recall@1. We evaluate cross-modal retrieval in the model's own CLIP space on the 5,000 images of the Karpathy COCO test split, pairing each image with its first reference caption. The model generates CLIP tokens independently from each caption and each image, using the corresponding generation samplers (Table 5). We decode the 196 tokens with the frozen 4M CLIP VQ-VAE decoder into a continuous CLIP ViT-B/16 feature map, average-pool over the 14×14 spatial positions, and ℓ_2-normalize the resulting vector. The dot product of the vectors for caption i and image j gives the similarity S_{ij}. Let $\mathcal{Q}$ denote the set of captions, $\mathcal{V}$ the set of images, $j^\star(i)$ the image of caption i, and $i^\star(j)$ the caption of image j. Text-to-image and image-to-text Recall@1 measure the percentage of queries whose counterpart ranks first among all candidates

$$\mathrm{R@1}_{\mathrm{t2i}} = \frac{100}{|\mathcal{Q}|} \sum_{i \in \mathcal{Q}} \mathbb{1}\left[\arg\max_{j \in \mathcal{V}} S_{ij} = j^\star(i)\right], \quad \mathrm{R@1}_{\mathrm{i2t}} = \frac{100}{|\mathcal{V}|} \sum_{j \in \mathcal{V}} \mathbb{1}\left[\arg\max_{i \in \mathcal{Q}} S_{ij} = i^\star(j)\right]. \quad (9)$$

Higher values indicate better retrieval. Because the embeddings are pooled CLIP patch features rather than CLIP's joint image-text embedding, and each image is paired with a single caption, these values are comparable across our models but not with published CLIP retrieval results. For the steganography diagnostics, we instead compute cosine distances between corresponding spatial positions of the decoded CLIP feature maps and average over positions, without spatially pooling the features.

Caption-based top-1 accuracy. Top-1 accuracy measures the percentage of images whose class is correctly predicted from their generated captions. We use a random 100-class ImageNet-1K subset, with 64 training images and 50 validation images per class, and evaluate two protocols.

For caption-trained probes, generated captions are embedded with the frozen CLIP ViT-B/16 text encoder. A linear classifier is trained using either 10 examples per class or all 64 training examples per class. For prompt-trained transfer, classifiers are trained only on fixed class-name prompts and evaluated on generated captions, using CLIP or MiniLM embeddings (Wang et al., 2020; Reimers & Gurevych, 2019). This second protocol prevents classifiers from learning model-specific phrasing from training captions.

Both protocols report accuracy on the same 5,000 validation images, with chance accuracy of 1%. Higher accuracy indicates that generated captions preserve more information useful for object classification.

Semantic drift, reconstruction gap, and semantic fidelity are defined in § 4.2.

A.6 Datasets

We use CC12M as the primary training corpus and CommonCatalog to evaluate learning from unpaired observations drawn from a different distribution. COCO supports evaluation of image generation, captioning, and retrieval, while an ImageNet subset measures the class information preserved in generated captions.

Conceptual 12M (CC12M). CC12M (Changpinyo et al., 2021) contains approximately 12 million web-sourced image-text pairs collected for vision-language pretraining. Its collection pipeline relaxes the filtering criteria of Conceptual Captions to increase coverage of diverse and long-tail visual concepts. We use CC12M to construct controlled training settings in which the fraction of paired observations varies while the underlying source corpus remains fixed.

For each paired fraction $\pi \in \{0.10, 0.175, 0.25\}$, we partition the corpus into disjoint paired and unpaired pools. The paired pool retains both modalities. The remaining samples are divided equally into image-only and caption-only subsets by discarding the corresponding caption or image, respectively. Thus, each sample in the unpaired pool contributes only one observed modality; its original counterpart is unavailable during training. A fully paired configuration with $\pi = 1$ provides a reference for the benefit of complete cross-modal supervision.

Images are processed at 224×224 resolution and tokenized using the frozen RGB DiVAE from 4M (Mizrahi et al., 2023), producing 196 tokens per image from a vocabulary of 16,384 entries. Captions use the T5 tokenizer (Raffel et al., 2020) and are truncated to 64 tokens. For samples containing an image, we additionally obtain CLIP ViT-B/16 patch features (Radford et al., 2021) and quantize them with the frozen 4M CLIP tokenizer. These features are derived from the observed image and do not require an additional paired measurement. Training data are pretokenized, and the tokenizers remain frozen.

CommonCatalog. CommonCatalog (Gokaslan et al., 2024) contains Creative-Commons-licensed Flickr images accompanied by synthetic captions generated with BLIP-2 (Li et al., 2023). We use its CC-BY subset, retaining approximately 9 million samples after conversion to our pretokenized format. This pool is roughly the same size as the unpaired 75% of CC12M that it replaces in the $\pi = 0.25$ configuration.

The paired pool remains fixed at 25% of CC12M, while CommonCatalog supplies all unpaired observations. As in the CC12M construction, half of these samples retain only the image and half retain only the caption. Images and captions use the same preprocessing and frozen tokenizers as CC12M, including the 64-token caption limit. This setting changes both the image source and the caption distribution, replacing web-sourced text with model-generated descriptions. It therefore tests whether unpaired observations remain useful when they are collected independently of the paired training corpus.

COCO. COCO (Lin et al., 2014) contains images of everyday scenes with multiple objects and human-written descriptions. Its caption annotations provide five reference captions per training and validation image (Chen et al., 2015). We use COCO to evaluate cross-modal generation and retrieval.

Text-to-image generation uses 30,000 COCO validation samples, evaluated with FID and CLIP-Score. Captioning and retrieval use the 5,000-image Karpathy test split (Karpathy & Fei-Fei, 2015), evaluated with CIDEr and Recall@1. Metric definitions and retrieval protocols are given in § A.5; generation settings appear in Table 5.

ImageNet. ImageNet-1K (Russakovsky et al., 2015) provides a complementary evaluation of whether generated captions retain information about object identity. We select a random subset of 100 classes, with 64 training images and 50 validation images per class, yielding 6,400 training and 5,000 validation images. Each model generates captions for all 11,400 images. The caption-based classification protocols are detailed in § A.5.

B Ablations

B.1 Introduction of unpaired data

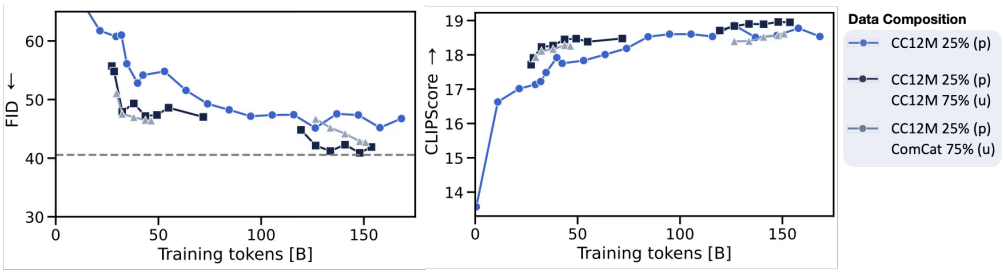


Figure 4: Effect of starting cycle training at 27B or 118B training tokens, using unpaired data from CC12M or CommonCatalog. *p* and *u* denote paired and unpaired samples, respectively.

Cycle starting point. At $\pi = 25\%$, we introduce cycle training after either 27B tokens, while the paired-only model is still improving rapidly, or 118B tokens (Figure 4). The early fork improves the model almost immediately. At 32B tokens, FID drops from 61.0 for paired-only training to 47.9 and CLIPScore rises from 17.2 to 18.2, levels that paired-only training reaches only after about 95B

and 75B tokens, respectively. The early gain then plateaus around an FID of 47. The late fork starts from a stronger paired model and yields the best results overall with an FID of 40.9, within 0.4 of the released 4M-B trained on fully paired data, and a CLIPScore of 19.0. Using CommonCatalog as the unpaired source gives similar results, on par with CC12M for the early fork and slightly behind it for the late fork (FID 42.7, CLIPScore 18.6). An early start accelerates the initial gains, while the later start achieves the best final performance.

B.2 Effect of EMA

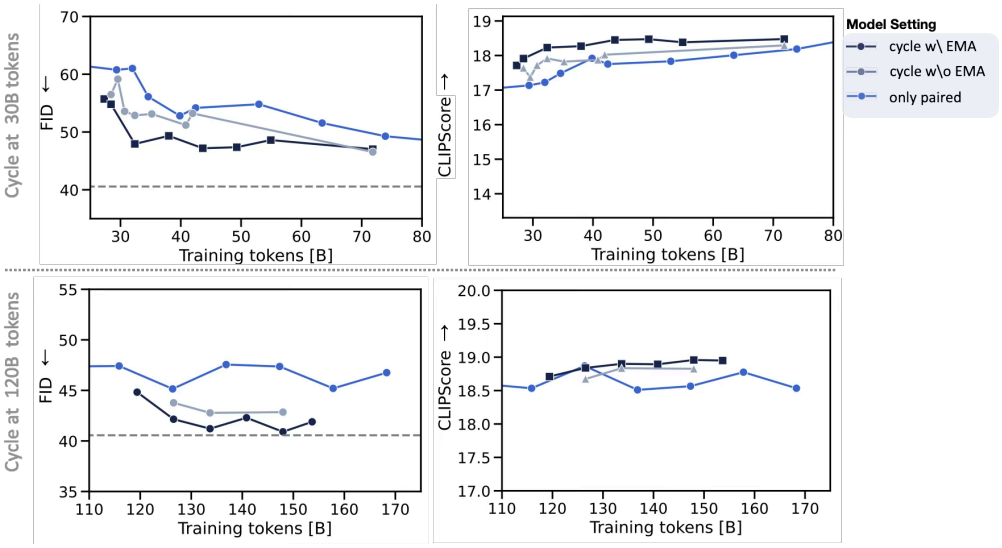


Figure 5: Effect of an EMA teacher on downstream performance for early and late starts of cycle training.

EMA versus no EMA. In our default configuration, an EMA-based teacher generates intermediate modalities that guide cycle reconstruction and serve as targets for the anchoring loss. The no-EMA variant instead uses the current model weights for first-pass generation. Figure 5 compares these settings for both cycle starting points. For the early fork, the EMA-based model achieves better performance than the non-EMA counterpart. At 32B tokens, EMA achieves FID 47.9 versus 52.9 without EMA, and CLIPScore 18.2 versus 17.9. The gap is largest within the first 15B tokens after the fork and closes by 72B tokens (FID 47.0 vs. 46.5). This pattern is consistent with EMA stabilizing the initial cycle updates. For the late fork, the EMA model keeps a smaller but consistent advantage of 1.5-2 FID (e.g. 40.9 vs. 42.8 at 148B tokens), while CLIPScore differs by less than 0.2. Notably, both the EMA and non-EMA settings outperform paired-only training at the same number of training tokens (FID 52.9 vs. 61.0 at 32B and 42.8 vs. 47.4 at 148B). EMA therefore improves early training and the final performance of the late-fork model, while gains from cycle training also persist without it.

B.3 Effect of Teacher Anchoring

Downstream performance. Figure 6 compares cycle training with and without teacher anchoring, both forked from the paired model after 27B tokens with $\pi = 25\%$. Without anchoring, cycle training provides little benefit. FID stays at the level of paired-only training (56-60 vs. 56-61), and image-to-text retrieval even falls below it (R@1 3.3-3.9 vs. 3.8-4.3). With anchoring, cycle training improves all four metrics from approximately 31B tokens onward. FID drops to about 53, CLIPScore reaches 17.9, CIDEr stays at 0.10-0.13 instead of 0.09-0.10, and R@1 rises to 4.3-4.6.

Figure 7 visualizes the evolution of the pass-2 CLIP intermediates c_2 for one validation caption after cycle training begins at 26.7B tokens. PCA projections at checkpoints from 29.6B to 40.8B tokens are shown alongside the image and its ground-truth CLIP representation. Without anchoring, the intermediates exhibit irregular patches and horizontal bands with little resemblance to the object's

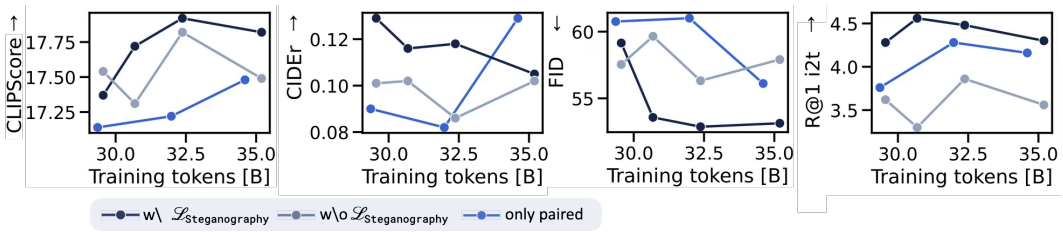


Figure 6: Effect of teacher anchoring on downstream performance after cycle training starts at 27B tokens. From approximately 31B tokens onward, anchoring improves all four evaluated metrics relative to the unanchored model.

spatial structure, even though the model can still reconstruct the caption. With anchoring, the intermediates retain a coherent foreground shape across checkpoints. This example illustrates how successful reconstruction can coexist with degraded intermediate representations, and how teacher anchoring helps preserve their semantic structure.

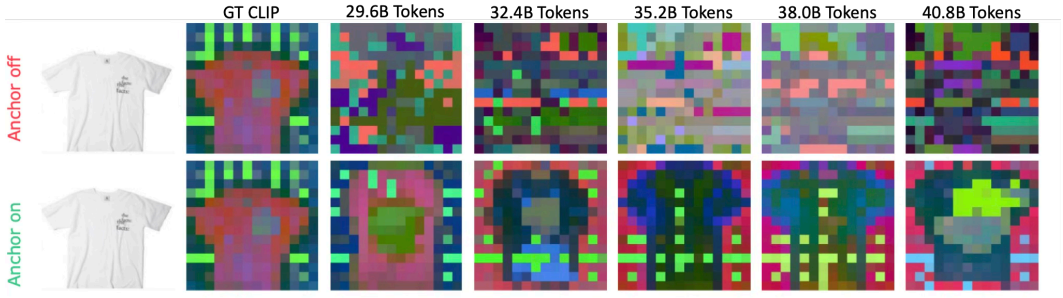


Figure 7: Pass-2 CLIP intermediates c_2 for one validation caption, without (top) and with (bottom) teacher anchoring. Columns show the real image, its ground-truth CLIP representation, and intermediates at 29.6B–40.8B total training tokens, after cycle training begins at 26.7B. CLIP features are visualized using PCA.

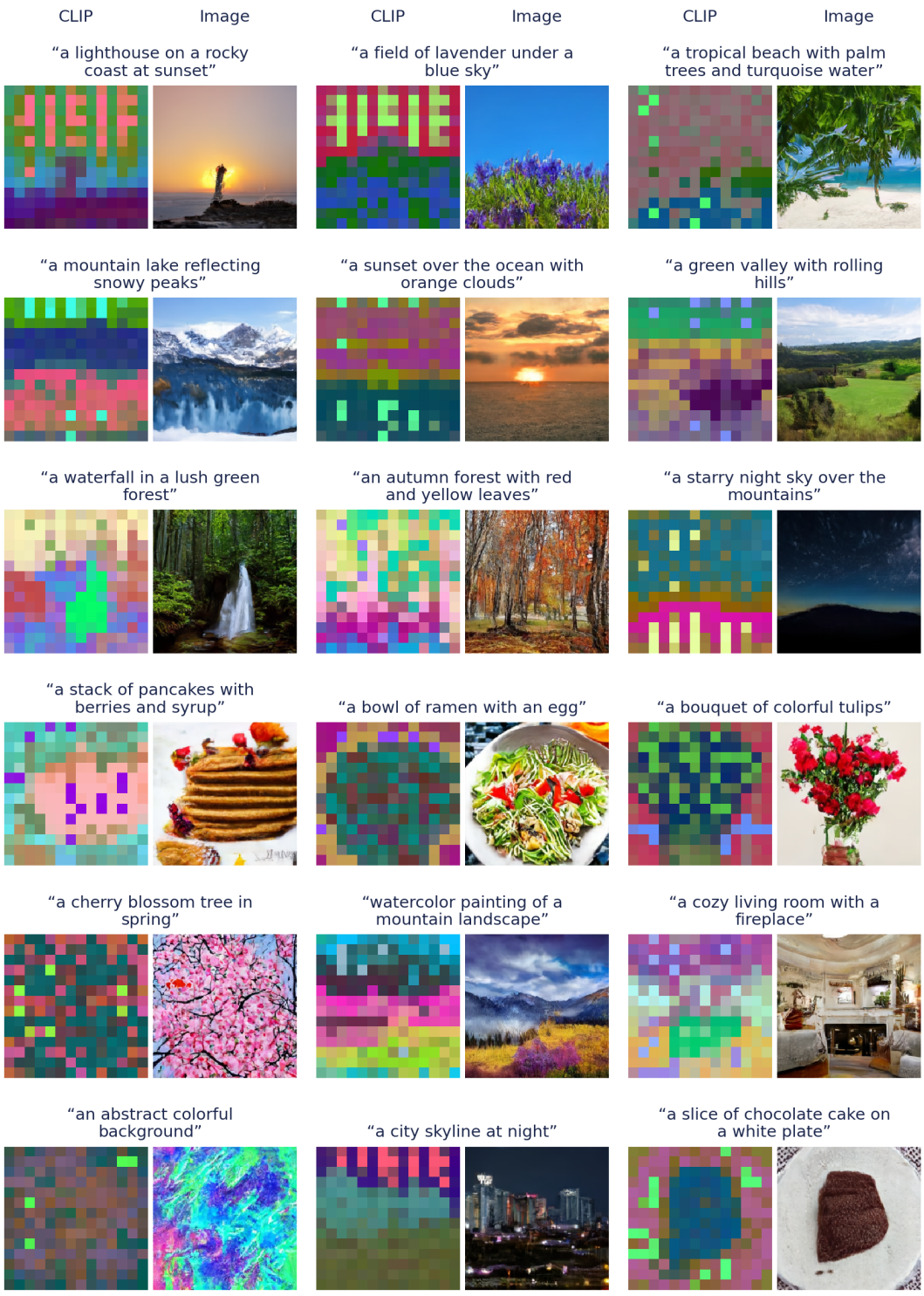


Figure 8: Qualitative examples from `unimo-x` at 25% pairing, trained with cycle reconstruction and teacher anchoring. Images are generated through the caption → CLIP → RGB chain.

C Code Availability

The code will be released publicly upon publication of this work.