FLAT: RESAMPLING IMAGE AND TEXT INTO 1D FLEXIBLE-LENGTH ALIGNED TRANSMODAL TOKENS FOR RETRIEVAL AND GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Traditional multimodal representation learning and generation are two stages: a contrastive or self-supervised visual encoder is trained first, followed by a separate downstream generative model. This setup bottlenecks generative performance behind frozen embeddings. To bridge this gap, we revisit joint multimodal representation learning and generation to produce linearly interpolatable embeddings that are directly consumable by generative decoders. We present FLAT (**F**lexible-**L**ength **A**ligned **T**ransmodal representations), a representation pre-training framework that jointly optimizes a shared multimodal encoder alongside downstream text-to-image (T2I) and image-to-text (I2T) decoders. By combining contrastive alignment with bidirectional cross-modal generative objectives, FLAT ensures its representations function as both discriminative semantic descriptors and generative conditions. Architecturally, FLAT maps visual and textual inputs into a unified continuous 1D sequence space, applying nested dropout over prefix-K tokens to enable dynamic output lengths. A single pre-training stage allows FLAT to perform cross-modal retrieval and generation across variable prefix K, especially with a T2I GenEval score of 71.1. Task-specific fine-tuning further aligns model performance with state-of-the-art baselines: 83.1 GenEval on T2I generation; 40.5 BLEU-4 and 138.6 CIDEr on MS-COCO image captioning; and Recall@5 scores of 88.4 (I2T) / 77.9 (T2I) on MS-COCO alongside 98.3 (I2T) / 93.6 (T2I) on Flickr30K. Finally, qualitative evaluations demonstrate that FLAT representations natively support linear interpolation, latent space arithmetic, and zero-shot composed retrieval.

1 INTRODUCTION

Traditional multimodal architectures separate modality representation learning from cross-modal generation. In representation learning, models like CLIP (Radford et al., 2021) align visual and textual features using contrastive objectives, while DINO (Caron et al., 2021) and JEPA (Assran et al., 2023) train visual encoders with self-supervision. In image-to-text (I2T) models, frozen visual representations are connected to a transformer decoder and re-aligned with the text embedding space (Liu et al., 2023). Conversely, text-to-image (T2I) models rely on pre-trained, frozen text encoders: Stable Diffusion and SDXL utilize CLIP text embeddings, SD3 combines CLIP and T5, and PixArt, Sana, and MetaQuery leverage frozen large language models (Radford et al., 2021; Rombach et al., 2022; Podell et al., 2024; Esser et al., 2024; Chen et al., 2024; Xie et al., 2025; Pan et al., 2025). This decoupling of representation learning and generative modeling bottlenecks generative performance behind frozen representations and requires re-alignment in generative model training.

To bridge this gap, recent unified multimodal models either jointly train representation encoders with transformer backbones using generative objectives, or bypass vision encoders to operate directly in pixel space (Agrawal et al., 2024; Diao et al., 2025; Liu et al., 2026). In this work, we revisit an alternative direction: retaining an explicit contrastive, linearly-interpolatable embedding space while simultaneously optimizing alignment and generation. We present **FLAT** (**F**lexible-**L**ength **A**ligned **T**ransmodal representations), a representation pre-training framework that jointly trains a multimodal encoder with T2I and I2T decoders using **dual bidirectional alignment and generative** losses. Our central insight is that retrieval and generation are mutually reinforcing (Yu

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.

et al., 2022)—effective representations should serve as both discriminative semantic descriptors and generative conditions for either direction. We demonstrate that these combined objectives produce representations that are linearly interpolatable and consistently enhance downstream tasks.

FLAT builds upon recent advances in 1D visual tokenization, which resample image grids into compact 1D token sequences to eliminate spatial redundancy (Yu et al., 2024; Bachmann et al., 2025). Similar techniques have been employed to compress lengthy text into summary vectors (Chevalier et al., 2023). Unlike prior visual tokenizers that focus on image reconstruction, FLAT resamples both visual and textual inputs into a unified 1D representation space via a shared multimodal encoder, producing continuous 1D representations that are contrastively aligned during generative training. FLAT incorporates nested dropout (Bachmann et al., 2025; Rippel et al., 2014; Kusupati et al., 2022) into sequence representation learning by optimizing over random prefix-K tokens per batch, enabling flexible sequence length selection at inference time.

We evaluate FLAT across T2I and I2T generation and retrieval. By varying the number of prefix-K tokens, one pre-trained FLAT model retrieves and synthesizes coarse-to-fine cross-modal outputs from the same encoder pass. After a single pre-training stage it reaches a GenEval score of 71.1 on T2I generation and, zero-shot, MS-COCO Recall@5 of 69.1 (I2T) and 64.6 (T2I). Task-specific fine-tuning then brings model performance on each task in line with published SOTA baselines: 83.1 GenEval on T2I generation, above every baseline we compare against including two 7B models that additionally rewrite the prompt; 40.5 BLEU-4 and 138.6 CIDEr on MS-COCO for I2T generation; and Recall@5 of 88.4 (I2T) / 77.9 (T2I) on MS-COCO and 98.3 (I2T) / 93.6 (T2I) on Flickr30K. On ImageNet linear probing the frozen representation reaches 81.8, above all published results for generative latent space. Finally, qualitative evaluations show that FLAT representations natively support interpolation, semantic arithmetic on latent space, and zero-shot composed retrieval.

2 Related Work

Decoupled Representation Learning and Generation. Contrastive representation learning methods in the CLIP family (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2023; Sun et al., 2023) have established the benchmark for cross-modal alignment and retrieval. However, downstream multimodal generative models typically build on separately pretrained encoders. In I2T architectures such as BLIP-2 (Li et al., 2023a) and LLaVA (Liu et al., 2023), frozen visual encoders are connected to large language models through trainable adapters. Conversely, T2I models such as unCLIP (Ramesh et al., 2022), SDXL (Podell et al., 2024), SD3 (Esser et al., 2024), PixArt-α (Chen et al., 2024), and Sana (Xie et al., 2025) use frozen text or multimodal encoders as conditioning modules. This separation persists in many unified multimodal systems, which keep pretrained visual encoders, tokenizers, or VAEs frozen while training the components for understanding and generation (Wu et al., 2024a; Chameleon Team, 2024; Xie et al., 2024; Zhou et al., 2025; Pan et al., 2025; Chen et al., 2025a).

Joint Representation Learning and Generation. BLIP (Li et al., 2022) and CoCa (Yu et al., 2022) combine contrastive image–text alignment with text generation. MAGE (Li et al., 2023b) combines masked image generation with self-supervised representation learning, while DREAM (Li et al., 2026) jointly optimizes image–text contrastive alignment and masked image generation. Meanwhile, recent unified multimodal models process understanding and generation within a shared architecture using generative training (Huang et al., 2025; Liu et al., 2026; Diao et al., 2026). Another line develops unified visual tokenizers or encoders that support both semantic understanding and visual reconstruction (Wu et al., 2024b; Ma et al., 2025; Qu et al., 2025; Zhao et al., 2025; Lin et al., 2025; Fan et al., 2025; Yue et al., 2026; Zhang et al., 2026). FLAT instead jointly optimizes a variable-length continuous representation for contrastive retrieval and bidirectional cross-modal generation, covering both I2T and T2I.

Resampled 1D Representations. TiTok (Yu et al., 2024), FlexTok (Bachmann et al., 2025) and GigaTok (Xiong et al., 2025) resample spatial image grids into compact 1D token sequences. AutoCompressor (Chevalier et al., 2023) resamples long text inputs into summary vectors. BLIP-2 (Li et al., 2023a) resamples image encoder outputs with learnable queries for image-text alignment. MiniGPT-5 (Zheng et al., 2023), DreamLLM (Dong et al., 2024), and MetaQuery (Pan et al., 2025) map or query multimodal LLM hidden states to condition image generation. In comparison, FLAT resamples both images and text into a shared continuous space as 1D sequence representations.

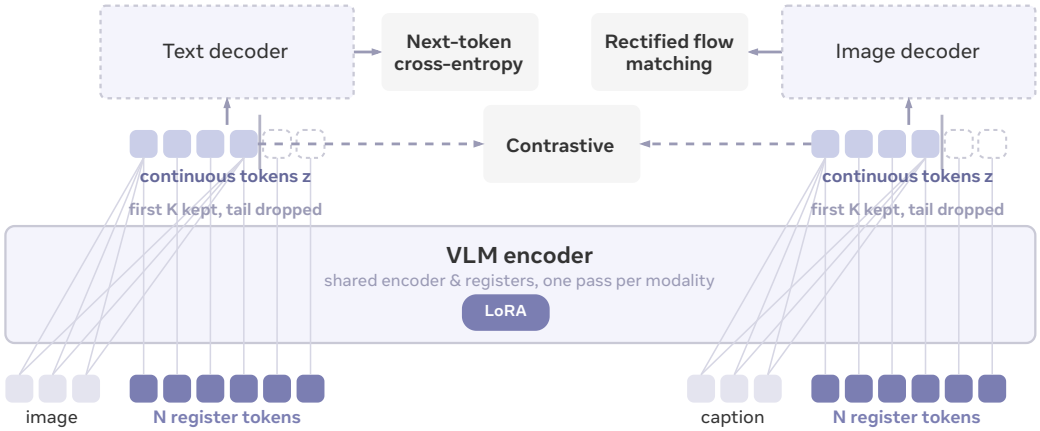


Figure 1: **FLAT architecture overview.** A shared VLM encoder processes visual and textual inputs through register tokens, applying nested dropout to truncate sequences to variable prefix lengths during joint training. FLAT's unified 1D continuous representation simultaneously drives contrastive alignment, autoregressive caption generation, and rectified-flow image synthesis.

3 METHOD

3.1 OVERVIEW

FLAT is a representation pre-training framework which learns a multimodal encoder and two transmodal decoders jointly with contrastive and generative objectives. Given an input text or image, FLAT maps the input to an ordered sequence of continuous tokens, which we call the *representation*. The representation is contrastively aligned and read by both transmodal decoders. An overview of FLAT architecture is shown in Fig. 1.

Representation Encoder. Let x denote an input of modality $m \in \{\text{img}, \text{txt}\}$. FLAT appends N learnable register tokens $R = \{r_n\}_{n=1}^{N}$ to each input and encodes the resulting sequence with a shared multimodal encoder f_{enc}, which is a pre-trained Vision-Language Model (VLM) with a system prompt *Represent the input*. The hidden states at the register positions are linearly projected to obtain the representation $\mathbf{z}$:

$$\mathbf{z} = W_{\text{lat}}\, f_{\text{enc}}(x, R) \in \mathbb{R}^{N \times d}, \tag{1}$$

where d is the register dimension and W_{lat} the projection. The two modalities share the same registers and the same projection. The causal encoder places registers after the input tokens, so each register attends to the complete input and the preceding registers while remaining independent of later registers.

Representation Dropout. To enable variable sequence lengths, we apply nested dropout (Rippel et al., 2014) directly over the representation $\mathbf{z}$. For each training batch, a keep-length K is sampled from a geometric ladder $\mathcal{K} = \{1, 4, 16, 64, 256\}$, and only the prefix sequence $\mathbf{z}_{:K}$ is passed to the contrastive objective and downstream decoders. Sampling from a geometric ladder rather than continuous integers avoids wasting optimization steps on imperceptible length differences. By default, FLAT samples uniformly from the $\mathcal{K}$ ladder. App. F.3 compares uniform sampling with other sampling strategies.

Compared to FlexTok (Bachmann et al., 2025), which maintains a fixed sequence length by padding tail positions with learned null embeddings, we find that directly truncating the representation to the prefix $\mathbf{z}_{:K}$ yields better results in FLAT training setups (App. I). To ensure consistency, the target length K is sampled once per training step and broadcast globally across all ranks. This guarantees that both contrastive and generative objectives operate on the exact same sequence prefix at every optimization step.

Decoders. The truncated representation sequence $\mathbf{z}_{:K}$ serves as the semantic condition for both the I2T and T2I decoders. Given a paired image-text input, the visual representation $\mathbf{z}_{:K}^{\text{img}}$ conditions caption generation, while the textual representation $\mathbf{z}_{:K}^{\text{txt}}$ conditions image synthesis. For each generative pathway, $\mathbf{z}_{:K}$ is treated as a sequence of soft tokens and mapped into the respective decoder's input dimension h:

$$\mathbf{e} = \text{RMSNorm}(\text{MLP}(\mathbf{z}_{:K})) \in \mathbb{R}^{K\times h}, \tag{2}$$

where the MLP projections consist of decoupled, task-specific parameters for the I2T and T2I decoders.

The I2T decoder is a standard autoregressive language model. The visual soft tokens $\mathbf{e}^{\text{img}}$ are prepended with a standard system prompt (*"Describe the input"*) and fed as input tokens to autoregressively generate the corresponding text caption. The T2I decoder is a standard rectified flow transformer that denoises continuous VAE latents. The textual soft tokens $\mathbf{e}^{\text{txt}}$ act as semantic conditioning signals integrated via cross-attention mechanisms, guiding the flow-matching model to synthesize VAE latents decodable into the target image. The detailed, multi-task training objectives governing these decoders are formally defined in the next section.

3.2 Joint Training

Given a paired image and text, FLAT is trained with both bidirectional contrastive loss and generation loss. Since the soft tokens $\mathbf{e}$ are projected from the representation $\mathbf{z}$, we use $\mathbf{z}$ consistently as the condition in the loss functions.

Contrastive Loss. We contrast corresponding register positions in the representation $\mathbf{z}_{:K}$. For an image–text pair (i, j) we define the similarity computed over their total K register positions as a late-interaction score between registers at matching positions:

$$s_{ij}^{(K)} = \tfrac{1}{K}\textstyle\sum_{n=1}^{K} \frac{(z_{i,n}^{\text{img}})^\top z_{j,n}^{\text{txt}}}{\|z_{i,n}^{\text{img}}\|_2 \, \|z_{j,n}^{\text{txt}}\|_2}, \tag{3}$$

and the contrastive loss $\mathcal{L}_{\text{align}}^{(K)}$ as

$$\mathcal{L}_{\text{i2t}}^{(K)} = -\tfrac{1}{B}\textstyle\sum_{i=1}^{B} \log \frac{\exp\left(s_{i,i}^{(K)}/\tau\right)}{\sum_{j=1}^{B}\exp\left(s_{i,j}^{(K)}/\tau\right)}, \quad \mathcal{L}_{\text{t2i}}^{(K)} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(s_{i,i}^{(K)}/\tau\right)}{\sum_{j=1}^{B}\exp\left(s_{j,i}^{(K)}/\tau\right)}, \tag{4}$$

$$\mathcal{L}_{\text{align}} = \tfrac{1}{2}\left(\mathcal{L}_{\text{i2t}}^{(K)} + \mathcal{L}_{\text{t2i}}^{(K)}\right). \tag{5}$$

where B denotes the number of gathered global candidates across all ranks and $\tau = 0.07$ is the contrastive temperature.

Generation Loss. For the same image–text pair, we apply symmetric generation losses for I2T and T2I, both using the representations $\mathbf{z}_{:K}$ as semantic conditions.

For I2T generation and an image input with representation $\mathbf{z}_{:K}^{\text{img}}$, the loss over target caption y and decoder parameter θ is defined as:

$$\mathcal{L}_{\text{txt}} = -\textstyle\sum_t \log p_\theta\left(y_t \,\middle|\, y_{<t}, \mathbf{z}_{:K}^{\text{img}}\right) \tag{6}$$

For T2I generation and a text input, the representation $\mathbf{z}_{:K}^{\text{txt}}$ are used as semantic conditions for a flow-matching decoder. Let $\hat{x}$ be the clean VAE image latent, $\epsilon \sim \mathcal{N}(0, I)$, and $t \in [0, 1]$. The noised VAE latent and target velocity used in flow matching are defined as

$$\hat{x}_t = (1-t)\hat{x} + t\epsilon, \qquad v^\star = \epsilon - \hat{x}. \tag{7}$$

The conditional flow-matching objective over decoder parameter ϕ is defined as:

$$\mathcal{L}_{\text{img}} = \mathbb{E}_{\hat{x},\epsilon,t}\left[\left\|v_\phi\left(\hat{x}_t, t, \mathbf{z}_{:K}^{\text{txt}}\right) - (\epsilon - \hat{x})\right\|_2^2\right]. \tag{8}$$

Table 1: **Text-to-image synthesis on GenEval.** FLAT achieves competitive scores while enabling inference-time sequence length selection. Evaluated from a single fine-tuned checkpoint across prefix lengths. † marks a system that rewrites the prompt before generating.

Method	Backbone	GenEval
Chameleon (Chameleon Team, 2024)	scratch 7B	0.39
Janus (Wu et al., 2024a)	DeepSeek 1.5B	0.61
Transfusion (Zhou et al., 2025)	scratch 7B	0.63
JanusFlow (Ma et al., 2024)	DeepSeek 1.5B	0.63
Emu3† (Wang et al., 2024)	scratch 7B	0.66
Show-o-512 (Xie et al., 2024)	Phi-1.5 1.3B	0.68
Janus-Pro-1B (Chen et al., 2025b)	DeepSeek 1.5B	0.73
MetaQuery-L (Pan et al., 2025)	Qwen2.5-VL 3B	0.78
Janus-Pro-7B (Chen et al., 2025b)	DeepSeek 7B	0.80
MetaQuery-XL† (Pan et al., 2025)	Qwen2.5-VL 7B	0.80
FLAT, K=1	Qwen3.5 2B	0.49
FLAT, K=4	Qwen3.5 2B	0.77
FLAT, K=16	Qwen3.5 2B	**0.83**
FLAT, K=64	Qwen3.5 2B	**0.83**
FLAT, K=256	Qwen3.5 2B	0.82

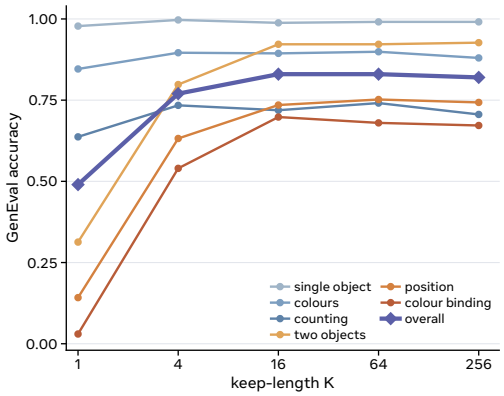


Figure 2: **GenEval accuracy across keep-lengths broken down by eval category.** Smaller K can capture core entity and color, but larger K is needed to capture spatial relations, multi-object layouts and attribute bindings.

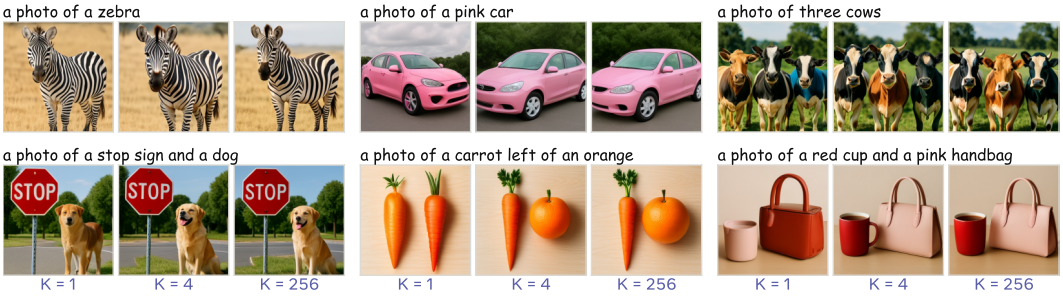


Figure 3: **Qualitative T2I synthesis across different prefix lengths.** Large K provides more fine-grained visual details.

Training Objective. In representation pre-training, FLAT optimizes the total loss:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{txt}}\mathcal{L}_{\text{txt}} + \lambda_{\text{img}}\mathcal{L}_{\text{img}} \tag{9}$$

In fine-tuning, we show that each task-specific objective can be optimized separately over one of the encoder and decoders to maximize model performance.

4 Experiments

We evaluate FLAT across generation, retrieval, and probing tasks to assess its unified 1D continuous representations. Our evaluation spans three core dimensions: (1) **Generative Evaluation,** demonstrating that the FLAT representation drives both T2I synthesis and I2T captioning competitively against specialized baselines; (2) **Discriminative Evaluation,** evaluating the FLAT representation directly with cross-modal retrieval, linear probing, and unsupervised clustering; and (3) **Geometry Analysis,** analyzing latent geometry to show that FLAT projects image and text into a unified representation space, natively unlocking emergent zero-shot linear interpolation, latent space arithmetic and composed retrieval. App. G further extends this evaluation to zero-shot generalization across multilingual prompts and video modalities.

4.1 Implementation Details

The representation encoder and text decoder are built on a Qwen3.5-2B backbone (Qwen Team, 2026) with different LoRA adapters, while the image decoder is initialized from SANA (Xie et al.,

Table 2: **Image captioning on the MS-COCO Karpathy test split.** FLAT achieves competitive performance with a compact latent sequence. Column *Latent Feat.* gives the latent feature shape used to condition the text decoder. The metrics follow previous work: B@4 (BLEU-4), M (METEOR), R (ROUGE-L), C (CIDEr), S (SPICE).

Method	Latent Feat.	B@4	M	R	C	S
CrossFlow (Liu et al., 2025)	$4\times32\times32$	36.4	27.8	57.1	116.2	20.4
SCD-Net (Luo et al., 2023)	$N\times512$	37.3	28.1	58.0	118.0	21.6
Florence-2-L (Xiao et al., 2024)	–	–	–	–	143.3	–
CoCa (Yu et al., 2022)	256×1408	40.9	33.9	–	143.6	24.7
BLIP-2 (Li et al., 2023a)	32×768	43.7	–	–	145.8	–
FLAT, $K=1$	1×64	35.3	28.7	57.7	122.0	22.0
FLAT, $K=4$	4×64	38.1	29.9	59.4	131.0	23.3
FLAT, $K=16$	16×64	39.6	30.8	60.4	136.2	24.1
FLAT, $K=64$	64×64	40.2	31.1	60.8	138.1	24.4
FLAT, $K=256$	256×64	**40.5**	**31.2**	**60.8**	**138.6**	**24.4**



Figure 4: **Image captioning with fine-tuned I2T decoder.** Small K generates semantically meaningful captions and larger K refines lexical details.

2025). The total parameter of FLAT is 3.7B. Full architectural, pre-training and fine-tuning configurations are provided in App. B. Quantitative evaluations are conducted as a function of prefix length K. Specifically, we benchmark T2I synthesis on GenEval (Ghosh et al., 2023) without prompt rewriting; I2T captioning on the MS-COCO Karpathy test split; and cross-modal retrieval on both the MS-COCO and Flickr30k Karpathy splits (Karpathy & Fei-Fei, 2015). We additionally examine zero-shot composed retrieval on the MMEB CIRR test split (Jiang et al., 2025; Liu et al., 2021). The full eval protocols are shown in App. B.1. While our main result focuses on the final fine-tuned checkpoints, we report pre-trained FLAT performance across all tasks in App. C and singe-stage task-specific training results (no pre-training) in App. D.

4.2 Generative Evaluation

Text-to-Image Generation. For T2I generation, we continue training the image decoder from the pre-trained checkpoint for 8k steps on 120K clean image–text pairs (Table 7). As shown in Table 1, FLAT achieves a GenEval score of **0.83** using a 2B encoder without prompt rewriting, outperforming MetaQuery-L which uses prompt rewrites. Categorical GenEval analysis further demonstrates a clear link between task complexity and required prefix length K (Fig. 2). While a single token ($K=1$) captures basic semantics like single object and color, relational and compositional tasks require larger prefixes. Complex categories—two objects, position, and color binding—score near zero at $K=1$ (e.g., 0.03 for color binding), but recover steeply by $K=4$ and $K=16$ (reaching 0.67). As shown in Fig. 3, $K=1$ generates only the main entity, while spatial arrangements, secondary entities, and attribute-object bindings emerge accurately as K increases.

Image-to-Text Generation. The I2T evaluation fine-tunes only the I2T-decoder LoRA with captioning loss on COCO Karpathy training+restval set for 5k steps (Table 7). As detailed in Table 2, performance improves steadily across all metrics as K increases. Qualitatively (Fig. 4), a single

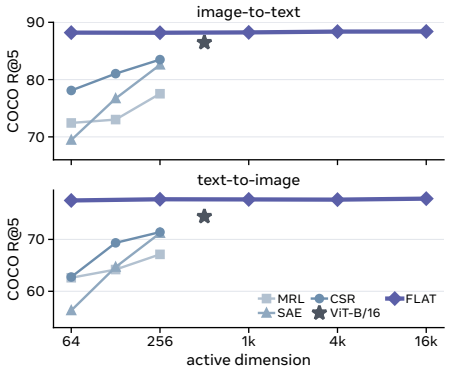


Figure 5: **MS-COCO retrieval accuracy across active dimensions.** Unlike other variable-width baselines whose performance degrades sharply at reduced dimensions, FLAT maintains stable retrieval accuracy across all dimensions.

Table 3: **T2I and I2T retrieval results across prefix K.** The first token captures key cross-modal discriminability, keeping retrieval accuracy virtually unaffected by prefix truncation. Performance is evaluated on the MS-COCO and Flickr30k Karpathy test splits.

	I→T			T→I		
K	R@1	R@5	R@10	R@1	R@5	R@10
MS-COCO Karpathy 5k						
1	65.52	88.20	93.84	50.19	77.51	85.92
4	65.52	88.18	94.06	**50.55**	77.74	**86.14**
16	65.76	88.24	**94.10**	50.42	77.71	86.13
64	65.80	88.38	94.06	50.42	77.67	86.11
256	**65.82**	**88.40**	93.92	50.49	**77.86**	86.07
Flickr30k Karpathy 1k						
1	90.50	98.10	99.50	76.94	93.52	96.48
4	**91.20**	**98.30**	99.50	**77.40**	93.58	**96.54**
16	90.80	98.30	**99.60**	77.32	93.60	96.54
64	90.90	98.20	99.40	76.88	**93.72**	96.54
256	90.80	98.20	99.30	76.26	93.16	96.40

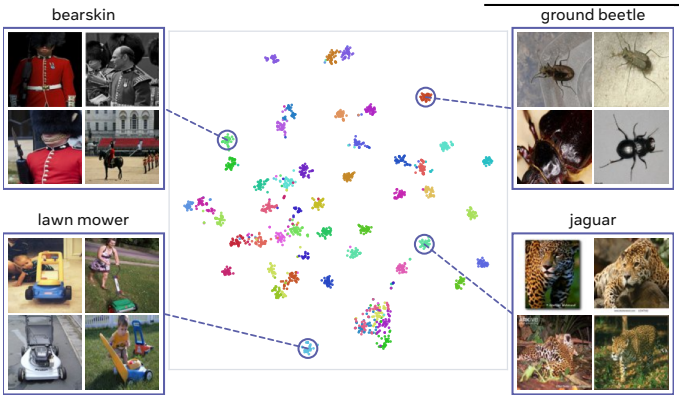


Figure 6: **Unsupervised k-means clustering with a single FLAT token.** Clusters formed from frozen $K=1$ representations (64 dimensions) align with ImageNet ground-truth classes without label supervision.

Table 4: **ImageNet-1K linear-probe top-1 accuracy.** A linear classifier is fit to each frozen representation; Qwen uses its last hidden state and FLAT concatenates the first K tokens.

Method	Dim	Top-1
Generative Latent Spaces		
SD-VAE latent	4	8.0
REPA (Leng et al., 2025)	–	62.5
MAE-B (He et al., 2022)	768	68.0
Frozen VLM Backbone		
Qwen3.5-2B (last)	2048	79.7
Jointly Contrastive and Generative		
DREAM (Li et al., 2026)	1024	72.7
FLAT, $K=1$	64	73.3
FLAT, $K=4$	256	77.0
FLAT, $K=16$	1024	81.2
FLAT, $K=64$	4096	**81.8**
FLAT, $K=256$	16384	81.8

token ($K=1$) generates well-formed, coherent captions, while additional tokens enhance lexical precision and concrete detail. These qualitative refinements directly mirror the quantitative gains in Table 2. Even at $K=1$, FLAT outperforms baselines such as CrossFlow (Liu et al., 2025) and SCD-Net (Luo et al., 2023) across CIDEr, METEOR, and SPICE scores.

4.3 Discriminative Evaluation

Cross-Modal Retrieval. For each benchmark, retrieval adapts the encoder LoRA, registers, and latent projection using contrastive loss on the corresponding training split. The reported checkpoints use 5k steps for COCO and 320 steps for Flickr30K (Table 7). Fig. 5 compares FLAT against representations that also support adaptive dimensionality from Wen et al. (2025), including MRL, SAE, and CSR. For FLAT, each token contains 64 dimensions, spanning from a 64-dimensional embedding ($K=1$) to a 16,384-dimensional embedding ($K=256$). Notably, FLAT's retrieval performance remains virtually invariant across varying K, whereas baseline accuracy degrades sharply as K decreases. Table 3 reports detailed retrieval results across the K spectrum. Remarkably, a single token is sufficient for retrieval: every metric stays within roughly one percentage point across a $256\times$ reduction in width. This mirrors our observations in T2I and I2T generation, confirming that the first token carries sufficient global semantics and discriminative power.

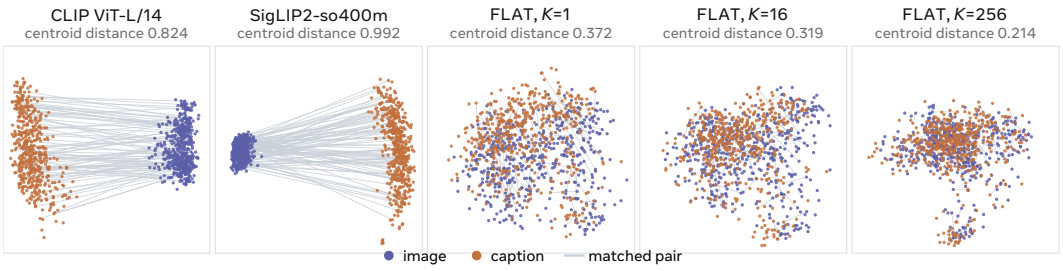


Figure 7: **PCA distributions for image-text pairs from MS-COCO.** Dual-encoder baselines exhibit clear spatial separation between image and text representations, whereas FLAT projects both modalities into a highly overlapping latent space. Increasing prefix K further shrinks the cross-modal centroid distance.



Figure 8: **Continuous cross-modal feature interpolations.** Linear interpolation between a pair of image–text produces smooth semantic transitions. Image and text decoders map the same representation to corresponding concepts.

Linear Probing and Clustering. To directly evaluate the representation space, we perform linear probing and unsupervised clustering on FLAT representations. A single linear classifier trained on a frozen $K{=}1$ token achieves 73.3% top-1 accuracy, scaling to 81.2% at $K{=}16$. This outperforms all purely generative latent spaces and surpasses DREAM (Li et al., 2026) which jointly trains the representation encoder with an image decoder, at equivalent dimensionality. Furthermore, without fitting any parametric heads, k-means clustering on a single FLAT token cleanly recovers clusters aligned with ImageNet classes (Fig. 6).

4.4 Geometry Analysis

Geometry Visualization. As shown in Fig. 7, while dual-encoder models like CLIP and SigLIP2 suffer from a persistent modality gap despite strong retrieval performance (Radford et al., 2021; Tschannen et al., 2025; Liang et al., 2022), FLAT projects both modalities into a tighter embedding space. FLAT achieves significantly greater cross-modal overlap: a single register token cuts CLIP's centroid distance by half, and the full 256 tokens reduces it to a quarter. This tightly aligned geometry enables the zero-shot latent operations described below.

Continuous Feature Interpolation. We perform linear interpolation between a pair of FLAT codes as $z_\alpha = (1-\alpha)z_1 + \alpha z_2$, and decode each intermediate latent representation using both generative heads. As shown in Fig. 8, this yields smooth semantic transitions across the input image–text pairs. Both the image and text decoders generate intermediate concepts along the interpolation trajectory. Furthermore, the prefix length K dictates transition granularity: longer prefixes smoothly vary fine-grained attributes, whereas $K{=}1$ forms a coarser semantic average (App. F.4).

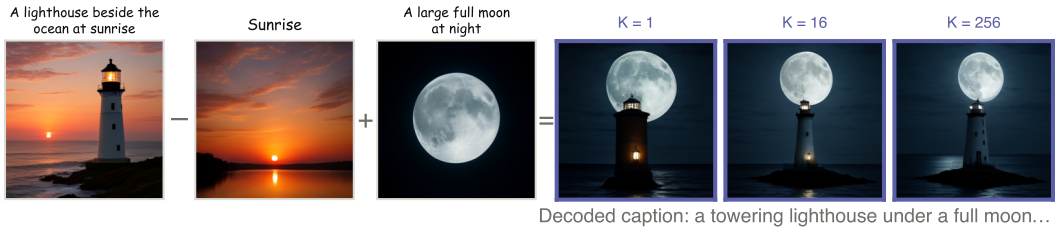


Figure 9: **Zero-shot token-wise arithmetic operation.** Linear addition and subtraction modify the target concept in both image and text outputs while preserving the surrounding scene.

Semantic Arithmetic. FLAT representations naturally support zero-shot arithmetic, $z_{\text{edit}} = z_1 - z_2 + z_3$, without explicit editing supervision. As shown in Fig. 9, subtracting *sunrise* and adding *a full moon at night* modifies the focal concept while preserving the surrounding scene context, with both decoders interpreting the composite representation consistently. The same space also supports zero-shot composed image retrieval on CIRR by directly adding separately encoded reference-image and instruction tokens (App. G.3).

4.5 Ablation Study on Training Loss Combinations

FLAT jointly optimizes a contrastive loss alongside bidirectional generation losses for I2T captioning and T2I synthesis. To evaluate the contribution of each objective, we train ablation variants across all seven non-empty loss combinations for 40k steps. Figure 10 and Table 5 present these variants along with their normalized and absolute eval scores averaged over $K \in \{1, 4, 16, 64, 256\}$. Overall, the three objectives provide complementary signals and make positive marginal contributions across prefix lengths.

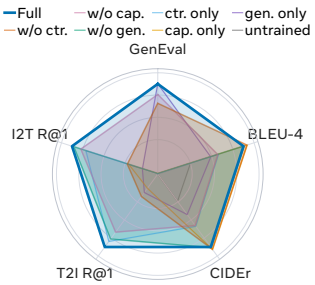


Figure 10: **Eval for different training loss combinations (normalized scores).**

Table 5: **Eval for different training loss combinations (absolute scores).**

Losses	Retrieval		Captioning		Generation
	I2T	T2I	B@4	CIDEr	GenEval
Full	54.2	58.8	38.6	133.0	0.325
w/o contrastive	37.7	35.4	39.2	134.5	0.296
w/o captioning	51.9	53.9	33.9	117.1	0.309
w/o generation	53.7	56.1	38.7	133.4	0.004
Contrastive only	50.7	57.0	34.0	117.8	0.002
Captioning only	35.5	22.5	39.3	134.9	0.004
Generation only	19.1	29.2	32.6	109.0	0.325
untrained	0.2	0.0	28.8	97.4	0.000

We further quantify how the three losses reinforce one another using a Shapley value decomposition (Shapley, 1953) over all coalitions. The attribution matrices in App. H show that while each task is driven primarily by its matched loss, the other objectives provide complementary signals.

5 Conclusion

We introduce FLAT, a representation pre-training framework that resamples images and text into 1D flexible-length, aligned, transmodal tokens. FLAT jointly optimizes a shared representation encoder alongside T2I and I2T decoders using bidirectional alignment and generation losses. This setup learns unified semantic representations not only useful for cross-modal retrieval but also as effective generative conditions for both decoders. Empirical results demonstrate that FLAT natively supports variable-length retrieval and synthesis, and with lightweight task-specific fine-tuning pushing T2I / I2T retrieval and generation performance to SOTA levels. Furthermore, qualitative analysis reveals that FLAT tokens enable zero-shot linear interpolation, latent space arithmetic, and composed retrieval.

AI USE STATEMENT

We used generative AI tools to assist with experiment planning and management, edit training and evaluation scripts, organize experiment reports, and summarize and interpret experimental results. We also used these tools to improve the readability of the manuscript, format references, and prepare figures and other presentation artifacts. All AI-assisted code, analyses, figures, and text were reviewed and verified by the authors. The authors take responsibility for the final content of this work.

REFERENCES

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Roman Bachmann, Jesse Allardice, David Mizrahi, Enrico Fini, Oğuzhan Fatih Kar, Elmira Amirloo, Alaaeldin El-Nouby, Amir Zamir, and Afshin Dehghan. FlexTok: Resampling images into 1d token sequences of flexible length. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2502.13967.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. In *arXiv preprint arXiv:2405.09818*, 2024.

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

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, 2023. doi: 10.1109/CVPR52729.2023.00276. arXiv:2212.07143.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. Adapting language models to compress contexts. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3829–3846, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Haiwen Diao, Xiaotong Li, Yufeng Cui, Yueze Wang, Haoge Deng, Ting Pan, Wenxuan Wang, Huchuan Lu, and Xinlong Wang. Evev2: Improved baselines for encoder-free vision-language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21014–21025. IEEE, 2025.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, et al. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. DreamLLM: Synergistic multimodal comprehension and creation. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.11499.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024.

Weichen Fan, Haiwen Diao, Quan Wang, Dahua Lin, and Ziwei Liu. The prism hypothesis: Harmonizing semantic and pixel representations via unified autoencoding. *arXiv preprint arXiv:2512.19693*, 2025.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2310.11513.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. arXiv:2111.06377.

Ziyuan Huang, DanDan Zheng, Cheng Zou, Rui Liu, Xiaolong Wang, Kaixiang Ji, Weilong Chai, Jianxin Sun, Libin Wang, Yongjie Lv, et al. Ming-univision: Joint image understanding and generation with a unified continuous tokenizer. *arXiv preprint arXiv:2510.06590*, 2025.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, pp. 4904–4916, 2021.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-V: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.05160.

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.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. REPA-E: Unlocking VAE for end-to-end tuning with latent diffusion transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Chao Li, Tianhong Li, Sai Vidyaranya Nuthalapati, Hong-You Chen, Satya Narayan Shukla, Jianpeng Cheng, Yonghuan Yang, Jun Xiao, Xiangjun Fan, Aashu Singh, Dina Katabi, and Shlok Kumar Mishra. Unifying contrastive and generative objectives for visual understanding and text-to-image generation. *arXiv preprint arXiv:2603.02667*, 2026. URL https://arxiv.org/abs/2603.02667.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, 2022. arXiv:2201.12086.

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 (ICML)*, 2023a. arXiv:2301.12597.

Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2142–2152, 2023b.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2203.02053.

Haokun Lin, Teng Wang, Yixiao Ge, Yuying Ge, Zhichao Lu, Ying Wei, Qingfu Zhang, Zhenan Sun, and Ying Shan. Toklip: Marry visual tokens to clip for multimodal comprehension and generation. *arXiv preprint arXiv:2505.05422*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Qihao Liu, Xi Yin, Alan Yuille, Andrew Brown, and Mannat Singh. Flowing from words to pixels: A noise-free framework for cross-modality evolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2412.15213.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. arXiv:2108.04024.

Zhiheng Liu, Weiming Ren, Xiaoke Huang, Shoufa Chen, Tianhong Li, Mengzhao Chen, Yatai Ji, Sen He, Jonas Schult, Belinda Zeng, et al. Tuna-2: Pixel embeddings beat vision encoders for multimodal understanding and generation. *arXiv preprint arXiv:2604.24763*, 2026.

Jianjie Luo, Yehao Li, Yingwei Pan, Ting Yao, Jianlin Feng, Hongyang Chao, and Tao Mei. Semantic-conditional diffusion networks for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. arXiv:2212.03099.

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. UniTok: A unified tokenizer for visual generation and understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2502.20321.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *arXiv preprint arXiv:2411.07975*, 2024.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with MetaQueries. *arXiv preprint arXiv:2504.06256*, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2307.01952.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K. Du, Zehuan Yuan, and Xinglong Wu. TokenFlow: Unified image tokenizer for multimodal understanding and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Qwen Team. Qwen3.5-Omni Technical Report. *arXiv preprint arXiv:2604.15804*, 2026. URL https://arxiv.org/abs/2604.15804.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

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.

Oren Rippel, Michael A. Gelbart, and Ryan P. Adams. Learning ordered representations with nested dropout. In *International Conference on Machine Learning (ICML)*, 2014.

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

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games, Volume II*, volume 28 of *Annals of Mathematics Studies*, pp. 307–318. Princeton University Press, 1953. doi: 10.1515/9781400881970-018.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. UniIR: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision (ECCV)*, pp. 387–404, 2024. doi: 10.1007/978-3-031-73021-4_23. arXiv:2311.17136.

Tiansheng Wen, Yifei Wang, Zequn Zeng, Zhong Peng, Yudi Su, Xinyang Liu, Bo Chen, Hongwei Liu, Stefanie Jegelka, and Chenyu You. Beyond matryoshka: Revisiting sparse coding for adaptive representation. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2503.01776.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024a.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. VILA-U: A unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024b.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2311.06242.

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 transformers. In *International Conference on Learning Representations (ICLR)*, 2025.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

Tianwei Xiong, Jun Hao Liew, Zilong Huang, Jiashi Feng, and Xihui Liu. Gigatok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18770–18780. IEEE, 2025.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5288–5296, 2016. doi: 10.1109/CVPR.2016.571.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive captioners are image-text foundation models. In *Transactions on Machine Learning Research*, 2022. arXiv:2205.01917.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Zhengrong Yue, Haiyu Zhang, Xiangyu Zeng, Boyu Chen, Chenting Wang, Shaobin Zhuang, Lu Dong, Yi Wang, Limin Wang, and Yali Wang. UniFlow: A unified pixel flow tokenizer for visual understanding and generation. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.10575.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023.

Letian Zhang, Sucheng Ren, Yanqing Liu, Xianhang Li, Zeyu Wang, Yuyin Zhou, Huaxiu Yao, Zeyu Zheng, Weili Nie, Guilin Liu, et al. Openvision 3: A family of unified visual encoder for both understanding and generation. *arXiv preprint arXiv:2601.15369*, 2026.

Yue Zhao, Fuzhao Xue, Scott Reed, Linxi Fan, Yuke Zhu, Jan Kautz, Zhiding Yu, Philipp Krähenbühl, and De-An Huang. QLIP: Text-aligned visual tokenization unifies auto-regressive multimodal understanding and generation. *arXiv preprint arXiv:2502.05178*, 2025.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*, 2023.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations (ICLR)*, 2025.

A Overview

This appendix first details the implementation and evaluation protocols (Appendix B), followed by zero-shot results from the pre-trained checkpoint (Appendix C) and controlled task-only adaptation comparisons (Appendix D). We then provide expanded qualitative galleries (Appendix E), analyze flexible-length representations and sampling choices (Appendix F), examine zero-shot transfer and latent composition (Appendix G), and report loss and truncation ablations (Appendices H and I).

B Implementation and evaluation details

FLAT is jointly pre-trained on 65.7M image–text pairs for 135k steps on 8 nodes using the config in Table 6. Table 7 summarizes the task-specific fine-tuning configs for checkpoints evaluated in Sec. 4

Table 6: FLAT model and pre-training configuration.

Component	Configuration
Encoder and text decoder	
Model	Qwen3.5-2B, bfloat16, FlashAttention-2, gradient checkpointing
LoRA	r=16, α=32, dropout 0.05; on Q/K/V/O and gate/up/down projections
Adapters	Two adapters sharing this configuration: encoder and caption decoder
Base weights	Frozen; 11.4M trainable across the adapter and register projections
Register sequence	
Register tokens	$N = 256$, latent dimension $d = 64$
Nested dropout	$p = 1.0$, $\mathcal{K} = \{1, 4, 16, 64, 256\}$ uniform; one shared K per optimization step
Truncation	Tail dropped, not null-padded, so a prefix is a shorter sequence (App. I)
Minimum tokens	1
Read-out head	MLP with norm matching and latent normalization
Quantizer	None; the representation stays continuous and is scored directly
Objectives	
Loss weights	Contrastive 1.0; text generation 1.0; image generation 1.0
Contrastive objective	InfoNCE, temperature 0.07, cross-device negatives
Image conditioning	Text registers at every step (`imagegen_image_cond_prob`=0)
Encoder instruction	*Represent the user's input.*
Caption instruction	*Describe the input.*
Image decoder	
Model	SANA-1600M at 512 px, fully fine-tuned
Connector	16 Transformer layers, caption channels 2304, input scale 2.3452
Flow matching	Flow objective, shift 3.0, logit-normal t with mean 0 and std 1
CFG dropout	0.1
Data	
Image resolution	512×512, giving a 16×16 grid of 256 visual tokens
Maximum text length	128 tokens
Training data	65.7M image–text pairs
Approx. composition	Long-caption (29.7M), short-caption (31.8M), prompt-style (4.15M), and curated high-quality (0.06M) pairs
Data workers	20 per rank, prefetch factor 6
Optimization	
Optimizer	AdamW, $\beta = (0.9, 0.999)$, weight decay 0.0, gradient clip 1.0
Learning rate	1.0×10^{-4}; 500-step linear warm-up then cosine decay to zero
Batch size	16 per rank $\times$ 64 ranks $\times$ 1 accumulation step $=$ 1024 global
Hardware	8 nodes $\times$ 8 H200 GPUs
Precision	bfloat16
Training steps	135k pretraining

Table 7: **Task-specific fine-tuning configs.** Each column describes the config of the task-specific fine-tuned checkpoint evaluated in Sec. 4. All runs use AdamW, bfloat16, zero weight decay, direct prefix truncation, and uniform sampling over $\mathcal{K} = \{1, 4, 16, 64, 256\}$. Batch sizes are global; retrieval uses the global batch as cross-device in-batch negatives.

Setting	T2I synthesis	I2T captioning	Image–text retrieval	
			COCO	Flickr30k
Training data	120K high-quality image–text pairs	COCO Karpathy train+restval (113,287 images)	COCO Karpathy train+restval	Flickr30k Karpathy train
Updated parameters	Image decoder	Text decoder LoRA	Encoder LoRA, registers, and latent projection	Encoder LoRA, registers, and latent projection
Objective	T2I generation	I2T generation	T2I and I2T contrastive	T2I and I2T contrastive
Checkpoint step	8,000	5,000	5,000	320
Global batch	1,024	256	1,024	2,048
Learning rate	4×10^{-4}	2×10^{-4}	1×10^{-4}	1×10^{-4}
Warm-up steps	5,000	2,213	500	128

Checkpoint step denotes the selected checkpoint rather than the end of the configured schedule. The T2I and I2T runs are configured for at most $50{,}000$ and $44{,}253$ steps, respectively, so their warm-up periods cover 10% and 5% of the planned schedules. T2I follows the MiniT2I optimization recipe, using β_2=0.95, a logit-normal flow-timestep distribution with mean -0.8 and standard deviation 0.8, and a maximum text length of 256; the remaining branches use $\beta = (0.9, 0.999)$. Each decoder first receives the registers through its task-specific MLP and RMSNorm projection. The T2I branch then applies the 16-layer Transformer connector listed in Table 6, whereas I2T prepends the projected registers to the language-model context.

B.1 Evaluation protocol

T2I Generation. We evaluate text-to-image (T2I) generation using GenEval (Ghosh et al., 2023) across all 553 prompts, generating 4 images per prompt with 20 sampling steps and a classifier-free guidance (CFG) scale of 4.5. To ensure controlled comparisons across prefix lengths, the initial noise vector is sampled once per prompt and reused. We do not apply prompt rewriting at any stage.

I2T Generation. We evaluate image captioning (I2T) on the MS-COCO Karpathy test split using greedy decoding with one beam, KV caching, and at most 64 newly generated tokens; decoding stops early at EOS. Generated captions are evaluated via `pycocoevalcap`, reporting standard metrics: BLEU-4 (B@4), METEOR (M), ROUGE-L (R), CIDEr (C), and SPICE (S).

- **BLEU-4 (B@4):** Measures n-gram precision (up to 4-grams) between generated and reference captions, incorporating a brevity penalty to discourage overly short outputs.
- **METEOR (M):** Computes an F-score based on unigram precision and recall, accounting for word stemming, synonymy, and paraphrase alignment.
- **ROUGE-L (R):** Evaluates structural word-order similarity using the Longest Common Subsequence (LCS) between predicted captions and ground-truth references.
- **CIDEr (C):** Computes TF-IDF–weighted n-gram cosine similarity to reward visually descriptive content while penalizing generic terms.
- **SPICE (S):** Converts captions into semantic scene graphs (objects, attributes, and relations) and calculates the F_1-score over the extracted graph tuples.

Retrieval. We evaluate both text-to-image (T2I) and image-to-text (I2T) retrieval on the COCO Karpathy 5K and Flickr30K Karpathy 1K test splits, reporting direct retrieval recall without post-hoc reranking.

Composed Retrieval. We use the MMEB-formatted CIRR test split (Jiang et al., 2025; Liu et al., 2021) as a zero-shot diagnostic. It contains 1,000 reference-image/instruction queries, each with a prescribed 1,000-image candidate set and one designated target. We evaluate the 8k-step T2I-adapted checkpoint with all model weights fixed. The complete modification instruction is encoded as a single text input, separately from the reference image. At each prefix length, we add corresponding tokens with fixed unit coefficients, $\mathbf{z}_{q,n} = \mathbf{z}_{\mathrm{img},n} + \mathbf{z}_{\mathrm{inst},n}$ for $n \leq K$, and ℓ_2-normalize each composed token. Candidate images are encoded and normalized independently, then ranked by the mean cosine similarity between aligned prefix positions, following Eq. 3.

Linear Probing. For linear probing on ImageNet-1K (Deng et al., 2009), images undergo a deterministic square center crop without random augmentation and are processed at the model's 512×512 input resolution. We cache frozen representations in fp16, then ℓ_2-normalize each register, concatenate the first K registers, and cast the resulting feature to fp32 when fitting the classifier. The probe is a single linear layer trained on all 1,281,167 ImageNet-1K training images for 20 epochs with AdamW, batch size 1,024, zero weight decay, and cosine learning-rate decay. We select the learning rate from $\{3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}\}$ and test an optional non-affine BatchNorm at $K{=}64$; the selected setting, learning rate 3×10^{-3} without BatchNorm, is then fixed for every K. We report top-1 accuracy on the complete 50,000-image validation split.

Clustering and Modality Geometry. For the unsupervised clustering diagnostic in Figure 6, we use 2,000 ImageNet validation images from 50 classes (40 images per class), reduce the concatenated prefix representation to 50 dimensions with PCA, and run k-means with 50 clusters. Class labels are used only to compute purity, NMI, and ARI after clustering. For Figure 7, we sample 500 matched COCO image–caption pairs. For modality m, we compute $\bar{\mathbf{z}}_i^m = \mathrm{norm}(K^{-1}\sum_{n=1}^{K}\mathrm{norm}(\mathbf{z}_{i,n}^m))$ and report $\|\boldsymbol{\mu}_{\mathrm{img}} - \boldsymbol{\mu}_{\mathrm{txt}}\|_2$, where $\boldsymbol{\mu}_m$ is the mean of $\bar{\mathbf{z}}_i^m$ over the 500 samples. CLIP and SigLIP2 use their normalized pooled embeddings in the same calculation. PCA is fitted jointly to the image and text vectors for visualization; the reported centroid distance is computed in the original feature space.

C Evaluation Results of the pre-trained checkpoint

The jointly pre-trained checkpoint supports generation and retrieval without task-specific adaptation. We first evaluate generation across both modalities, followed by bidirectional retrieval.

T2I Generation. Table 8 reports GenEval results using the same prompt, sampling, and fixed-noise protocol as the fine-tuned checkpoint. Without task-specific tuning, the pre-trained model reaches an overall score of 0.71, with performance scaling from 0.32 using a single token to 0.71 with 256 tokens.

Table 8: **GenEval for the pre-trained checkpoint.** No prompt rewriting is applied.

K	1	4	16	64	256
GenEval	0.317	0.613	0.703	0.701	**0.711**

I2T Generation. We evaluate the pre-trained checkpoint directly on I2T captioning using the MS-COCO Karpathy test split. As shown in Table 9, performance steadily improves with larger prefix lengths, reaching 63.8 CIDEr and 16.3 SPICE at $K{=}256$. Although these absolute scores are lower than those of the fine-tuned checkpoint (Table 2), this gap is largely driven by stylistic differences of training data: the pre-training corpus contains diverse descriptive text, whereas COCO favors short, literal captions. This distributional gap inherently penalizes overlap-based metrics, as semantically correct but longer or editorial descriptions introduce phrasing that diverges from compact reference captions.

Figure 11 qualitatively illustrates this phenomenon. Zero-shot generations reliably capture the central entities and actions, but often express them as detailed descriptions or title-style captions. For instance, at $K{=}64$, 99.6% of generated sequences properly terminate with an EOS token and average 10.8 words, yet 27.1% begin with quotation marks characteristic of image titles. These results

indicate that while pre-trained representations already encode rich visual semantics, fine-tuning primarily serves to align decoder length and stylistic tone with COCO expectations. Consequently, part of the metric gains observed after fine-tuning reflects output-format alignment rather than improved semantic understanding alone.

Table 9: **Zero-shot I2T captioning evaluation for the pre-trained checkpoint.** Greedy decoding is used to generate the captions.

K	B@4	M	R	C	S
1	11.0	17.9	37.2	46.1	12.5
4	13.4	20.2	39.8	57.2	14.6
16	14.5	21.3	41.3	62.1	15.7
64	14.9	21.6	41.6	63.5	16.2
256	**15.0**	**21.7**	**41.7**	**63.8**	**16.3**



COCO reference

A lone man sitting on a bench on a hill next to a large tree.

Zero-shot, $K = 64$

"Solitary figure under the tree, seeking solace in nature."

Fine-tuned, $K = 64$

A person sitting on a bench under a tree.



COCO reference

A painting of a table with fruit on top of it.

Zero-shot, $K = 64$

"Still life with fruit, pitcher, and candle in ornate frame."

Fine-tuned, $K = 64$

A painting of a pitcher, oranges and a candle.



COCO reference

A commercial airplane ascending into the sky.

Zero-shot, $K = 64$

A sleek airplane soars through a clear sky, leaving a trail of light.

Fine-tuned, $K = 64$

A large jetliner flying through a gray sky.



COCO reference

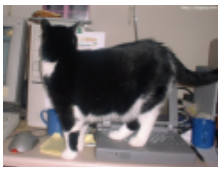
A black and white photo of a train sitting in a station.

Zero-shot, $K = 64$

"Classic train at station, black and white."

Fine-tuned, $K = 64$

A black and white photo of a train at a station.



COCO reference

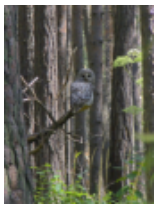
A black and white cat standing on a laptop.

Zero-shot, $K = 64$

A black and white cat stands on a laptop keyboard in an office setting.

Fine-tuned, $K = 64$

A black and white cat standing on a laptop.



COCO reference

An owl sitting on a branch in a forest.

Zero-shot, $K = 64$

A curious owl perches on a tree branch in a forest setting.

Fine-tuned, $K = 64$

An owl sitting on a branch in a forest.

Figure 11: **A comparison between the generated captions between pre-trained and fine-tuned checkpoints.** Each example compares a COCO reference caption against outputs from the pre-trained and fine-tuned checkpoints. While zero-shot predictions from the pre-trained model reliably capture core entities and actions, fine-tuning aligns output length and phrasing with the COCO annotation style.

Retrieval. Table 10 evaluates the pre-trained checkpoint on the full MS-COCO and Flickr30K Karpathy test splits without retrieval-specific fine-tuning. Even in this zero-shot setting, the learned

representations effectively support bidirectional retrieval. Zero-shot performance is strongest at smaller prefix lengths and decreases as K increases. Task-specific fine-tuning substantially mitigates this sensitivity to K (Table 3).

Table 10: **Zero-shot retrieval for the pre-trained checkpoint.** Results are recall (%) on the COCO Karpathy 5k and Flickr30k Karpathy 1k test splits.

	Image → text			**Text → image**		
K	R@1	R@5	R@10	R@1	R@5	R@10
MS-COCO Karpathy 5k						
1	**44.90**	**69.10**	**78.54**	**39.70**	**64.59**	**74.51**
4	44.48	68.26	78.08	37.85	62.31	72.08
16	41.94	66.96	76.32	36.32	61.02	70.84
64	39.96	65.02	74.64	34.33	59.07	69.37
256	35.48	60.04	70.64	31.45	55.87	66.57
Flickr30k Karpathy 1k						
1	**64.10**	**89.70**	**94.90**	**64.40**	**86.04**	**91.16**
4	63.50	89.10	94.60	61.52	84.62	89.76
16	61.00	89.00	94.60	60.44	83.98	89.30
64	59.30	88.10	93.50	58.68	83.00	88.64
256	55.70	84.30	91.90	55.74	81.30	87.20

D Effect of FLAT pre-training on task adaptation

We compare task-specific adaptation with and without FLAT pre-training. The task-specific training jobs use the same architecture, training data, optimization budget, and evaluation protocol as the corresponding fine-tuned FLAT models. The FLAT-specific parameters are randomly initialized instead of pre-trained. Table 11 reports the complete prefix-length sweep. For T2I synthesis, the task-specific model remains near 0.60 GenEval across all prefix lengths. FLAT pre-training improves GenEval by 0.16–0.24 for $K \geq 4$ and reaches 0.83 at $K{=}16$ and $K{=}64$. In contrast, the pre-trained initialization gains 0.34 from $K{=}1$ to $K{=}16$, showing that pre-training enables the image decoder to exploit wider prefixes for compositional generation. Pre-training also improves captioning by 9.0–10.9 BLEU-4 and 31.9–41.2 CIDEr across K. Pre-training also improves retrieval in both directions: image-to-text R@1 improves by 3.1–3.6 points and text-to-image R@1 improves by 1.6–2.1 points across K.

Table 11: **Task-specific adaptation with and without FLAT pre-training.** T2I uses matched 8k-step recipes on 120K image–text pairs; captioning and retrieval use matched 5k-step recipes on the COCO Karpathy train+restval split. T2I is evaluated on GenEval without prompt rewriting. Captioning and retrieval are evaluated on the Karpathy 5k test split.

Metric	Initialization	$K{=}1$	$K{=}4$	$K{=}16$	$K{=}64$	$K{=}256$
T2I GenEval	Random (No pre-training)	**0.60**	0.61	0.61	0.60	0.59
	Pre-trained	0.49	**0.77**	**0.83**	**0.83**	**0.82**
Captioning B@4	Random (No pre-training)	24.4	27.3	30.6	31.0	30.8
	Pre-trained	**35.3**	**38.1**	**39.6**	**40.2**	**40.5**
Captioning CIDEr	Random (No pre-training)	80.8	91.5	104.3	105.3	105.0
	Pre-trained	**122.0**	**131.0**	**136.2**	**138.1**	**138.6**
I2T retrieval R@1	Task-only	62.44	62.38	62.28	62.72	62.24
	FLAT pre-trained	**65.52**	**65.52**	**65.76**	**65.80**	**65.82**
T2I retrieval R@1	Task-only	48.62	48.54	48.49	48.53	48.43
	FLAT pre-trained	**50.19**	**50.55**	**50.42**	**50.42**	**50.49**

E QUALITATIVE RESULTS FOR FLAT

The following galleries present a qualitative evaluation of FLAT. Text-to-image samples share the same initial noise across prefix lengths, while retrieval evaluations utilize task-specific checkpoints and full galleries as described in Table 7. Figures 12 and 13 evaluate performance on controlled compositional prompts, whereas Figure 14 extends this comparison to longer, free-form prompts. Figure 15 illustrates bidirectional retrieval results, and Figure 16 highlights representative failure cases across the primary tasks.



Figure 12: **Text-to-image gallery, part I.** Samples are drawn from the fine-tuned checkpoint across the full prefix-length ladder. The initial noise is fixed within each row. These examples cover a single object, two-object composition, and counting.

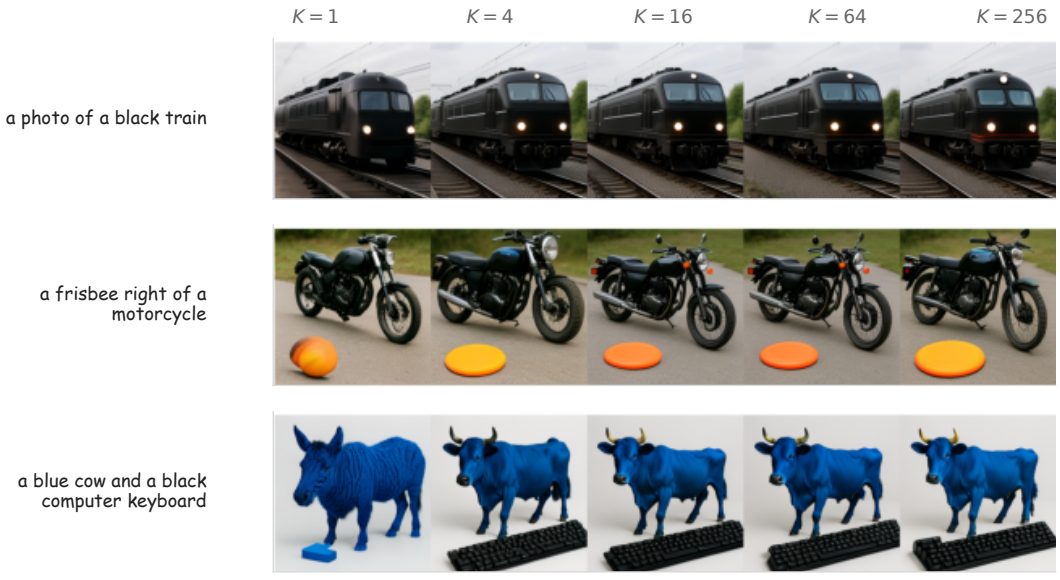


Figure 13: **Text-to-image gallery, part II.** Additional samples covering color, spatial relations, and attribute–object binding, with fixed initial noise across prefix lengths.

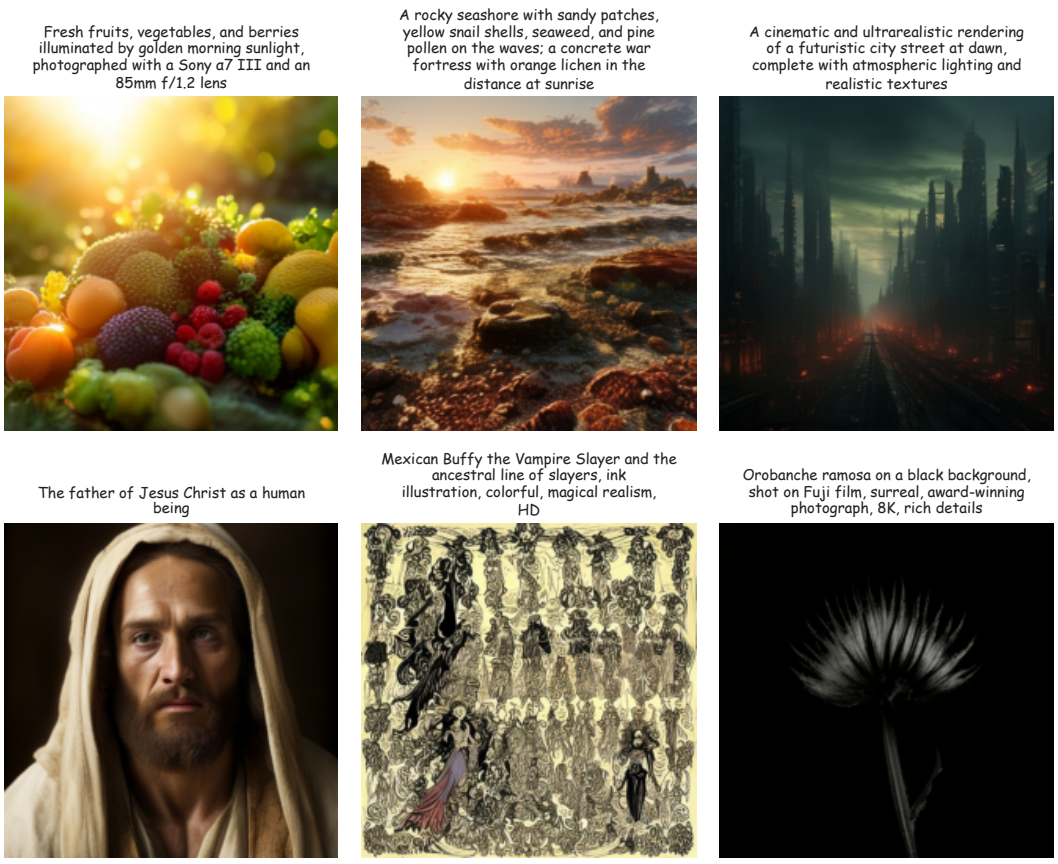


Figure 14: **Text-to-image gallery with free-form prompts.** Samples at $K{=}64$ for longer prompts from MJHQ-30K, spanning photographic and illustrative styles across diverse subjects and scenes.

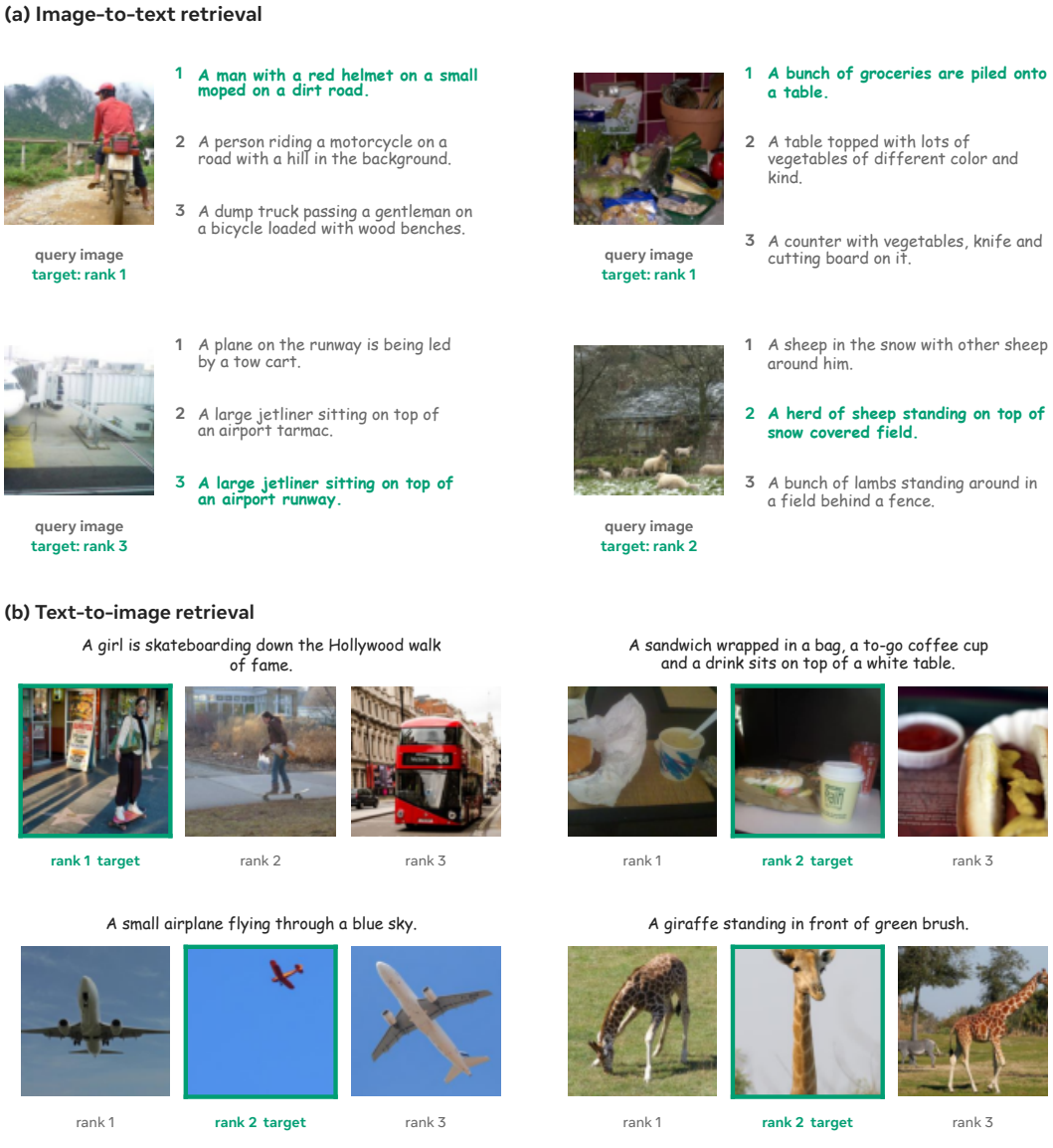


Figure 15: **Bidirectional retrieval galleries.** Image-to-text (top) and text-to-image (bottom) retrieval results at $K{=}16$ on MS-COCO. A green box highlights the ground-truth target, while remaining images/texts show top-ranked retrieved candidates. Cases where the ground truth appears at rank 2 or 3 illustrate near-miss failures, where leading alternatives remain semantically consistent with the query.

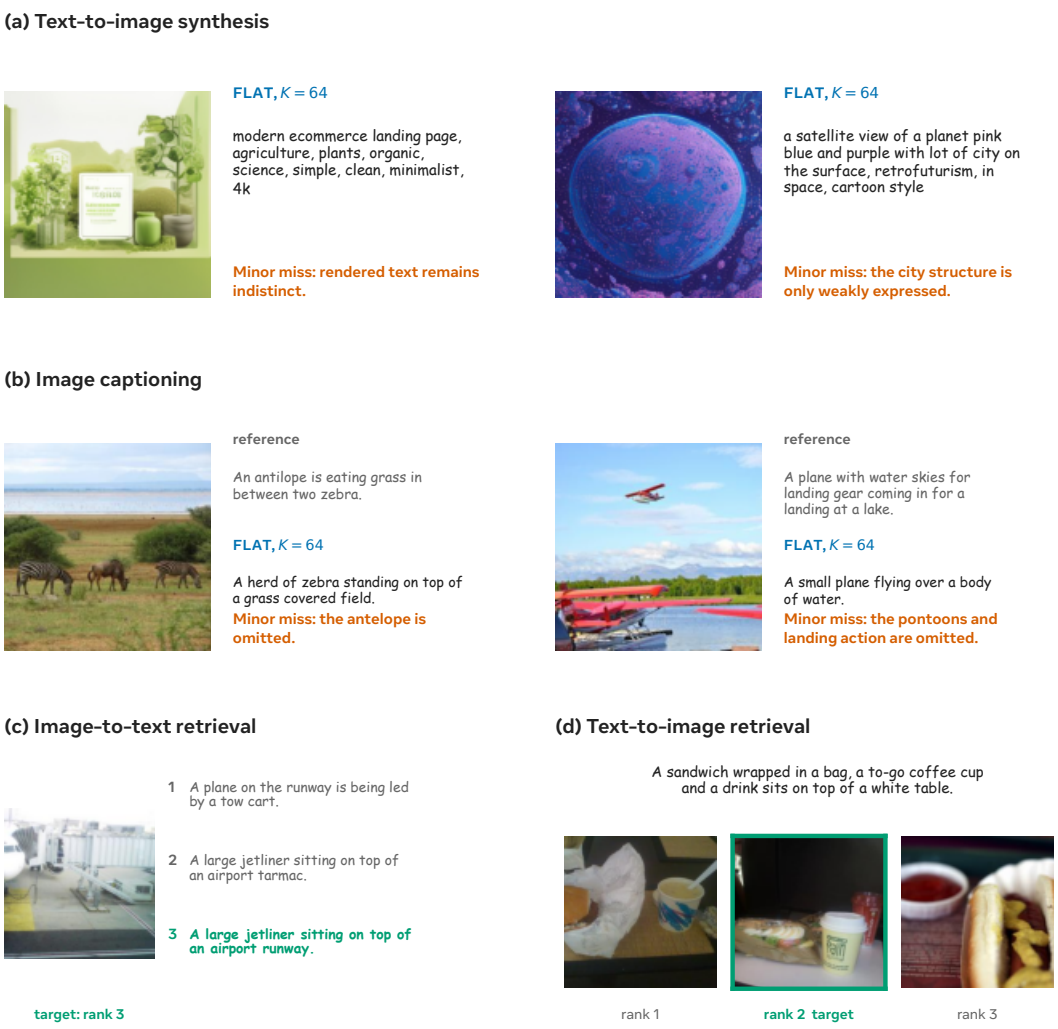


Figure 16: **Representative failures across tasks.** For text-to-image synthesis (top), the overall subject matter and style are preserved, though fine-grained layout or typography exhibits minor imperfections. For captioning (middle), the primary scene is described accurately, but secondary objects or actions are occasionally omitted. In bidirectional retrieval, ground-truth targets consistently rank near the top, preceded only by semantically close alternatives.

F FLEXIBLE-LENGTH REPRESENTATION ANALYSIS

During training, FLAT samples K only from a list $\mathcal{K} = \{1, 4, 16, 64, 256\}$. We examine whether arbitrary choices of K can be flexibly applied at test time.

F.1 GENERALIZATION TO UNSEEN PREFIX LENGTHS

In short, direct prefix truncation enables inference at arbitrary integer sequence lengths within 256. Figure 17 illustrates outputs generated from the same prompt and initial noise across both trained lengths and unseen intermediate lengths. Generations at unseen prefix lengths remain coherent and evolve smoothly with K.

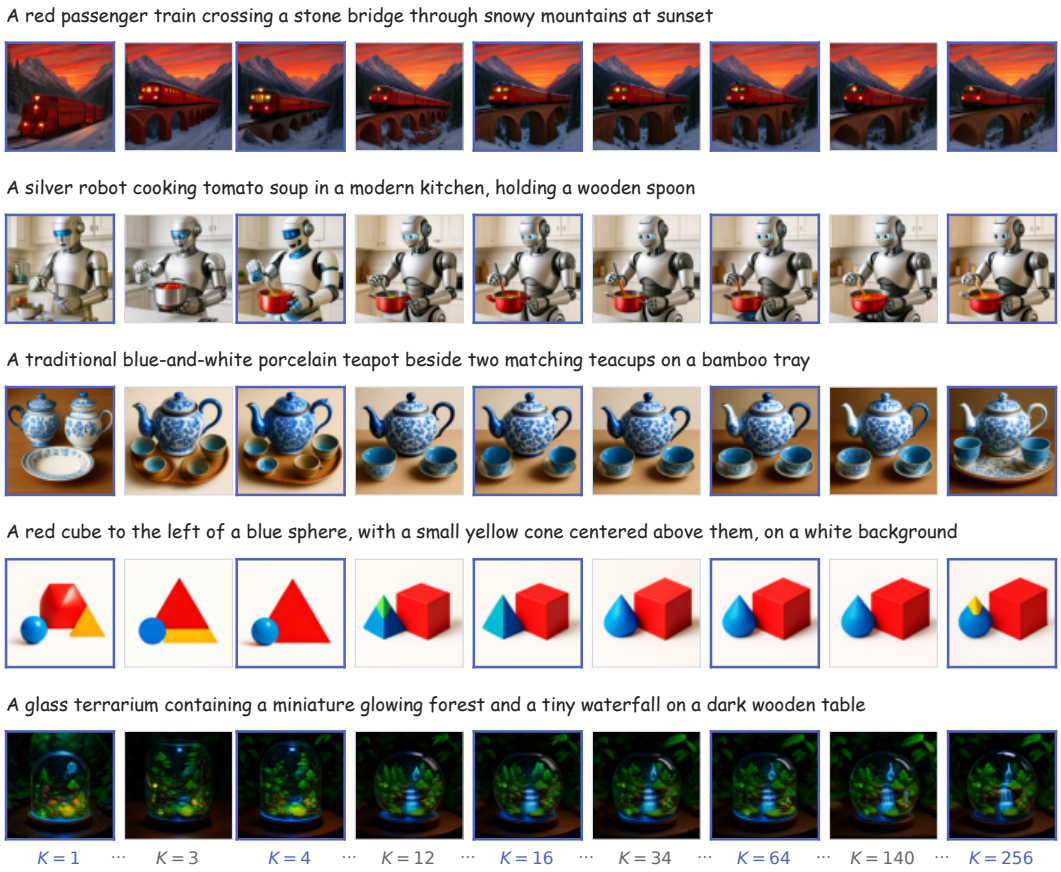


Figure 17: **Generation at trained and unseen prefix lengths.** Blue borders indicate sequence lengths sampled during training, whereas gray borders denote unseen intermediate lengths. All columns share the same prompt, sampling configuration, and initial noise vector.

F.2 PREFIX-DEPENDENT COMPUTE

Because the representation encoder is causal, the hidden states of the first K registers are independent of subsequent register positions. At inference time, we can therefore process only the first K registers rather than encoding all $N = 256$ registers and truncating post-hoc. Table 12 reports single-batch inference FLOPs for the encoder and both decoders, counting each multiply–add operation as two FLOPs. For the representation encoder, we report image and text inputs separately. The image pass uses a 512×512 input that produces 256 visual tokens and 18 instruction/chat tokens before the registers, while the text pass uses a 128-token sequence including its instruction scaffold. Reducing K from 256 to 1 lowers image-encoder and text-encoder compute by 32.3% and 66.6%, respectively. The fixed vision-tower cost moderates the image-side reduction, whereas registers ac-

count for a larger fraction of the text sequence. The corresponding reductions are 71.1% for the I2T decoder and 13.8% for the T2I decoder, whose cost is dominated by spatial denoising and VAE decoding.

Table 12: **Encoder and decoder compute under prefix truncation.** Absolute cost is reported in TFLOPs; relative cost normalizes each component to its K=256 setting. Encoder measurements include the vision tower when present, the causal language backbone, the first K registers, and the 64-dimensional latent projection. The image and text inputs contain 274 and 128 content tokens before the registers, respectively; all passes use bfloat16 with eager attention. I2T uses cached greedy decoding for 64 output tokens. T2I uses 20 denoising steps at 512 px with CFG 4.5 (two denoiser calls per step), including semantic projection, connector, denoiser, and VAE decoding.

K	Encoder (Image)		Encoder (Text)		I2T decoder		T2I decoder	
	TFLOPs	Rel.	TFLOPs	Rel.	TFLOPs	Rel.	TFLOPs	Rel.
1	1.502	67.7%	0.358	33.4%	0.289	28.9%	29.964	86.2%
4	1.511	68.1%	0.366	34.2%	0.298	29.7%	30.020	86.4%
16	1.544	69.6%	0.400	37.3%	0.331	33.1%	30.245	87.1%
64	1.679	75.7%	0.533	49.8%	0.465	46.4%	31.144	89.6%
256	2.218	100.0%	1.070	100.0%	1.002	100.0%	34.741	100.0%

F.3 Prefix-length sampling strategies

We sample the prefix K from the ladder $\mathcal{K} = \{1, 4, 16, 64, 256\}$ using three strategies: 1) **Uniform** assigns equal probability to every K; 2) **Geometric** samples K_r with $p(K_r) \propto \beta^r$, where $\beta = 1.3$; and 3) **Replay** spends half of training on shorter prefixes and the other half on the full prefix.

Figure 18 compares the training loss through 40k steps. Uniform and Geometric sample every K throughout training, whereas Replay progressively unlocks longer prefixes.

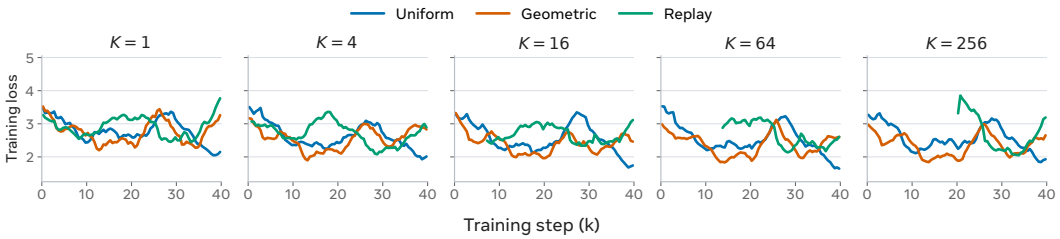


Figure 18: **Training loss under different prefix-length sampling strategies.** Training lasts 40k steps.

Metric	Sampling	K=1	K=4	K=16	K=64	K=256
GenEval (×100) ↑	Uniform	**33.0**	32.5	32.4	31.9	32.7
	Geometric	26.0	**43.2**	**49.9**	**49.9**	**50.2**
	Replay	32.2	32.6	32.2	32.0	32.3
I2T R@1 ↑	Uniform	57.8	53.9	54.3	52.9	**52.2**
	Geometric	56.1	52.6	53.1	52.2	51.8
	Replay	**59.1**	**55.7**	**55.8**	**54.5**	52.1
T2I R@1 ↑	Uniform	**62.4**	**60.1**	58.4	56.7	56.3
	Geometric	60.9	56.9	56.6	55.8	55.5
	Replay	62.2	**60.1**	**59.7**	**59.2**	**59.1**
CIDEr ↑	Uniform	122.7	130.4	136.8	**137.6**	137.5
	Geometric	**123.0**	129.6	**137.0**	137.3	**138.3**
	Replay	121.5	**131.3**	136.5	137.5	137.5

Table 13: **Downstream performance under different prefix-length sampling strategies.**

Table 13 evaluates matched 40k-step checkpoints for different sampling strategies. As shown, Geometric favors generation for $K \geq 4$, Replay favors retrieval. However, no single strategy is universally optimal across all metrics and values of K.

F.4 LATENT INTERPOLATION

Figures 19 and 20 examine latent interpolation through both decoders. Figure 19 compares five prefix lengths, while Figure 20 adds six $K{=}256$ text–text walks; captions are condensed to keywords.

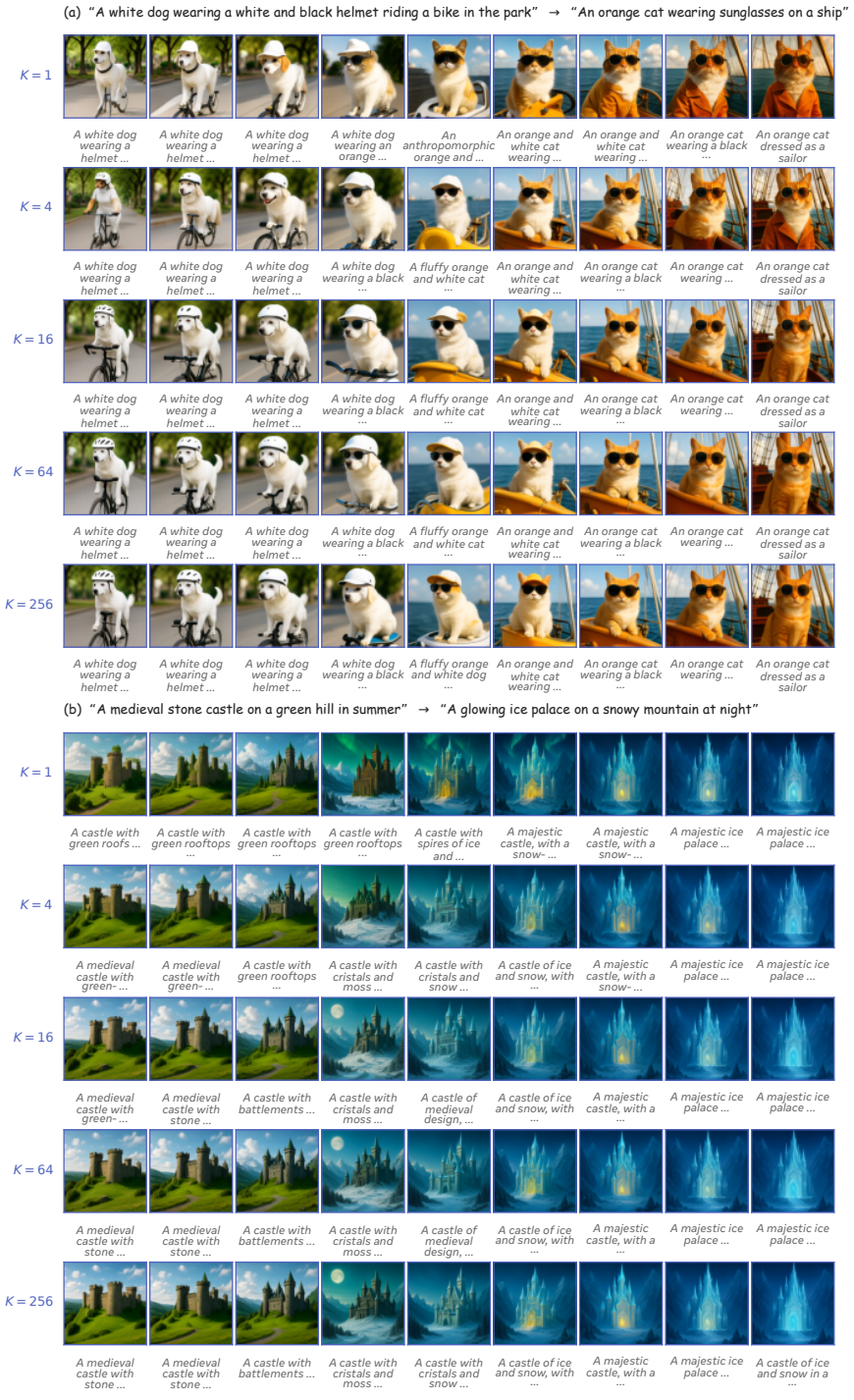


Figure 19: **Latent walks across prefix lengths.** Two text-prompt endpoint pairs are interpolated at $K \in \{1, 4, 16, 64, 256\}$. Italic text under every image is decoded directly from the corresponding intermediate representation by the I2T decoder.

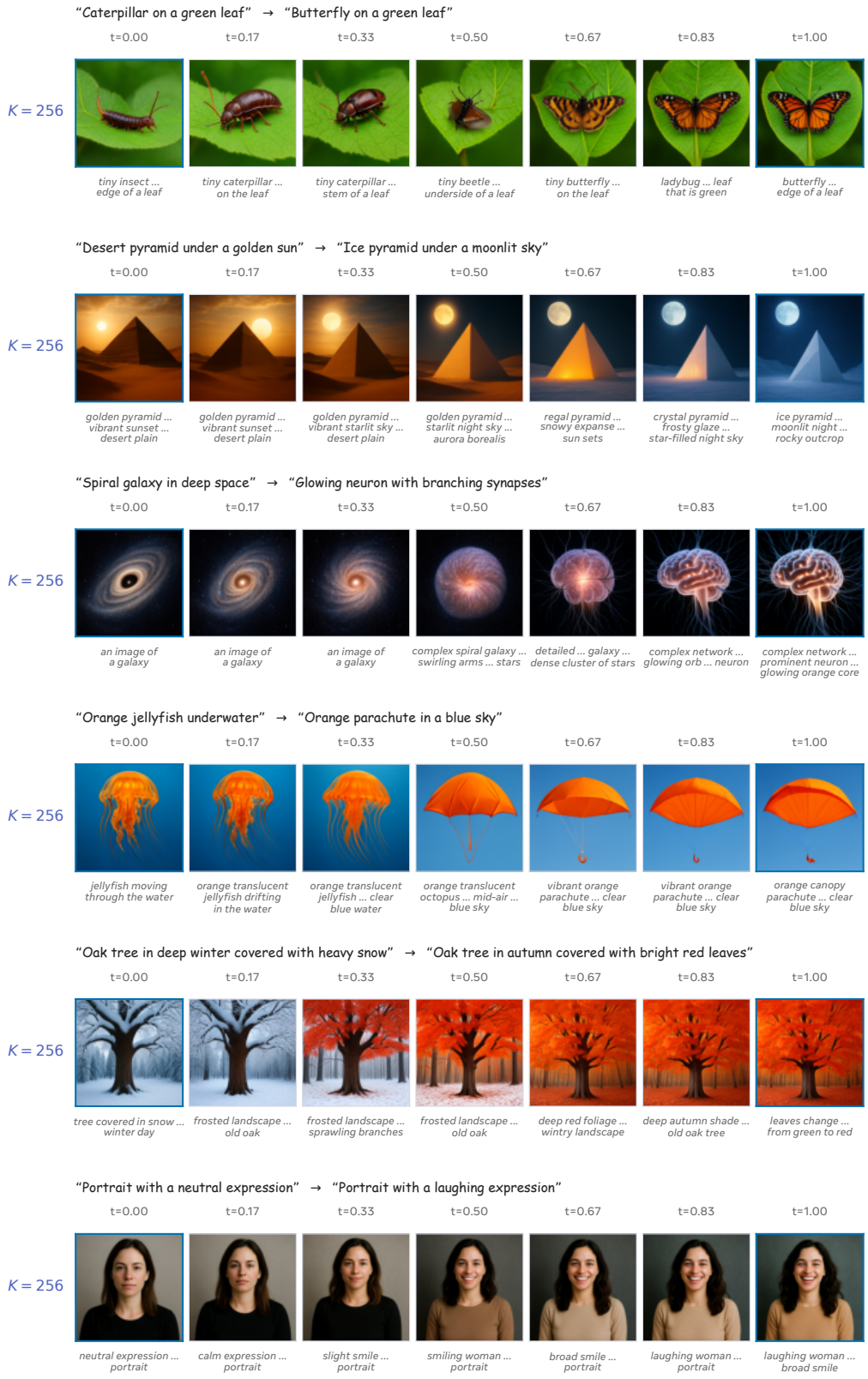


Figure 20: **Additional latent walks at K=256.** Blue borders mark the two endpoints. Italic text under every image is produced by the I2T decoder.

Interpolation protocol. Each endpoint is encoded once, and every intermediate representation is passed directly through the task-specific image and text projections; generated images are not re-encoded before captioning. The text–text walk in Figure 8 uses the Euclidean interpolation defined in Sec. 4.4. Its image–text walk and the appendix examples use per-register spherical interpolation: register directions follow a spherical path while their endpoint norms are linearly interpolated. No normalization is applied after Euclidean interpolation. Image synthesis uses 20 steps and CFG 4.5, with the initial noise fixed within each walk.

G ZERO-SHOT TRANSFER AND LATENT COMPOSITION

In this section, we demonstrate that the pre-trained FLAT representation natively supports zero-shot multilingual text understanding, multi-frame video understanding, and composed retrieval.

G.1 MULTILINGUAL TEXT AND EMOJI UNDERSTANDING

We first show that the representation encoder inherits multilingual capabilities from its LLM backbone. Consequently, non-English prompts occupy a similar representation space to English, despite the model being trained exclusively on English data. Figure 21 demonstrates that Spanish and Chinese prompts decode into images consistent with their English counterparts, while emoji sequences reliably decode into their corresponding visual referents.

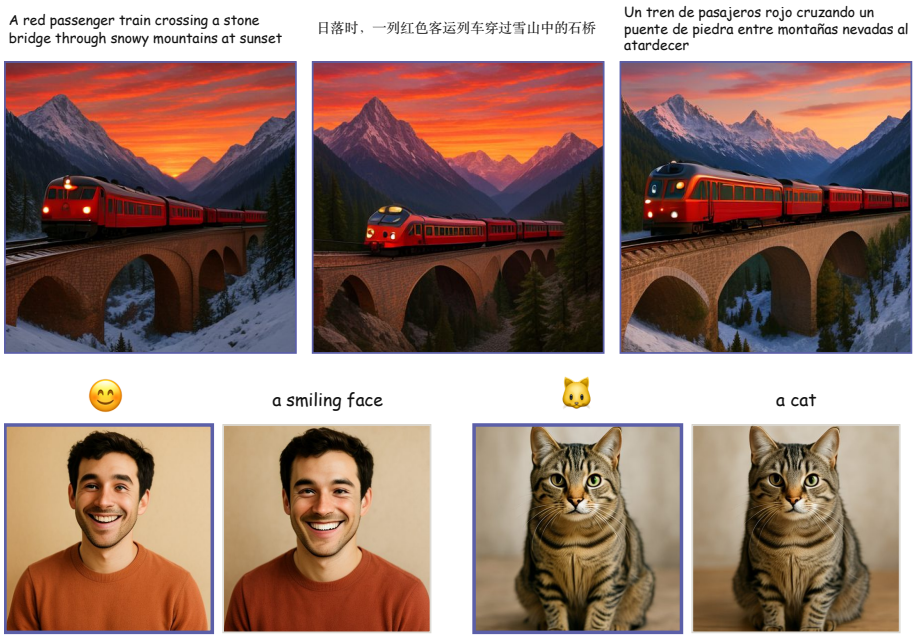


Figure 21: **Zero-shot multilingual text and emoji conditioned image generation.** Although trained exclusively on English data, FLAT generalizes its understanding to non-English languages and emojis.

G.2 MULTI-FRAME VIDEO UNDERSTANDING

We next evaluate FLAT's ability to process multi-frame video inputs despite being trained exclusively on static images. For each MSR-VTT (Xu et al., 2016) clip, we uniformly sample 4 frames in chronological order, resize each frame to the encoder's fixed image resolution, and place the frames as four image items in one user message. The standard Qwen processor inserts the image-boundary and positional metadata; the resulting frame tokens are concatenated in one context, followed by the shared register sequence. A single encoder pass produces the representation, and its first $K{=}64$ registers condition the unchanged image-generation head. No video-specific module or parameter

update is introduced. Figure 22 shows examples from an aquarium scene and a cooking sequence. The generated outputs preserve persistent objects and scene context while integrating information across the sampled frames.

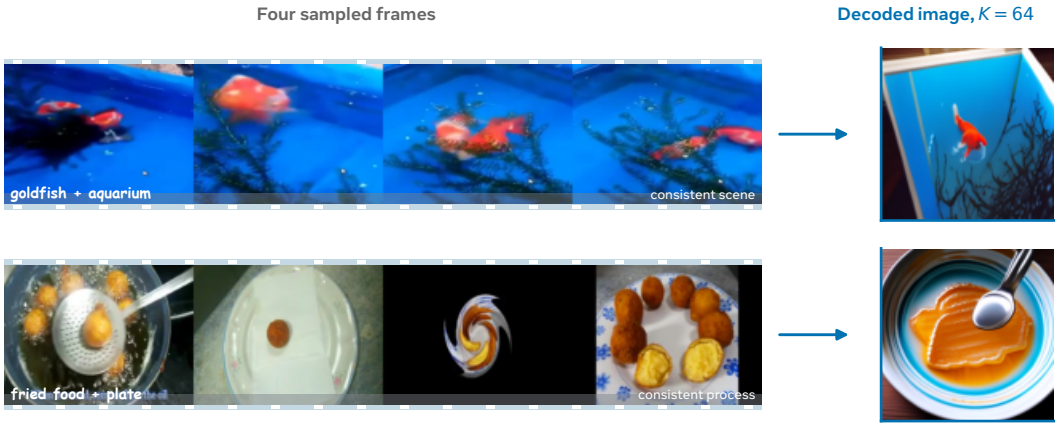


Figure 22: **Zero-shot multi-frame video semantic aggregation.** Each row jointly encodes 4 sampled frames and decodes one image at K=64. The generated outputs preserve persistent objects and scene context across the sampled frames.

G.3 Composed retrieval

This experiment tests whether FLAT's shared latent space supports composed retrieval through the same token-wise arithmetic used for semantic editing in the main paper. For every query and prefix length, we apply the fixed operator $\mathbf{z}_q = \mathbf{z}_{\text{img}} + \mathbf{z}_{\text{inst}}$ and rank the designated target within its 1,000-image candidate set. Table 14 reports Hit@k across prefix lengths. The first register gives the strongest ranking, reaching 23.4 Hit@1 and 79.8 Hit@10; the remaining prefixes also retrieve the target within the top 10 for 70.9–77.4% of queries. These results show that the latent directions induced by complete editing instructions can be composed directly with visual representations. Figure 23 illustrates successful object, attribute, pose, and scene modifications at K=256.

Table 14: **Zero-shot CIRR composed retrieval through direct feature addition.** The designated target is ranked within the prescribed 1,000-image candidate set for each query. The reference image and complete modification instruction are encoded separately and added token-wise with fixed unit coefficients.

Method	K	Hit@1	Hit@5	Hit@10
In-domain				
UniIR (Wei et al., 2024)	–	**51.3**	–	–
VLM2Vec (Jiang et al., 2025)	–	49.9	–	–
Zero-shot				
CLIP (Radford et al., 2021)	–	12.6	–	–
OpenCLIP (Cherti et al., 2023)	–	15.4	–	–
SigLIP (Zhai et al., 2023)	–	15.1	–	–
BLIP-2 (Li et al., 2023a)	–	9.8	–	–
E5-V (Jiang et al., 2024)	–	6.1	–	–
FLAT	1	**23.4**	**67.9**	**79.8**
	4	18.0	63.7	77.4
	16	13.9	55.5	72.7
	64	14.2	55.3	70.9
	256	17.3	60.8	74.4

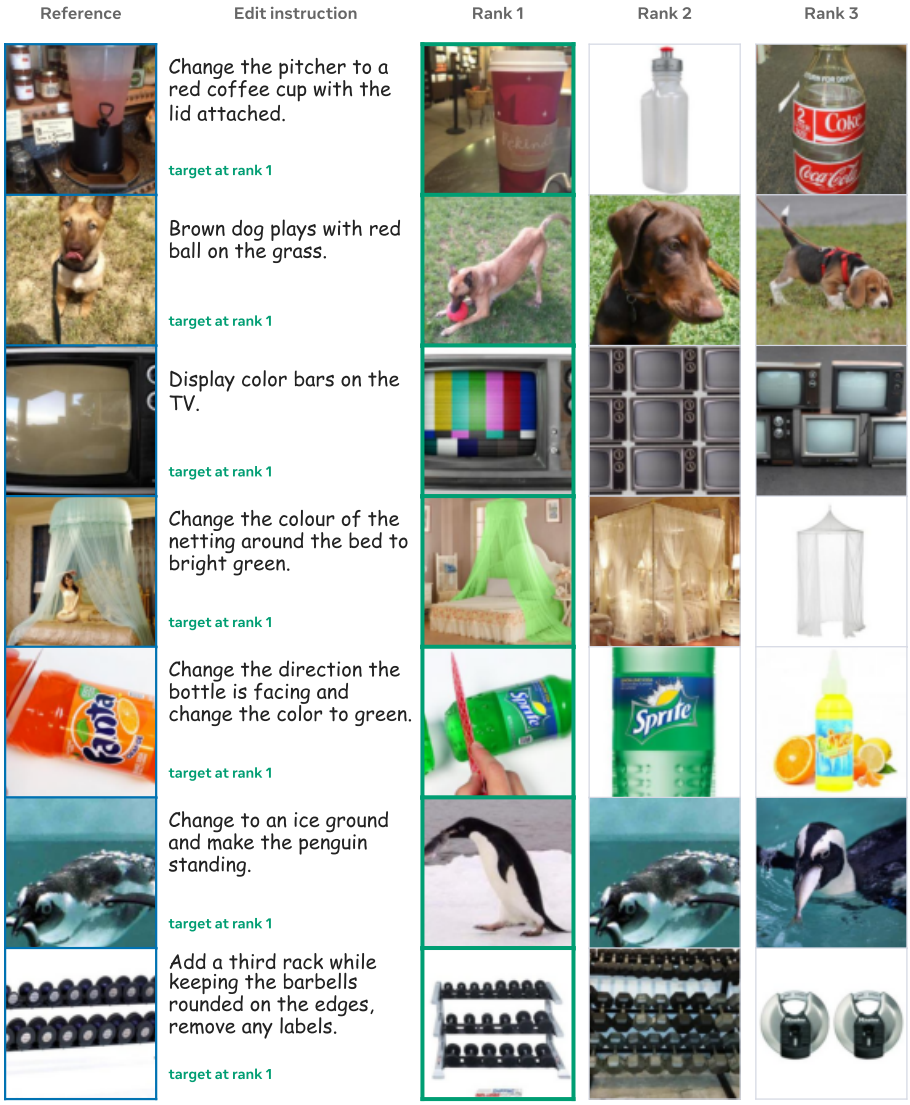


Figure 23: **Examples of zero-shot composed image retrieval through direct feature addition.** The separately encoded reference image and editing instruction are added token-wise. Green borders indicate ground-truth targets.

H Ablation on the training loss contributions

FLAT pre-training incorporates three loss terms targeting retrieval, T2I generation, and I2T generation. We perform an ablation study by training models on different subsets of these losses to evaluate their individual and joint contributions. Table 15 reports downstream scores across all evaluated prefix lengths and loss combinations. Each variant uses the same initialization and 40k-step schedule, varying only the active loss components. Incorporating all three losses provides a clear advantage, yielding balanced, optimal performance across all metrics and prefix lengths K. Figure 24 plots the normalized evaluation scores for each ablated variant across all prefix lengths.

Table 15: **Training loss ablation using all prefix lengths.** $\mathcal{L}_{\text{ctr}}$, $\mathcal{L}_{\text{cap}}$, and $\mathcal{L}_{\text{gen}}$ denote contrastive, captioning, and image-generation loss respectively. The final block averages each raw score over $K \in \{1, 4, 16, 64, 256\}$.

K=1

Objectives	I2T	T2I	B@4	CIDEr	Gen.
$\emptyset$	0.1	0.2	24.4	80.8	0.000
$\mathcal{L}_{\text{ctr}}$	50.7	57.2	33.9	117.3	0.001
$\mathcal{L}_{\text{cap}}$	48.2	46.3	35.9	123.3	0.004
$\mathcal{L}_{\text{gen}}$	41.7	49.6	32.1	106.6	0.328
$\mathcal{L}_{\text{ctr}} + \mathcal{L}_{\text{cap}}$	56.1	59.2	35.3	122.2	0.004
$\mathcal{L}_{\text{ctr}} + \mathcal{L}_{\text{gen}}$	52.6	54.1	33.8	117.2	0.280
$\mathcal{L}_{\text{cap}} + \mathcal{L}_{\text{gen}}$	54.2	55.4	36.1	123.9	0.298
All three	57.8	62.4	35.4	122.7	0.329

K=4

Objectives	I2T	T2I	B@4	CIDEr	Gen.
$\emptyset$	0.1	0.0	27.3	91.5	0.000
$\mathcal{L}_{\text{ctr}}$	50.8	57.0	34.1	118.1	0.002
$\mathcal{L}_{\text{cap}}$	26.4	21.2	38.7	133.7	0.005
$\mathcal{L}_{\text{gen}}$	41.8	50.1	32.1	107.0	0.323
$\mathcal{L}_{\text{ctr}} + \mathcal{L}_{\text{cap}}$	55.1	56.2	38.3	131.9	0.005
$\mathcal{L}_{\text{ctr}} + \mathcal{L}_{\text{gen}}$	52.4	54.7	34.1	117.4	0.309
$\mathcal{L}_{\text{cap}} + \mathcal{L}_{\text{gen}}$	36.2	36.6	38.6	133.1	0.298
All three	53.9	60.1	37.6	130.4	0.325

K=16

Objectives	I2T	T2I	B@4	CIDEr	Gen.
$\emptyset$	0.2	0.0	30.6	104.3	0.000
$\mathcal{L}_{\text{ctr}}$	50.4	56.7	34.1	117.7	0.002
$\mathcal{L}_{\text{cap}}$	28.0	19.8	40.3	138.3	0.004
$\mathcal{L}_{\text{gen}}$	10.1	40.6	32.9	110.4	0.325
$\mathcal{L}_{\text{ctr}} + \mathcal{L}_{\text{cap}}$	53.5	56.0	40.0	137.2	0.003
$\mathcal{L}_{\text{ctr}} + \mathcal{L}_{\text{gen}}$	51.5	53.4	33.9	117.0	0.328
$\mathcal{L}_{\text{cap}} + \mathcal{L}_{\text{gen}}$	41.6	41.0	40.1	137.8	0.294
All three	54.3	58.4	39.7	136.8	0.324

K=64

Objectives	I2T	T2I	B@4	CIDEr	Gen.
$\emptyset$	0.2	0.0	31.0	105.3	0.000
$\mathcal{L}_{\text{ctr}}$	50.4	56.9	34.0	117.7	0.003
$\mathcal{L}_{\text{cap}}$	35.6	21.0	40.7	139.3	0.004
$\mathcal{L}_{\text{gen}}$	1.1	4.4	33.0	110.4	0.327
$\mathcal{L}_{\text{ctr}} + \mathcal{L}_{\text{cap}}$	52.7	54.8	39.9	137.8	0.004
$\mathcal{L}_{\text{ctr}} + \mathcal{L}_{\text{gen}}$	51.7	53.7	33.8	116.8	0.299
$\mathcal{L}_{\text{cap}} + \mathcal{L}_{\text{gen}}$	42.7	37.3	40.5	138.7	0.296
All three	52.9	56.7	40.1	137.5	0.319

K=256

Objectives	I2T	T2I	B@4	CIDEr	Gen.
$\emptyset$	0.2	0.0	30.8	105.0	0.000
$\mathcal{L}_{\text{ctr}}$	51.0	57.0	34.2	118.0	0.003
$\mathcal{L}_{\text{cap}}$	39.4	4.0	40.8	139.7	0.004
$\mathcal{L}_{\text{gen}}$	0.6	1.1	33.1	110.8	0.323
$\mathcal{L}_{\text{ctr}} + \mathcal{L}_{\text{cap}}$	51.2	54.5	40.0	137.8	0.003
$\mathcal{L}_{\text{ctr}} + \mathcal{L}_{\text{gen}}$	51.5	53.6	33.9	117.1	0.329
$\mathcal{L}_{\text{cap}} + \mathcal{L}_{\text{gen}}$	13.6	6.7	40.7	139.2	0.297
All three	52.2	56.3	40.2	137.5	0.327

Overall (average over K)

Objectives	I2T	T2I	B@4	CIDEr	Gen.
$\emptyset$	0.2	0.0	28.8	97.4	0.000
$\mathcal{L}_{\text{ctr}}$	50.7	57.0	34.0	117.8	0.002
$\mathcal{L}_{\text{cap}}$	35.5	22.5	39.3	134.9	0.004
$\mathcal{L}_{\text{gen}}$	19.1	29.2	32.6	109.0	0.325
$\mathcal{L}_{\text{ctr}} + \mathcal{L}_{\text{cap}}$	53.7	56.1	38.7	133.4	0.004
$\mathcal{L}_{\text{ctr}} + \mathcal{L}_{\text{gen}}$	51.9	53.9	33.9	117.1	0.309
$\mathcal{L}_{\text{cap}} + \mathcal{L}_{\text{gen}}$	37.7	35.4	39.2	134.5	0.296
All three	54.2	58.8	38.6	133.0	0.325

We also compute Shapley values (Shapley, 1953) to quantify the marginal contribution of each loss objective. For a loss component i and downstream metric v, the Shapley attribution is defined as:

$$\phi_i(v) = \sum_{S \subseteq \mathcal{O} \setminus \{i\}} \frac{|S|!(|\mathcal{O}| - |S| - 1)!}{|\mathcal{O}|!} \left[v(S \cup \{i\}) - v(S) \right], \tag{10}$$

where $\mathcal{O}$ is the set of all three losses. Figure 25 visualizes these Shapley values across prefix lengths. The heatmaps normalize each attribution by $v(\mathcal{O}) - v(\emptyset)$, ensuring that each metric column sums to 100%. The task-aligned attribution structure becomes sharper as the prefix widens: contrastive learning increasingly accounts for retrieval performance, captioning dominates both text-generation metrics from K=16 onward, and image generation consistently accounts for at least 95% of GenEval. Minor negative entries at K=16 and K=256 indicate that auxiliary loss contributions are not uniformly positive at every prefix length. For the Overall panel, we first average the coalition scores across all five values of K and then recompute the Shapley values; all resulting contributions are positive.

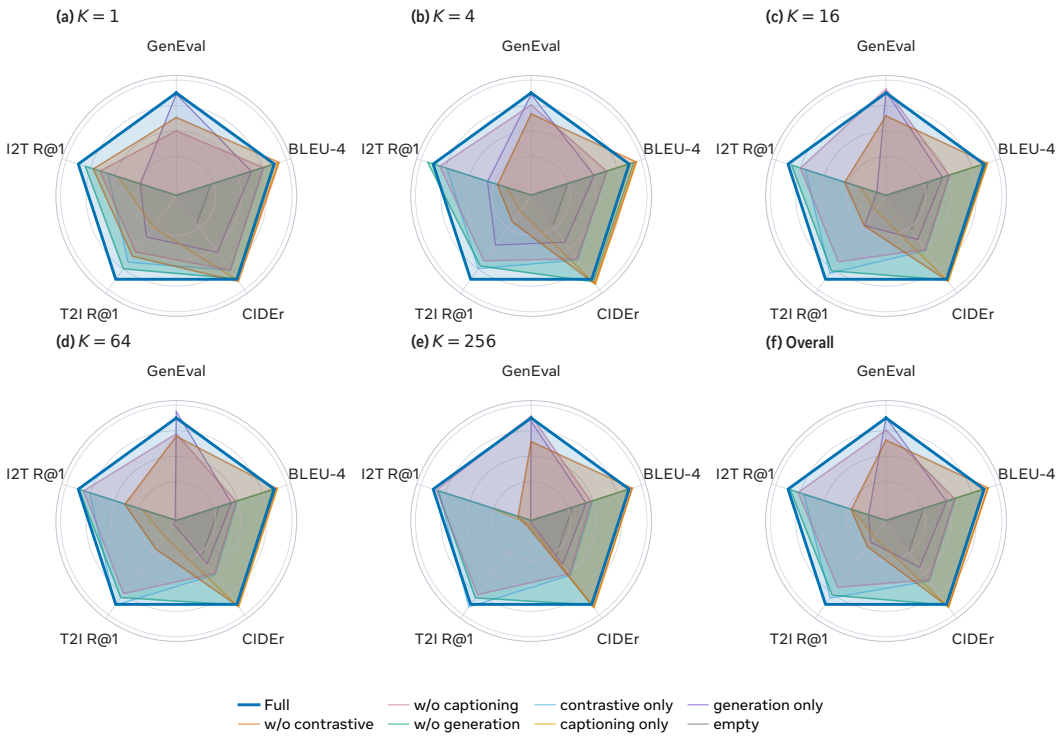


Figure 24: **Ablation on training loss. Eval performance at $K \in \{1, 4, 16, 64, 256\}$ and its average.** Each axis is normalized within the corresponding panel. The Overall panel first averages each loss combination's raw score across all five prefix lengths.

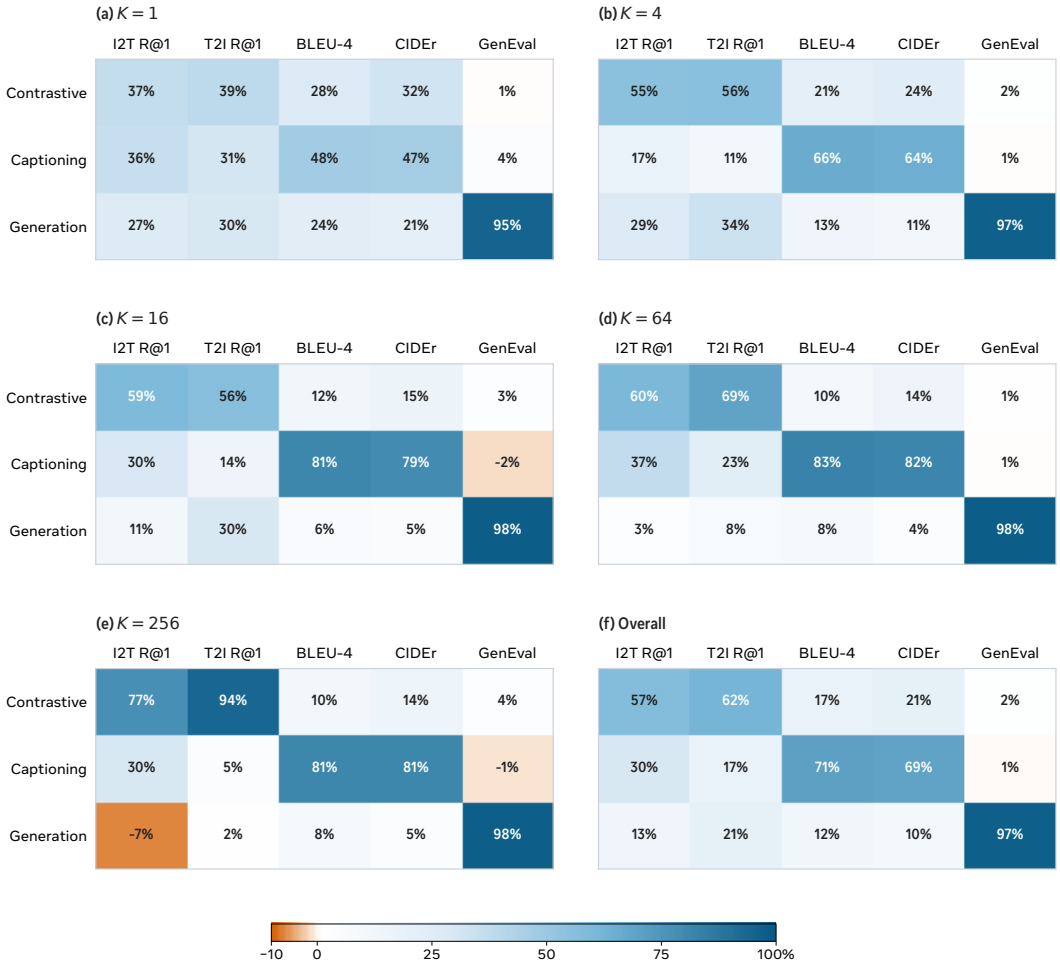


Figure 25: **Relative Shapley attribution at different prefix lengths and overall.** Cell values are percentages of the total improvement over the untrained checkpoint for each metric. Overall attributions are recomputed from coalition scores averaged across all five prefix lengths. Blue denotes positive attribution and orange denotes negative attribution.

I TRUNCATED PREFIXES AND NULL PADDING

Section 3.1 describes truncating the register sequence to its first K positions. An alternative strategy retains a fixed sequence length N and replaces positions $i \geq K$ with learned null embeddings, as in FlexTok (Bachmann et al., 2025). Empirically, we find that null-padded token sequences introduce optimization instabilities during FLAT joint training. Figure 26 illustrates these training dynamics. Under null padding, the gradient norm grows rapidly and spikes near 33k steps, causing the contrastive loss to collapse toward $\ln(1024)$—the theoretical random-baseline value for the global batch. In contrast, prefix truncation remains stable throughout the same training window. As shown in Figure 27, null-padded conditioning loses semantic correspondence between the source image and its generated counterpart, whereas prefix truncation preserves recognizable semantic relationships. Because these runs follow their respective original training recipes (including differing K-sampling granularities and learning-rate schedules), we interpret these findings as qualitative evidence of training dynamics.

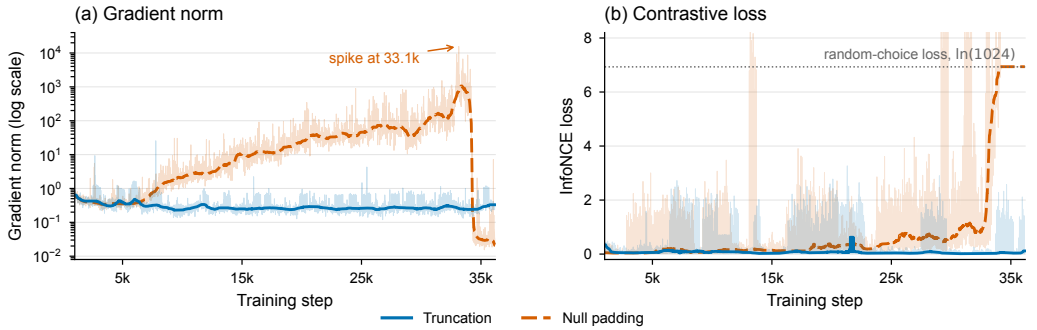


Figure 26: **Optimization behavior under truncation and null padding.** Curves are smoothed with a 1,010-step rolling median; faint lines show the unsmoothed measurements. The null-padding run develops large gradient excursions before its contrastive loss approaches the random-choice value.

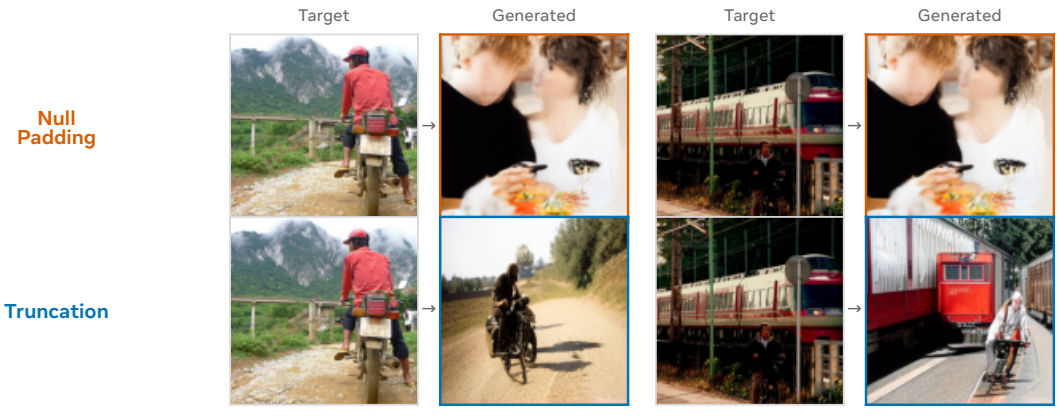


Figure 27: **Images generated by prefix truncation v.s. null padding at step 40k.** Each column uses the same COCO target–caption pair and diffusion noise.