GAMUT: UNIFIED VISUAL REPRESENTATION FOR GENERATION AND MULTIMODAL ALIGNMENT

Anonymous authors
Paper under double-blind review

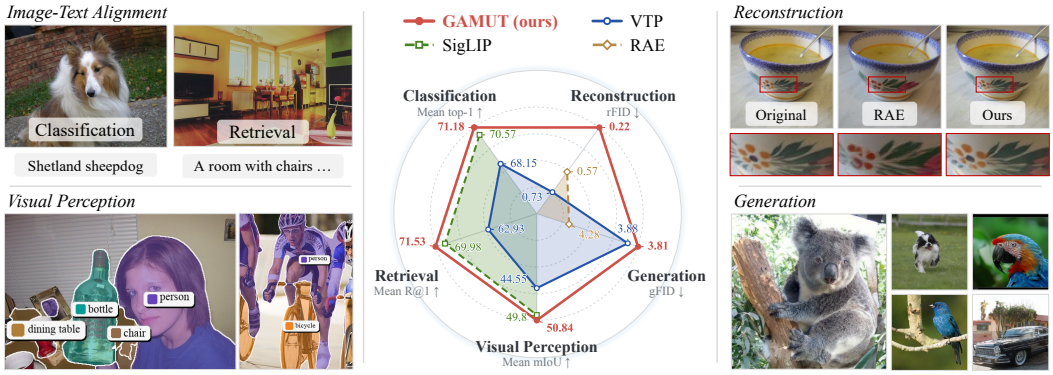


Figure 1: **One representation for the full gamut of visual tasks.** Through joint pretraining, GAMUT learns a unified visual representation that supports both understanding and generation.

ABSTRACT

Visual representation learning has evolved through largely separate paradigms, including self-supervised learning, multimodal alignment, and visual tokenization. Despite substantial progress within each paradigm, this separation is increasingly at odds with the demands of modern vision and multimodal systems, which require diverse information to support open-ended tasks rather than representations specialized for a single capability. To address this issue, we introduce GAMUT, which learns a unified visual representation and brings these paradigms into a single joint pretraining framework. We develop three key techniques to ensure stable and efficient joint pretraining. First, we construct a three-stage training pipeline that progressively shapes the representation and avoid the prohibitive computational cost of prolonged multi-objective training. Second, we propose the latent channel regularization, which effectively mitigates the channel-wise energy distribution imbalance and prevents representation collapse. Third, we introduce latent interpolation decoding to further improve the suitability of such representation for generative tasks. Together, these designs enable GAMUT to capture high-level semantics while preserving fine-grained visual details. As a result, GAMUT performs strongly across a diverse range of visual tasks, outperforming existing unified tokenizers and task-specific baselines on multiple benchmarks.

1 INTRODUCTION

Visual representations are the foundation of modern vision and multimodal systems, and have evolved along separate learning paradigms with distinct strengths. Multimodal alignment methods, such as CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023), learn vision-language correspondences that support open-vocabulary recognition and multimodal understanding. Self-supervision methods, such as MAE (He et al., 2022) and DINO (Caron et al., 2021; Oquab et al., 2024), capture spatial structure useful for recognition and dense perception. Visual tokenization methods use continuous (Kingma & Welling, 2014; Rombach et al., 2022) or discrete (van den Oord et al., 2017;

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.

Esser et al., 2021; Yu et al., 2024) autoencoders that support high-fidelity reconstruction to compress pixel-space vision data and provide efficient latent spaces for generative modeling.

Although these paradigms have each made significant progress, their task-specific visual representations are becoming less desirable in latest vision and multimodal systems. Traditionally, visual understanding and generation are treated as distinct problems: the former prioritizes high-level semantic abstraction, whereas the latter emphasizes modeling low-level visual details. This division, however, is increasingly at odds with contemporary tasks, whose boundaries and information requirements substantially overlap. Visual understanding tasks such as GUI Agent (Xie et al., 2024) depend on precise visual details, while conditional generation and image editing (Wu et al., 2025) also require semantic understanding of object relationships and scene context. However, existing representations tailored to one objective often tend to discard the properties essential to the other, preventing them from meeting the full requirements of downstream systems and tasks.

To address this issue, recent studies have begun exploring unified, general-purpose visual representations. One line of research seeks to **adapt semantic representations into latents suitable for generation**. RAE (Zheng et al., 2026) constructs diffusion-based generative models (Ho et al., 2020) directly on the frozen pretrained encoders such as DINOv2 (Oquab et al., 2024), demonstrating that high-dimensional semantic representations can also serve as effective latents for diffusion. However, this paradigm is constrained by the rigidity of frozen representations, such as the poor preservation of visual details, preventing it from achieving fully comprehensive. Although the subsequent work PS-VAE (Zhang et al., 2026b) improves detail preservation through further fine-tuning, it introduces a VAE bottleneck to perform compression, sacrificing the rich semantics in the high-dimensional representations. PS-VAE also shows that eliminating compression can cause collapse and ultimately impair generation. Conversely, another line of research explores **enhancing the semantic modeling capabilities of generative tokenizers**. VA-VAE (Yao et al., 2025) introduces the alignment with pretrained encoders as an additional objective into the VAE training process. VTP (Yao et al., 2026) takes a further step by replacing the post-hoc alignment with joint optimization using image-text contrastive and self-supervised objectives. However, similar to the discussion above, by prioritizing generation and compressing representations into low-dimensional latent spaces, these methods weaken the semantic capacity afforded by high-dimensional representations. Taken together, these pioneering works suggest the potential of unifying the representations, yet achieving a fully comprehensive representation without compromising either capability remains an open challenge.

In response to this challenge, we propose GAMUT, which learns a unified visual representation within a joint pretraining framework. GAMUT is trained entirely from scratch by jointly optimizing image-text alignment, self-supervised learning, and pixel reconstruction objectives. Specifically, we develop a three-stage training pipeline that progressively shapes the representation while improving training efficiency and stability. The first stage employs masked image modeling for efficient warm-up, avoiding the prohibitive computational cost of prolonged multi-objective training. The second stage incorporates the three objectives and performs joint pretraining to equip the model with comprehensive capabilities. Latent channel regularization is also introduced to mitigate representation collapse, which manifests as an abnormal energy distribution across channels. The third stage incorporates latent interpolation decoding and other techniques to further enhance reconstruction fidelity and generation robustness. We also employ several simple yet effective architectural modifications, including non-affine normalization (Ba et al., 2016) and query-key normalization (Henry et al., 2020), which improves multi-objective compatibility and training stability. Together, these techniques enable effective and efficient training of GAMUT, leading to a unified representation supporting visual perception, multimodal alignment, and generation-friendly tokenization. As shown in Fig. 1, GAMUT achieves strong performance across a broad range of tasks, whereas existing methods either support only a subset of these capabilities or remain suboptimal in overall performance.

In summary, our key contributions are: (1) We propose GAMUT and develop a unified visual pretraining framework, demonstrating that visual perception, multimodal alignment, and generation-friendly tokenization can be well supported by a single unified visual representation. (2) We identify and address serveral key challenges in unified visual pretraining, including training instability, representation collapse, and the gap between reconstruction and generative decoding, enabling effective and efficient representation learning. (3) Experimental results demonstrate that GAMUT achieves strong performance across diverse visual tasks, outperforming both existing unified tokenizers and the specialized baselines. We hope that our practical exploration will inspire further research and help pave the way toward more comprehensive and capable unified visual representations.

2 Related Work

Self-Supervised Representation. Early self-supervised learning evolved from handcrafted pretext tasks (Noroozi & Favaro, 2016; Gidaris et al., 2018; Zhang et al., 2016) to preserving representation consistency across multiple augmented views (He et al., 2020; Chen et al., 2020; Caron et al., 2020; Grill et al., 2020). More recently, self-distillation and masked image modeling have proved effective at scale and become the dominant paradigms. The representative self-distillation method, DINO (Caron et al., 2021), trains the vision encoder to match the output logits of a momentum teacher. In parallel, MAE (He et al., 2022) introduces masked image modeling, learning representations by predicting the pixels of masked patches from their visible context. iBOT (Zhou et al., 2022) and DINOv2 (Oquab et al., 2024) further combine these two paradigms by introducing patch-level masked logits prediction while retaining image-level self-distillation. Together, these methods learn representations that transfer effectively to both visual recognition and dense perception tasks.

Multimodal-Aligned Representation. CLIP (Radford et al., 2021) popularize large-scale multimodal contrastive pretraining. It maps paired images and texts into a shared embedding space and employs a softmax-based contrastive loss on the image-text embedding similarity logits. Such language supervision endows the visual encoder with strong recognition and multimodal understanding capabilities. SigLIP (Zhai et al., 2023) further replaces the softmax-based objective with a sigmoid-based one, which eliminates global batch-wise normalization, providing a more efficient contrastive objective. More recent studies combine self-supervised learning into the multimodal contrastive framework. TIPS (Maninis et al., 2025) jointly applies image-text contrastive learning and the objective of DINOv2, enhancing spatial awareness and dense visual understanding. SigLIP2 (Tschannen et al., 2025) further integrates captioning into the training framework, producing strong representations and becoming a widely adopted visual encoder for recent multimodal systems.

Visual Tokenization. Visual tokenizers employ the autoencoder architecture, which encodes high-dimensional pixels into compact latent representations, and decodes the latents back to pixels. Discrete tokenizers (Esser et al., 2021; Yu et al., 2024) originates from VQ-VAE (van den Oord et al., 2017), which encode images into sequences of token indices, enabling autoregressive generation through next token prediction (Sun et al., 2024; Chang et al., 2022). Continuous tokenizers follow the VAE (Kingma & Welling, 2014) paradigm, which map images into a low-dimensional, Gaussian-distributed latent space. The generative models (Rombach et al., 2022) then learn diffusion process within this space. Both types of tokenizers are trained with reconstruction objectives and latent-space constraints. More recently, studies begin to incorporate semantic alignment into the generation-oriented training. REPA (Yu et al., 2025) aligns intermediate diffusion features with a pretrained encoder, while VA-VAE (Yao et al., 2025) introduces such alignment into the VAE training process. Both show that semantic alignment effectively improves generation performance.

Unified Representations for Understanding and Generation. Existing efforts toward unified representations broadly follow two routes. The first integrates tokenizers with semantic encoders through parallel fusion (Xie et al., 2025) or sequential enhancement (Huang et al., 2025; Zhang et al., 2026a). Although these methods incorporates characteristics of both representation types, they still retain multiple encoding modules. The second route directly seeks a single universal encoder. RAE (Zheng et al., 2026) and SVG (Shi et al., 2026) employ frozen pretrained semantic encoders to produce latent and learns an additional decoder to reconstruct images. Both demonstrate the generative potential of pretrained semantic spaces, but remain constrained by the fixed representation and exhibit poor preservation of visual details. VTP (Yao et al., 2026), the method most closely related to our GAMUT, jointly optimizes image-text contrastive, self-supervised, and reconstruction objectives. However, it prioritizes generation and compresses high-dimensional features into low-dimensional latents through bottleneck layers, sacrificing semantic richness. Taken together, these methods entail various trade-offs, and a truly versatile representation remains elusive.

3 Method

We propose GAMUT, which learns unified visual representation through a joint pretraining framework. In the following, Sec. 3.1 first introduces some necessary preliminaries and then provides a detailed description of GAMUT's model architecture. Sec. 3.2 and Sec. 3.3 then delve into the details of our proposed three-stage pretraining framework. An overview of GAMUT is shown in Fig. 2.

3.1 ARCHITECTURE OVERVIEW

Preliminaries. As GAMUT adopts some practical designs of SigLIP, DINO, *etc.*, we first briefly review these methods. SigLIP (Zhai et al., 2023) following CLIP (Radford et al., 2021) adopts a dual-encoder architecture composed of separate vision and text encoders. The two encoders process images and text independently and project their outputs into a shared embedding space. On the vision side, multihead attention pooling (MAP) (Zhai et al., 2022) aggregates spatial patch embeddings into a image-level embedding, while the text encoder produces a sentence-level embedding. The image and text embeddings are then aligned using a sigmoid-based binary classification loss.

DINO (Caron et al., 2021) employs an asymmetric student-teacher architecture and introduces self-distillation. The student and teacher share the same architecture, with encoder and projection-head producing high-dimensional logits, while the teacher are updated as an exponential moving average (EMA) of the student. DINO crops smaller local views and larger global views from each image, and aligns the student-encoded both views with the teacher-encoded global views. This cross-view alignment serves as its primary supervision signal. Building on this architecture, iBOT (Zhou et al., 2022) and DINOv2 (Oquab et al., 2024) employ an additional masked image modeling branch. The teacher processes the complete image and serves as an online tokenizer, whereas the student receives a partially masked image and is supervised to align with the teacher logits at the masked positions.

GAMUT Architecture. GAMUT primarily consists of a vision encoder and a pixel decoder, along with other objective-specific components such as the text encoder and EMA teacher. The vision encoder adopts the ViT-B/16 (Dosovitskiy et al., 2021) configuration from the SigLIP family. For the SigLIP and DINO objectives used in the training framework, which require image-level embeddings, we employ separate MAP heads for them. Both heads read the same patch embeddings produced by the vision encoder, but learn their own pooling parameters to reduce conflicts between language alignment and visual self-supervision. The pixel decoder mirrors the encoder architecture, while introducing an output linear layer to project the features back into pixel space. Moreover, we introduce query-key noralization (Dehghani et al., 2023) in the pixel decoder. During training, we observe that the query and key activations in the first few decoder blocks can grow rapidly, accompanied by spikes in the loss and gradient norm, and even prevents the model from convergence. Normalizing them before attention controls this numerical issue and stabilizes training. Since the vision encoder remains numerically stable, we apply this modification only to the pixel decoder.

Furthermore, similar to the approach in RAE, we replace the standard layer normalization (Ba et al., 2016) following the encoder's last Transformer block (Vaswani et al., 2017) with a non-affine one, removing its learnable scale and bias parameters. By constraining patch embeddings to lie on a fixed-energy hypersphere, non-affine layer normalization stabilizes the geometry of the representation space. Meanwhile, it leaves the embedding directions unconstrained, allowing encoding diverse visual characteristics and semantic information through directional variations. Moreover, hypersphere and high-dimensional standard Gaussian exhibit closely related geometries. Samples from a high-dimensional Gaussian tend to concentrate within a thin shell near the surface of a hypersphere, whereas the normalized embeddings lie exactly on that hypersphere. This correspondence makes the non-affine design particularly well suited to producing latents for generative modeling.

3.2 END-TO-END JOINT PRETRAINING

GAMUT builds a three-stage training framework. Stage1 serves as an efficient warmup for the random initialized vision encoder through masked image modeling, in order to establish basic visual modeling capabilities. Stage2 introduces the end-to-end joint pretraining with complementary objectives to learn a unified representation. Stage3 implements serveral augmented decoding training strategies to further improve the perceptual quality of decoding both real and generated latents.

MIM Warmup. Recent state-of-the-art vision-language encoders typically require substantial training budgets. For example, SigLIP2 is trained for 40B seen samples, while PE (Bolya et al., 2025) increases this number to 58B–86B. To accelerate convergence and avoid the cost of learning basic visual structures from scratch, we first pretrain the randomly initialized GAMUT encoder through masked image modeling (MIM), following EVA (Fang et al., 2023). Specifically, we randomly mask the input image patches and train the encoder to predict the corresponding SigLIP2 features at the masked positions. Although MIM requires no annotations, it benefits from images with rich spatial context, which enhance spatial correspondence learning. We therefore curate 27.9M visually rich

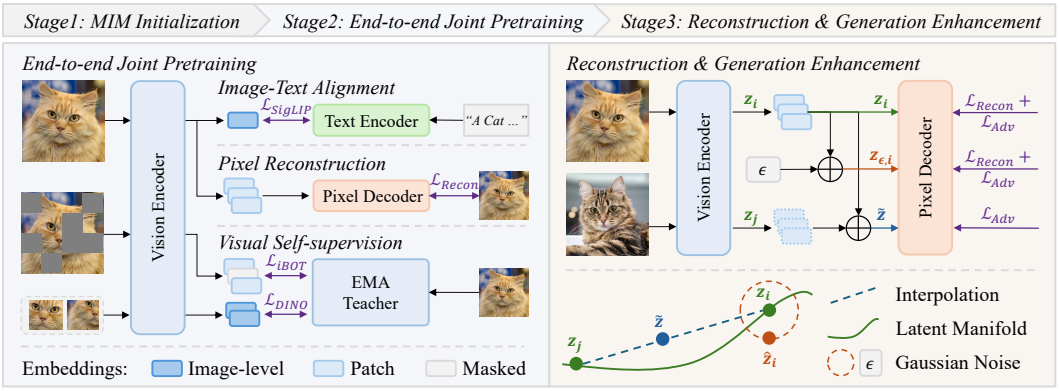


Figure 2: **Overview of the GAMUT training pipeline.** Stage 1 employs MIM for vision encoder initialization. Stage 2 jointly optimizes a shared visual representation through complementary objectives. Stage 3 proposes a three-branches decoding strategy to enhance the pixel decoder.

images from ImageNet-21K (Deng et al., 2009), SA-1B (Kirillov et al., 2023), Objects365 (Shao et al., 2019), and OpenImages v7 (Kuznetsova et al., 2020). This stage converges rapidly within training for 2B seen samples and provides a strong initialization for the subsequent training stages.

End-to-end Joint Pretraining. In this stage, we jointly optimize the visual representation with image-text alignment, visual self-supervision, and pixel reconstruction objectives, allowing it to benefit from the distinct modeling emphases of each objective. For image-text alignment, we employ the sigmoid-based objective from SigLIP and introduce the text encoder, which also adopts the Base-size configuration matching the vision encoder. We curate the required image-text training pairs from several public sources, including LAION-5B (Schuhmann et al., 2022), DataComp-1B (Gadre et al., 2023), and CC12M (Changpinyo et al., 2021). After global deduplication across sources and filtering, we retain 2.4B high-quality image-text pairs as the training data. The image-text alignment objective, along with the information-rich training data, enable our GAMUT to capture the language-grounded visual semantics and acquire basic multimodal world knowledge.

For visual self-supervision, following DINOv2 and iBOT, we combine image-level self-distillation with patch-level masked prediction, and introduce the EMA teacher encoder. These objectives effectively shape the geometry of patch embeddings, complementing the non-affine layer normalization to produce representations with fixed energy and semantically meaningful direction variations, as illustrated in Fig. 6. In contrast to the standard DINOv2 configuration, which performs both local-to-global and cross-global alignment, we remove the cross-global alignment objective and simply align the local-view student logits with a single global-view teacher logits. Consequently, self-distillation introduces only one costly full-resolution forward pass instead of four, substantially reducing the computational overhead. Moreover, this efficient strategy is particularly well suited to our training framework. The preceding training stage and the currently parallel text supervision together equips the encoder with visual modeling capabilities, enabling the self-supervision to focus exclusively on the more challenging local-to-global alignment. In contrast, we found when this configuration is applied to a randomly initialized model, it struggles to capture the necessary visual correspondences, collapse to produce uniform distribution, and remains unchanged thereafter, as shown in Fig. 4.

To preserve fine-grained visual information in the representation, we introduce the pixel decoder along with pixel reconstruction objective. The pixel decoder mirrors the vision encoder architecture while with half as many layers as the encoder to maintain training efficiency. Following the standard practice (Esser et al., 2021; Rombach et al., 2022), we apply the ℓ_1 loss and LPIPS perceptual loss (Zhang et al., 2018) between the pixel decoder output and original images. Since these objectives can already encourage the representation to retain fine-grained visual details that can be reliably recovered, we defer the adversarial training (Goodfellow et al., 2014) to the subsequent stage. This separation avoids the computational overhead and optimization complexity of alternating updates introduced by the adversarial during the lengthy pretraining stage.

Latent Channel Regularization. Although non-affine layer normalization fixes the total energy of each patch embedding, it does not constrain how that energy is distributed across channels. In our preliminary experiments, we observe an pronounced distribution imbalance: a single channel ex-

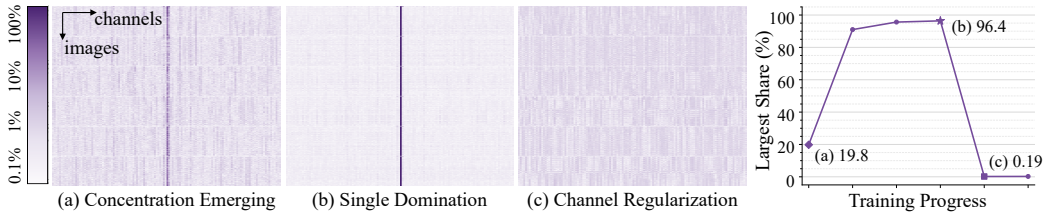


Figure 3: **Visualization of the emergence and mitigation of representation collapse.** The heatmaps show channel-wise energy distributions of representations across images at successive checkpoints during training (a-c). The right plot tracks the energy share of the largest channel.

hibits consistently large magnitudes across different images and spatial positions, eventually dominating the normalized embeddings, as shown in Fig. 3. As a result, patch embeddings with distinct visual characteristics and semantics become increasingly aligned in direction, distorting the semantically meaningful geometry of the representation. We observe that the pixel reconstruction objective benefits from this imbalance (with more details provided in Appendix Sec. A.3 and Fig. 5), suggesting that it may exploit the dominant channel as a shortcut for pixel recovery. In principle, the image-text alignment and self-supervision objectives should oppose this tendency to preserve directional variations, which benefit their objectives. In practice, however, they merely downweight the abnormal channel in the projection layers, ignoring the imbalance rather than correcting it.

Given that the existing objectives do not inherently suppress energy concentration, we propose latent channel regularization to directly constrain the channel distribution. Specifically, drawing inspiration from VICReg (Bardes et al., 2022), we regularize the channel-wise statistics of the representation. Let $\mathbf{z} \in \mathbb{R}^{B\times N\times D}$ denote the normalized patch embeddings, where B, N, and D are the batch size, number of patch embeddings, and channel dimension, respectively. For each channel, we first compute its mean μ_d and second moment $m_{2,d}$, and then define the regularization loss as:

$$\mu_d = \frac{1}{BN}\sum_{b=1}^{B}\sum_{n=1}^{N} z_{b,n,d},\ m_{2,d} = \frac{1}{BN}\sum_{b=1}^{B}\sum_{n=1}^{N} z_{b,n,d}^2,\ \mathcal{L}_{\text{LCR}} = \frac{1}{D}\sum_{d=1}^{D}(\mu_d^2 + (m_{2,d}-1)^2). \quad (1)$$

The μ_d term prevents each channel from biasing to a common direction, while the $m_{2,d}$ term promotes a balanced energy distribution across channels. In practice, we compute the statistics over each local batch, rather than separately for each image or across the global batch. Per-image statistics impose an overly rigid constraint that ignores the natural spatial correlations within an image, while global-batch statistics provide no noticeable improvement but incur substantial communication overhead. Beyond preventing representation collapse, this regularization also facilitates downstream generative modeling by encouraging the representation distribution to approach a standard Gaussian, yielding an effect similar to that of KL-regularization used in the VAE training process.

3.3 Reconstruction and Generation Enhancement

The first two training stages focus on optimizing the representations themselves. However, to serve as latents for generative models, the ability to decode latents into high-quality images is also important. We therefore introduce this additional stage to enhance the decoding process. Specifically, we freeze the encoder and scale up the pixel decoder to ViT-XL, following the configuration of RAE, to improve the quality and robustness of pixel decoding. We propose a three-branch-decoding training strategy, comprising direct reconstruction, noisy reconstruction, and latent interpolation decoding.

Direct and Noisy Reconstruction. The direct reconstruction branch follows the standard tokenizer training setup: the decoder reconstructs images from the encoder-produced latent $\mathbf{z}$ under the ℓ_1 loss, LPIPS perceptual loss, and adversarial loss. This branch primarily improves the decoder's ability to reconstruct high-fidelity images through latents on the real-data latent manifold. In downstream generation, generative models may produce imperfect latents that deviate from the real-data manifold. Following recent works (Zheng et al., 2026; Yang et al., 2026), we therefore introduce a noisy reconstruction branch to improve the decoder's robustness to such deviations. Specifically, we perturb the latent with Gaussian noise, $\mathbf{z}_\epsilon = \mathbf{z} + \sigma\boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where σ controls the noise magnitude and is randomly sampled from $\mathcal{U}(0, 0.6)$ during training. We train the decoder to recover the original image through the perturbed latents using the same objectives as in the first branch.

Latent Interpolation Decoding. Although noisy reconstruction improves the robustness of decoding off-manifold latents, we find that Gaussian noise injection primarily simulates local, statistically isotropic deviations from the real manifold. However, the imperfect generated latents often combine multiple plausible yet competing visual patterns (Xu et al., 2026), resulting in structured deviations that produce hallucinations and visual artifacts. Since such latents are unavailable until the generative model being fully trained, we introduce latent interpolation decoding, as shown in Fig. 2. This branch performs interpolation between nearby representations to simulate the mixtures of visual patterns, providing an efficient proxy for structured errors in generation. Specifically, for each encoded latent $\mathbf{z}_i$, we find its nearest neighbor $\mathbf{z}_j$ in the batch according to latent similarity, and construct:

$$\tilde{\mathbf{z}}_i = (1-\alpha)\mathbf{z}_i + \alpha\mathbf{z}_j, \quad j = \arg\max \mathrm{sim}(\mathbf{z}_i, \mathbf{z}_{k|k\neq i}), \tag{2}$$

where α is the interpolation coefficient and is drawn from $\mathcal{U}(0,1)$ during training. Since an interpolated latent has no corresponding reconstruction target, we apply only the adversarial loss for this branch, omitting the two reconstruction losses. This encourages the decoder to produce realistic images from latents between neighboring representations, thereby improving its robustness to the structured errors of generation. Moreover, this observation aligns with a recent finding iFID (Xu et al., 2026), which show that the FID (Heusel et al., 2017) of decoded images produced by latent interpolation correlates with the gFID of downstream generative models. We go a step further by using this interpolation as a proxy task to directly improve the robustness of the visual tokenizer.

4 EXPERIMENTS

We provide the full implementation details in Appendix A. Below, we report the evaluation results of GAMUT along three complementary axes. First, zero-shot image classification and bidirectional image-text retrieval assess whether it captures language-grounded visual semantics. Second, image reconstruction and generation evaluate its effectiveness as a generative tokenizer in terms of reconstruction fidelity and generation quality. Third, dense and open-vocabulary perception tasks probe the spatial structure encoding capability of the representation. Together, these evaluations cover the core understanding and generation capabilities expected of a unified visual representation.

Image-text Alignment. We report top-1 accuracy of zero-shot classification on ImageNet-1K (Deng et al., 2009), other ImageNet variants (Recht et al., 2019; Beyer et al., 2020; Hendrycks et al., 2021b;a; Wang et al., 2019), ObjectNet (Barbu et al., 2019), and several fine-grained classification datasets (Xiao et al., 2010; Cimpoi et al., 2014; Nilsback & Zisserman, 2008; Parkhi et al., 2012). For zero-shot image-text retrieval, we report Recall@1 on the widely used COCO (Lin et al., 2014) and Flickr30K (Young et al., 2014) datasets. Comparison methods include the specialized baselines of vision-language encoder, such as CLIP, OpenCLIP (Cherti et al., 2023), MetaCLIP (Xu et al., 2024), DataComp (Gadre et al., 2023), DFN (Fang et al., 2024a), SigLIP, and the current state-of-the-art model SigLIP2, as well as unified tokenizers that support image-text alignment and are comparable in model scale to our GAMUT, including VTP and UniTok (Ma et al., 2025).

As shown in Tab. 1, GAMUT obtains an overall average score of 71.3, substantially outperforming existing unified tokenizers by 4.4 points over VTP and 7.5 points over UniTok. The gains are particularly pronounced on retrieval tasks, demonstrating that GAMUT not only learns broad text-aligned recognition knowledge, but also effectively captures natural-language descriptions. Compared with other vision-language encoders, GAMUT is the first unified method that outperforms all specialized baselines. Although it remains 2.2 points behind the current SOTA, SigLIP2, it is acceptable given the substantial training scale gap: 2.4B *v.s.* 10B dataset samples and 14B *v.s.* 40B seen samples.

Image Reconstruction and Generation. We evaluate reconstruction on the ImageNet-1K validation set at 256×256 resolution using rFID, LPIPS (Zhang et al., 2018), PSNR, and SSIM (Wang et al., 2004). We compare with the baseline methods VQGAN, SD-VAE, the semantic augmented method VA-VAE, and unified tokenizers SVGTok (Shi et al., 2026), RAE, and VTP. To evaluate GAMUT's potential as a tokenizer for generative models, we train class-conditional generation in the representation space of GAMUT at 256×256 resolution on ImageNet. LightningDiT (Yao et al., 2025) is adopted as generative backbone since its standard architecture incorporates only essential components without specific modifications, thereby enabling a clean assessment of GAMUT's general capabilities. This backbone is also employed in prior unified tokenizers, facilitating a controlled comparison that effectively isolates the contribution of representation. The comparison generative backbones also include LlamaGen (Sun et al., 2024), REPA, and SVG. We evaluate generation

Table 1: Image-text alignment: zero-shot classification and bidirectional image-text retrieval results.

Method	Classification (Top-1 Acc)													Retrieval ($R_{I,T}$@1)				
	IN_{val}	IN_{v2}	IN_{real}	Obj.	IN_A	IN_R	IN_O	IN_S	SUN	DTD	Flo.	Pets	**Avg**	COCO		Flickr		**Avg**
Specialist Baselines																		
CLIP	68.3	61.9	75.1	55.3	50.0	77.7	42.3	48.2	64.3	45.0	69.1	88.9	62.2	32.9	52.6	63.0	83.0	57.9
OpenCLIP	70.2	62.2	77.2	56.0	38.1	80.6	46.1	56.2	70.9	56.3	71.3	90.6	64.6	42.4	59.2	69.7	86.2	64.4
MetaCLIP	72.1	65.1	79.0	61.3	49.4	84.2	41.5	60.2	68.8	62.2	75.0	91.8	67.6	41.4	59.2	70.4	86.0	64.3
DataComp	73.5	66.0	79.9	63.6	48.3	83.6	42.9	60.5	70.5	57.9	75.8	92.6	67.9	40.2	57.4	67.1	85.0	62.4
DFN	76.2	68.2	83.0	63.4	48.1	83.1	49.3	62.1	71.4	57.9	82.0	93.7	69.9	43.2	60.3	69.1	85.1	64.4
SigLIP	76.6	69.2	83.0	57.1	48.8	90.8	37.9	68.1	70.3	65.8	85.0	94.2	70.6	48.4	66.3	74.9	90.3	70.0
Specialist State-of-the-art																		
SigLIP2	79.3	71.9	85.3	62.4	60.1	91.9	35.9	69.5	73.3	67.1	85.6	95.6	73.2	54.1	70.9	80.3	93.4	74.7
Unified Tokenizers																		
UniTok	71.3	63.6	77.8	65.5	51.2	82.5	37.7	59.1	69.7	52.1	70.7	90.3	66.0	36.2	53.1	61.1	78.5	57.2
VTP	73.3	65.8	80.0	57.2	49.3	85.6	41.5	63.1	70.6	62.0	78.3	91.1	68.2	41.2	58.8	68.2	83.5	62.9
GAMUT	76.4	69.1	83.0	57.1	51.7	87.8	43.0	65.1	72.5	67.3	86.6	94.6	**71.2**	50.2	67.3	77.0	91.6	**71.5**

Table 2: Image reconstruction and class-conditional generation results at 256×256 resolution on ImageNet.

Tokenizer	Reconstruction				Generation		
	rFID ↓	LPIPS ↓	PSNR ↑	SSIM ↑	Model	Epoch	gFID ↓
Specialist Baselines							
VQGAN	7.96	0.306	20.20	0.52	L.Gen	300	9.38
SD-VAE	0.61	0.130	26.90	0.74	REPA	800	5.78
Unified Tokenizers							
SVGTok	0.65	0.190	23.89	0.65	SVG	80	6.57
VA-VAE	0.28	**0.096**	**27.96**	0.79	L.DiT	80	4.29
RAE	0.57	0.256	18.86	0.48	L.DiT	80	4.28
VTP	0.73	0.156	25.26	0.72	L.DiT	80	3.88
GAMUT	**0.22**	0.099	27.43	**0.80**	L.DiT	80	**3.81**

Table 3: Semantic and open-vocabulary segmentation results.

mIoU ↑	SigLIP	VTP	**GAMUT**
Visual Perception			
VOC	82.43	79.69	**84.25**
ADE	41.54	38.15	**43.98**
Open Vocabulary			
A-150	13.82	8.68	**14.53**
A-847	19.03	13.41	**20.24**
PC-59	33.01	28.61	**34.04**
PC-459	51.89	42.94	**51.96**
VOC-20	85.55	76.32	**86.10**
VOC-21	71.14	68.56	**71.58**

using the gFID without classifier-free guidance (Ho & Salimans, 2022) under the class-balanced ImageNet-1k protocol used in (Zheng et al., 2026).

The results are shown in Tab. 2, in which "L.Gen" and "L.DiT" refer to LlamaGen and LightningDiT. For image reconstruction, GAMUT achieves the best rFID of 0.22 and SSIM of 0.80 among all compared methods. Its LPIPS and PSNR are within 0.003 and 0.5 of the best-performing VA-VAE, respectively, while outperforming all other methods. This result demonstrates that GAMUT effectively reconciles the preservation of fine-grained visual details with the strong language-grounded semantics observed above. For image generation, the most direct comparison is among the unified tokenizers employing the same LightningDiT backbone for 80 epochs training. Under this controlled setting, GAMUT obtains the best gFID of 3.81, outperforming VTP, RAE, and VA-VAE. It also outperforms all other methods, demonstrating the representation is generation-friendly and can be effectively modeled by a standard diffusion architecture.

Visual Perception. We evaluate two dense visual perception tasks: semantic segmentation and open-vocabulary segmentation. For semantic segmentation, we freeze each encoder and train a DPT (Ranftl et al., 2021) head for mask prediction on PASCAL VOC 2012 (Everingham et al., 2015) and ADE20K (Zhou et al., 2017). This controlled setup isolates the ability of the original representations to encode semantic and spatial structure. For open-vocabulary segmentation, we adopt the CAT-Seg (Cho et al., 2024) framework, replacing its vision and text encoders with those under evaluation while keeping all other components unchanged. Following its protocol, we train on COCO-Stuff (Caesar et al., 2018) and evaluate on ADE20K, PASCAL Context (Mottaghi et al., 2014), and PASCAL VOC (Everingham et al., 2010). This task assesses both spatial modeling and the understanding of language-grounded visual semantics. All models are trained at 256×256 resolution and evaluated on full images using sliding-window inference with the same-size crops. As the results shown in Tab. 3, GAMUT consistently outperforms both SigLIP and VTP across all eight benchmarks. Compared with SigLIP, results show that our joint optimization framework learns more

comprehensive representation than image-text contrastive learning alone. The improvements over VTP further support our motivation: preserving high-dimensional representations retains a broader range of capabilities and avoids the information loss introduced by projection and compression.

Ablation Studies. We conduct two groups of ablations to evaluate how the different objectives and components of our training framework affect visual understanding and generation, respectively, as shown in Tab. 4. For understanding, we progressively construct four configurations with each introducing one additional components while fixing the training budget at 4B seen samples. Results demonstrate that, starting from only employing the SigLIP objective[(1)], introducing MIM warmup[(2)] yields a substantial 5.22-point improvement in early-stage performance, confirming its effect on accelerating convergence. Adding self-supervision[(3)] yields a further 1.16-point improvement, demonstrating its effectiveness on enhancing visual modeling. Finally, incorporating pixel reconstruction[(4)] leaves semantic recognition essentially unchanged, showing that preserving fine-grained visual details does not compromise the semantic structure. As a result, the three objectives can be well jointly optimized in a single representation space.

Table 4: Ablations on semantic modeling and generation.

Setup	IN-1K Acc. ↑
(1). SigLIP Only	68.14
(2). MIM + SigLIP	73.36
(3). (2) + SS	74.52
(4). (3) + Recon	74.51
Setup	**gFID ↓**
(a). SigLIP + Recon	11.51
(b). (a) + SS	8.30
(c). w/o LCR	383.28
(d). w/ LCR	8.92
(e). w/o LID	7.35
(f). w/ LID	5.47

For generation, we train the generative backbone for 80 epochs on representation of different configurations and evaluate gFID. We first isolate the effect of self-supervision[(b)] on top of the baseline with only SigLIP and pixel reconstruction[(a)]. The clear improvement shows that the semantically meaningful geometry induced by self-supervision (as shown in Appendix Fig. 6) benefits generation. This is intuitive that a latent space whose distances reflect visual and semantic similarity should be easier to learn than an unstructured representation space. We then evaluate the effectiveness of latent channel regularization (LCR)[(c), (d)] and latent interpolation decoding (LID)[(e), (f)] under two groups of independent controlled comparison. Without LCR, the single-channel energy dominance severely distorts the representation distribution, and the downstream generator can barely converge to produce valid images. Restoring LCR resolves this failure mode, establishing it as a necessary component for a generation-compatible representation. LID is also demonstrated effective to provide an additional gain by improving the robustness for decoding the generated latents.

5 Conclusion and Discussion

We presented GAMUT, a visual pretraining framework that learns a unified representation supporting visual perception, multimodal alignment, and generation-friendly tokenization. GAMUT achieves strong performance across all these tasks and demonstrates that unified visual pretraining is feasible, but also that it requires more than a simple superposition of multiple objectives. To achieve unified this, we design the three-stage training strategy to improve efficiency and reduce optimization complexity, introduce latent channel regularization to prevent representation collapse, and propose latent interpolation decoding to enhance the perceptual quality of generation. We also introduce several simple yet effective architectural modifications to better accommodate the unified framework and stabilize training, including non-affine normalization in the encoder and query–key normalization in the decoder. Our experiments further clarifies the role of each training objective. Image-text alignment supplies recognition knowledge and language-grounded visual semantics, forming the foundation for visual understanding; pixel reconstruction preserve the necessary visual details for generation; and self-supervision benefits both sides by strengthening visual spatial structures modeling, as well as shaping a generation-friendly geometry for the representation space.

Limitations and Future Work. Our current study remains limited in scale. As an exploratory work, our primary goal is to establish the feasibility of unified representation learning and validate the complete training pipeline at a manageable computational cost. We therefore instantiate GAMUT at the ViT-Base scale and evaluate reconstruction and generation only employing the ImageNet protocol. The model scale is substantially smaller than the encoders used in modern multimodal systems, which commonly contain at least several hundred million parameters. Understanding how GAMUT scales and assessing its practical performance on more complex tasks, such as vision–language dialogue and text-to-image generation, remain important directions for our future work.

AI USE STATEMENT

We used AI to assist with language editing of the manuscript, including spelling, grammar, punctuation, and spacing corrections. The language edits were intended to preserve the technical content and experimental results. The authors are responsible for reviewing all AI-assisted edits and take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed information on the data, training procedures, and evaluation protocols used in this work. The training data sources and dataset construction are described in Sec. 3.2 and Appendix A.2. Model architectures, training objectives, hyperparameters, and implementation details for all three training stages are provided in Appendix A.1–A.4. Appendix A.5 further specifies the evaluation protocols, including preprocessing, metrics, and evaluation tools, to facilitate reproduction of the reported results.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 60884–60937, 2025. doi: 10.52202/085713-2036.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1209–1218, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11315–11325, 2022.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning (ICML)*, pp. 1597–1607, 2020.

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.

Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. CAT-Seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4113–4123, 2024.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3606–3613, 2014.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme Ruiz, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd Van Steenkiste, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Collier, Alexey A. Gritsenko, Vighnesh Birodkar, Cristina Nader Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetic, Dustin Tran, Thomas Kipf, Mario Lucic, Xiaohua Zhai, Daniel Keysers, Jeremiah J. Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7480–7512, 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)*, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12873–12883, 2021.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. doi: 10.1007/s11263-014-0733-5.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. In *International Conference on Learning Representations (ICLR)*, 2024a.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: Exploring the limits of masked visual representation learning at scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19358–19369, 2023.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024b.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. DataComp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. doi: 10.52202/075280-1179.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020.

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)*, pp. 16000–16009, 2022.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15262–15271, 2021b.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 4246–4253, 2020. doi: 10.18653/v1/2020.findings-emnlp.379.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ziyuan Huang, DanDan Zheng, Cheng Zou, Rui Liu, Xiaolong Wang, Kaixiang Ji, Weilong Chai, Jianxin Sun, Libin Wang, Yongjie Lv, Taozhi Huang, Jiajia Liu, Qingpei Guo, Ming Yang, Jingdong Chen, and Jun Zhou. Ming-univision: Joint image understanding and generation with a unified continuous tokenizer. *arXiv preprint arXiv:2510.06590*, 2025.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The Open Images dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. doi: 10.1007/s11263-020-01316-z.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

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 (ECCV)*, pp. 740–755, 2014.

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.

Kevis-Kokitsi Maninis, Kaifeng Chen, Soham Ghosh, Arjun Karpur, Koert Chen, Ye Xia, Bingyi Cao, Daniel Salz, Guangxing Han, Jan Dlabal, et al. Tips: Text-image pretraining with spatial awareness. In *International Conference on Learning Representations*, volume 2025, pp. 64256–64279, 2025.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, pp. 722–729, 2008.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision (ECCV)*, pp. 69–84, 2016.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024.

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

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 (ICML)*, pp. 8748–8763, 2021.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12179–12188, October 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning (ICML)*, pp. 5389–5400, 2019.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. doi: 10.52202/068431-1833.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8430–8439, 2019.

Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder. In *International Conference on Learning Representations (ICLR)*, 2026.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

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

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3485–3492, 2010.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 47490–47518, 2025. doi: 10.52202/085713-1584.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems*, volume 37, pp. 52040–52094, 2024. doi: 10.52202/079017-1650.

Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *International Conference on Learning Representations (ICLR)*, 2024.

Tongda Xu, Mingwei He, Shady Abu-Hussein, Jose Miguel Hernandez-Lobato, Chunhang Zheng, Kai Zhao, Chao Zhou, Ya-Qin Zhang, and Yan Wang. Making reconstruction FID predictive of diffusion generation FID. *arXiv preprint arXiv:2603.05630*, 2026.

Jiawei Yang, Tianhong Li, Lijie Fan, Yonglong Tian, and Yue Wang. Latent denoising makes good tokenizers. In *International Conference on Learning Representations (ICLR)*, 2026.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Jingfeng Yao, Yuda Song, Yucong Zhou, and Xinggang Wang. Towards scalable pre-training of visual tokenizers for generation. In *European Conference on Computer Vision (ECCV)*, 2026.

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

Lijun Yu, José Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David Ross, and Lu Jiang. Language model beats diffusion – tokenizer is key to visual generation. In *International Conference on Learning Representations (ICLR)*, 2024.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations (ICLR)*, 2025.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12104–12113, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *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, Zhiding Yu, and Cihang Xie. Openvision 3: A family of unified visual encoder for both understanding and generation. *arXiv preprint arXiv:2601.15369*, 2026a.

Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *European Conference on Computer Vision (ECCV)*, pp. 649–666, 2016.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

Shilong Zhang, He Zhang, Zhifei Zhang, Chongjian Ge, Shuchen Xue, Shaoteng Liu, Mengwei Ren, Soo Ye Kim, Yuqian Zhou, Qing Liu, Daniil Pakhomov, Kai Zhang, Zhe Lin, and Ping Luo. Both semantics and reconstruction matter: Making representation encoders ready for text-to-image generation and editing. In *International Conference on Machine Learning (ICML)*, 2026b.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations (ICLR)*, 2026.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 633–641, 2017.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan L. Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

A Implementation Details

A.1 Architecture and Common Setup

GAMUT uses a ViT-B/16 vision encoder that processes images at 256×256 resolution. The encoder comprises 12 Transformer blocks with a width of 768 and 12 attention heads, producing a 16×16 grid of 768-dimensional patch embeddings. We omit the class token and employ the MAP head to acquire the image-level embedding. The patch embeddings after the final non-affine layer normalization serve as the unified visual representation. The encoder employs learnable absolute positional embeddings, which are bicubically interpolated when processing input at different resolutions, *e.g.*, the local crops used in self-supervision. Each of the two objective-specific MAP heads contains one learned query and 8 attention heads while preserving the 768-dimensional output dimension. The text encoder also contains 12 Transformer blocks with a width of 768 and 12 attention heads. We set the context length to 64 and use the tokenizer from SigLIP2 (Tschannen et al., 2025), which has 256K vocabulary. The QK-Norm is applied with learnable affine parameters to every attention layer in the pixel decoders. All three stages use `bfloat16` mixed-precision training.

A.2 Stage 1: MIM Warmup

The first stage equips the randomly initialized encoder with basic visual modeling capabilities. Following the masked feature prediction recipe of EVA-02 (Fang et al., 2024b), we apply random resized cropping with a scale range of $[0.2, 1.0]$ and resize each crop to 256×256. We employ `siglip2-giant-opt-patch16-256` as the teacher and mask 75% of the student patch tokens using block-wise masking. A linear prediction head maps the student outputs to the teacher feature space, and we apply a cosine similarity loss only at the masked positions.

We train with a global batch size of 32,768. After 1,000 warmup steps, the learning rate is cosine-decayed from 5×10^{-3} to 1×10^{-5}, and the gradient norm is clipped at 3.0. The training data consists of ImageNet-21K (Deng et al., 2009), SA-1B (Kirillov et al., 2023), Objects365 (Shao et al., 2019), and OpenImages v7 (Kuznetsova et al., 2020), which are all open-source visual perception datasets, forming a total of 27.9M visually rich images collection.

A.3 Stage 2: End-to-end Joint Pretraining

Stage 2 jointly optimizes the shared representation through image-text alignment, self-supervision, pixel reconstruction, along with the latent channel regularization. The complete objective is

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{SigLIP}} + \underbrace{\mathcal{L}_{\text{DINO}} + \mathcal{L}_{\text{iBOT}}}_{\text{self-supervision}} + \underbrace{\mathcal{L}_{\ell_1} + \mathcal{L}_{\text{LPIPS}}}_{\text{pixel reconstruction}} + \mathcal{L}_{\text{LCR}}. \tag{3}$$

All six terms in Eq. (3) receive unit coefficients and are summed before backpropagation. This preserves the balanced weighting of the alignment and self-supervised objectives used in VTP (Yao et al., 2026), while increasing the reconstruction weight from 0.1 to 1.0, as we find that the larger reconstruction weight does not conflict with the other objectives. After 2,000 warmup steps, the learning rate is cosine-decayed from 5×10^{-4} to 1×10^{-6}, and the gradient norm is clipped at 1.0.

Image-text Alignment. Both image and text embeddings are 768-dimensional. We initialize the learnable logit scale and bias of the SigLIP objective to $\ln(10)$ and -10, respectively, and cap the logit scale at $\ln(100)$ during training. Negatives are aggregated across all devices through the all-gather operation. The specialized distributed communication design reported by SigLIP provides no measurable efficiency benefit in our training environment, thus we use the simpler all-gather implementation. We train with a global batch size of 32,768. Image augmentation is limited to random resized cropping with scale $[0.08, 1.0]$ and aspect ratio $[0.75, 4/3]$; we avoid stronger transformations that could alter image-text correspondence and introduce bias.

Self-supervision. The self-supervised branch operates on the same image batch used for text-alignment. Following VTP, we randomly sample 4,096 images from the 32,768 global batch for this branch. This subsampling preserves data diversity while controlling the additional computation cost of self-supervision. Moreover, 4,096 is already a sufficient global batch size for the self-supervised objective. The remaining hyperparameters largely follow DINOv2 (Oquab et al., 2024) and TIPS (Maninis et al., 2025). The student and teacher use identically structured projection heads,

each implemented as a three-layer MLP with a hidden dimension of 2048, a bottleneck dimension of 256, and an output logits dimension of 65,536.

For each selected image, we sample one 256×256 global crop with scale $[0.4, 1.0]$ and six 96×96 local crops with scale $[0.05, 0.4]$. The image-level DINO objective aligns the student's predictions for the local crops with the teacher's prediction for the global crop. For the patch-level iBOT objective, we independently mask 50% of the patches in the global view. We set the student temperature to 0.1 and compute the teacher assignments using three iterations of Sinkhorn–Knopp normalization.

As discussed in Sec. 3.2, our modification to the DINO training data flow is a specifically design compatible with our GAMUT and is made possible by its multi-stage training strategy and the joint supervision. Here, we provide further evidence to support this. Fig. 4 shows the DINO loss curves under two training settings: our second-stage training and a controlled baseline without MIM warmup or the SigLIP objective. Loss of the second one remains nearly constant at $\ln(65{,}536) \approx 11.09$, the cross-entropy of a uniform prediction over the 65,536-dimension output logits. This plateau indicates that the model collapses to uniform outputs and does not escape this degenerate solution. In contrast, with MIM warmup and joint optimization, the DINO loss decreases steadily and converges as expected. This observation suggests a practical lesson: when training DINO from scratch, cross-global-view supervision is needed as an auxiliary signal, as local-to-global alignment alone is too challenging for a randomly initialized model.

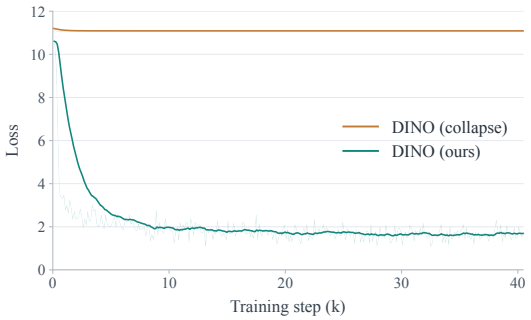


Figure 4: **DINO loss curves under different training setups.** Our full training setup converge-sas expected. Without the Stage 1 warmup and joint optimization, the loss remains nearly constant, indicating uniform-output collapse.

Pixel Reconstruction. The pixel decoder in Stage 2 follows the width and attention configuration of the vision encoder but uses only 6 Transformer blocks, half the encoder depth. Since the decoder receives a sequence of visual representations rather than an image, we replace the patch-embedding layer with a linear layer that independently projects each encoder patch embedding. The output head predicts the pixels within each patch, which are then rearranged to form the reconstructed image.

The reconstruction branch reuses the representation computed for image-text alignment and targets the corresponding encoder input after normalization to $[-1, 1]$. As in the self-supervised branch, we subsample the full image-text batch. A global batch size of 1,024 is sufficient for reconstruction and substantially reduces decoder computation. Reusing the same representation ensures that alignment and reconstruction operate on exactly the same visual distribution rather than on task-specific inputs.

Latent Channel Regularization. We introduce latent channel regularization after observing the emergence of single-channel energy concentration during joint training. Instead of restarting pre-training, we activate the regularizer from the existing checkpoint. As shown in Fig. 5 (a), latent channel regularization effectively reverses the concentration trend and restores a balanced distribution of channel energy, demonstrating that it can recover a representation from an emerging collapse, in addition to preventing collapse when applied from the start of training. Furthermore, as shown in Fig. 5 (b) and (c), we observe that the SigLIP loss remains stable across the introduction of latent channel regularization, whereas the image reconstruction loss exhibits a sharp spike. This contrasting behavior supports our analysis in Sec. 3.2: the reconstruction objective had exploited the dominant channel as a shortcut for pixel recovery, and suppressing this imbalance temporarily disrupts that shortcut before the decoder adapts to the more balanced representation.

A.4 Stage 3: Reconstruction and Generation Enhancement

In the final stage, we freeze the vision encoder and train a larger pixel decoder from random initialization. The decoder employs the ViT-XL configuration, containing 28 Transformer blocks with a width of 1,152, 16 attention heads, and an MLP hidden dimension of 4,096. We train it on the ImageNet-1K training set for 64 epochs with a global batch size of 2,048.

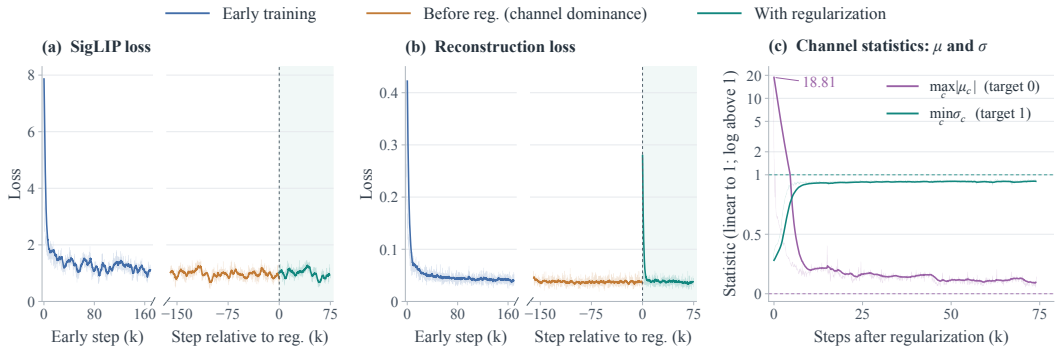


Figure 5: **Effect of activating latent channel regularization during joint pretraining.** Training dynamics before and after employing latent channel regularization is introduced, including the SigLIP and pixel reconstruction losses, along with the channel statistics.

In each training step, we first encode the input images and get their representations. The resulting representations are then transformed into inputs for the direct reconstruction, noisy reconstruction, and interpolation branches, which are decoded in parallel to improve efficiency. For noisy reconstruction, each sample draws a noise level $\sigma \sim \mathcal{U}(0, 0.6)$ that is shared across all spatial tokens. For interpolation decoding, we compute the pairwise similarity at corresponding spatial positions and average it over the token grid to identify the nearest neighbor of each image. The interpolation coefficient is sampled independently as $\alpha \sim \mathcal{U}(0, 1)$.

The direct and noisy branches are supervised by the ℓ_1 and LPIPS losses throughout this stage, with adversarial supervision added after the initial reconstruction-only phase. Because interpolated representations have no pixel-aligned targets, the interpolation branch is trained only with the adversarial loss. Following RAE (Zheng et al., 2026), the discriminator consists of a DINO-S/8 backbone followed by a randomly initialized trainable head. We first train the decoder for 16 epochs using only the ℓ_1 and LPIPS reconstruction losses employed in Stage 2. Discriminator updates begin at epoch 16, and adversarial supervision is added to the decoder at epoch 18. Thereafter, the decoder and discriminator are updated at a 1:1 ratio. We multiply the adaptive adversarial weight by an additional fixed coefficient of 0.75. The discriminator learning rate is cosine-decayed from 5×10^{-5} to 5×10^{-6}, while the decoder learning rate is cosine-decayed from 1×10^{-5} to 1×10^{-6}.

A.5 Evaluation Protocols

We evaluate zero-shot classification and image-text retrieval with the `CLIP-benchmark` toolkit.[1] This toolkit provides a consistent implementation of dataset preprocessing, prompt ensembling, and retrieval metrics across all image-text benchmarks.

We evaluate reconstruction on all images in the ImageNet-1K validation set. Reconstructed and reference images are processed with the same ADM protocol (Dhariwal & Nichol, 2021), and the real-image statistics for rFID are computed from this validation set. This consistent protocol avoids discrepancies caused by mismatched resizing procedures or reference statistics.

For generation, we train LightningDiT-XL (Yao et al., 2025) for 80 epochs. We enlarge its input projection to accommodate the 768-dimensional representation while leaving the main Transformer architecture unchanged. Following JiT (Li & He, 2026), the network directly predicts the clean latent representation and converts this prediction to velocity space for loss computation. We also adopt the dimension-dependent noise schedule shift used by RAE (Zheng et al., 2026) to account for the high dimensionality of the representation. We report gFID without classifier-free guidance under the class-balanced ImageNet-1K sampling protocol. The gFID reference statistics and preprocessing follow the ADM protocol (Dhariwal & Nichol, 2021).

[1] `https://github.com/LAION-AI/CLIP_benchmark`

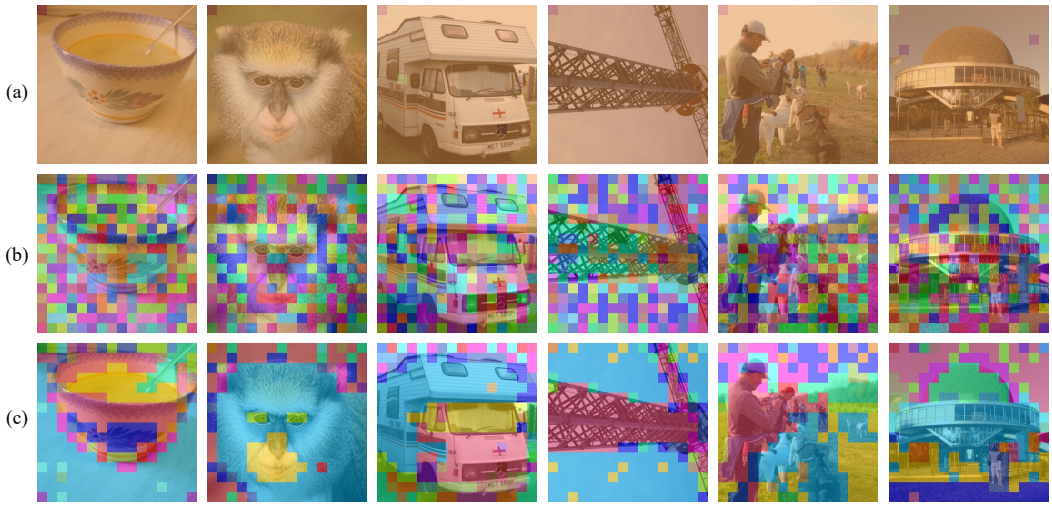


Figure 6: **Representation clustering illustration.** In each triplet, the three images visualize representations from different model checkpoints: the left corresponds to the checkpoint exhibiting single-channel energy dominance, the middle to an ablation trained only with the SigLIP and reconstruction objectives, and the right to the full GAMUT configuration.

We implement the frozen-encoder semantic segmentation experiments with `MMSegmentation`.[2] For open-vocabulary segmentation, we use the official CAT-Seg implementation.[3] We replace only the vision and text encoders with the models under evaluation and otherwise retain the original CAT-Seg training and inference protocol to isolate differences in visual representation.

B QUALITATIVE RESULTS

In Fig. 6, we visualize the clustering structure of representations from three checkpoints of GAMUT: one exhibiting single-channel energy dominance, an ablation trained with only the SigLIP, pixel reconstruction, and latent channel regularization objectives, and the full GAMUT configuration. Single-channel dominance nearly erases the spatial structure of the representation space, causing embeddings from different spatial positions to become almost indistinguishable. In contrast, the latter two checkpoints recover a diverse representation distribution, confirming that latent channel regularization effectively prevents energy concentration. Their comparison further supports our observation that self-supervision strengthens spatial structure encoding and organizes the representation space into a more semantically meaningful geometry. The clustering is performed within each image using agglomerative clustering with cosine distance and average linkage, applying a fixed distance threshold of 0.10 to both checkpoints.

Fig. 7 presents a qualitative comparison of image reconstruction among GAMUT, RAE, and VTP, where GAMUT better preserves and recovers fine-grained visual details. Figs. 8 and 9 further show qualitative results from the generative model trained in GAMUT 's representation space.

[2] `https://github.com/open-mmlab/mmsegmentation`
[3] `https://github.com/cvlab-kaist/CAT-Seg`



(a) Original (b) RAE (c) VTP (d) GAMUT (Ours)

Figure 7: **Qualitative comparisons on image reconstruction.**



Figure 8: **Qualitative results of class-conditional image generation**



Figure 9: **Qualitative results of class-conditional image generation**