The Prism Hypothesis: Harmonizing Semantic and Pixel Representations via Unified Autoencoding

Weichen Fan[1], Haiwen Diao[1], Quan Wang[2]
Hualin Da[2], and Ziwei Liu[1]

[1] S-Lab, Nanyang Technological University
weichen002@e.ntu.edu.sg, {haiwen.diao,ziwei.liu}@ntu.edu.sg
[2] SenseTime Research
{wangquan,dhlin}@sensetime.com

Abstract. Deep representations across modalities are inherently intertwined. In this paper, we systematically analyze the spectral characteristics of various semantic and pixel encoders. Interestingly, our study uncovers a highly inspiring and rarely explored correspondence between an encoder's feature spectrum and its functional role: semantic encoders primarily capture low-frequency components that encode abstract meaning, whereas pixel encoders additionally retain high-frequency information that conveys fine-grained detail. This heuristic finding offers a unifying perspective that ties encoder behavior to its underlying spectral structure. We define it as **the Prism Hypothesis**, where each data modality can be viewed as a projection of the natural world onto a shared feature spectrum, just like the prism. Building on this insight, we propose **Unified Autoencoding (UAE)**, a model that harmonizes semantic structure and pixel details via an innovative frequency-band modulator, enabling their seamless coexistence. Extensive experiments demonstrate that UAE effectively unifies semantic abstraction and pixel-level fidelity within a single latent space, achieving state-of-the-art performance. Moreover, we show that UAE can be directly applied to pixel-space modeling, significantly improving both FID and IS over the vanilla JIT baseline. Code is avaliable at: https://github.com/WeichenFan/UAE.

1 Introduction

Trained on massive corpora, recent foundation models have profoundly reshaped perception and generation systems, generalizing well across diverse downstream tasks [2,24,26,30]. Yet, early advances in perception and generation evolve along largely separate trajectories. Their objectives are typically distributed across distinct network structures, *e.g.*, employing pretrained semantic encoders [2,26,30], to capture high-level meaning, or pixel encoders [16,32] to compress fine-grained visual detail. While each module excels within its own domain, this fragmentation compels subsequent unification efforts [9, 29, 38] to depend simultaneously

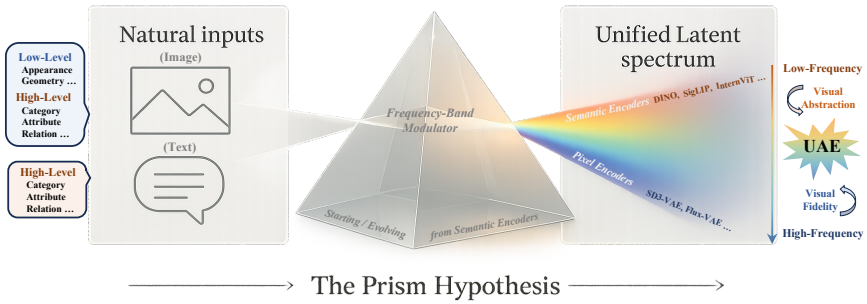

Fig. 1: The Prism Hypothesis. Our conceptual "prism" decomposes various natural inputs into spectral components along frequency. Low frequency bands capture global semantics and abstract meaning, while high frequency bands encode local detail and fine visual texture. This motivates our Unified Autoencoding (UAE), which harmonizes semantic and pixel representations within a single latent space.

on semantic and pixel encoders, forcing networks to reconcile fundamentally heterogeneous representations. The sharp mismatch lowers training efficiency and induces representational conflicts, with these incompatible features often interfering rather than complementing one another.

This fragmentation uncovers a deep-seated tension between abstraction and fidelity, which is a driving force in shaping subsequent foundation models. To alleviate it, recent studies [34, 49, 50] attempt to transfer semantic encoders into visual generation domains alongside strong pixel decoders. This strategy substantially accelerates convergence and improves semantic correspondence, yet remains limited in recovering fine-grained visual details. In parallel, another research direction seeks to endow pixel encoders with semantic awareness through text supervision [44, 45, 47], semantic encoder distillation [3, 10], and hierarchical feature integration [44]. While these efforts move toward unifying understanding and generation representations in a single module, they often achieve coexistence through trade-offs rather than genuine integration.

At the core of this development lies a fundamental question: How is information about the world represented such that multimodal inputs share a common semantic meaning while preserving their native granularity of detail?

Impressively, we empirically observe that pre-trained semantic features, whether derived from text or vision, tend to reside at the coarse end of the decoupled feature spectrum, primarily capturing low-frequency structures such as categories, attributes, and relations. In contrast, pre-trained pixel-level features extend toward the finer end of the spectrum, representing higher-frequency components that convey intricate appearance and geometric detail. Strikingly, these complementary representations can be harmoniously integrated within a unified encoder, extremely aligning well with the spectral arrangement and fostering a progressive synergy from semantic perception to detailed reconstruction.

We posit this as ***the Prism Hypothesis***: as shown in Figure 1, the real-world inputs project onto a continuous shared feature spectrum, and what we perceive as different modalities are distinct slices of this underlying continuum.

From this perspective, we introduce **Unified Autoencoding (UAE)**, a tokenizer that learns a shared latent space and harmonizes semantic structure and pixel-level detail. Specifically, UAE features a frequency decomposition that factorizes real-world content into a fundamental semantic band and residual pixel bands with controllable fine granularity. This design is not only grounded in comprehensive empirical evidence but also validated across diverse reconstruction and perception tasks. With its compact yet semantically expressive representations, UAE outperforms concurrent RAE [50], SVG [34], and UniFlow [49] across rFID, PSNR, gFID and Accuracy metrics on ImageNet, demonstrating that its learned latent space is both semantically representative and pixel-faithful.

Our work makes the following contributions.

1. We introduce the **Prism Hypothesis**, a spectral perspective that explains multimodal data through a shared fundamental band and modality-specific bands, supported by extensive empirical observations.
2. We propose **Unified Autoencoding (UAE)**, a simple yet effective tokenizer via frequency transformation that integrates seamlessly with diffusion transformers and enables flexible generative modeling in the latent space.
3. UAE is effective in **both pixel and latent spaces**. In the latent space, UAE achieves competitive performance in both generative modeling and pixel-level reconstruction. In the pixel space, integrating UAE into JIT [18] leads to faster convergence and improved FID and IS.

2 Related Work

Unified Tokenizers and Unified Representations. Unifying the representations between pixel and semantic embeddings has become a central objective for existing foundation models. Joint embedding approaches align images and text in a shared representation and enable strong zero-shot transfer [13,30], and have been extended to many modalities, *e.g.*, audio, depth, thermal, and inertial signals [7]. In parallel, modality-agnostic backbones aim to build a unified architecture that can process diverse input modalities and generate task-specific outputs through learned queries or a shared token representation [1, 12, 21, 22, 42].

On the tokenizer side, discrete codebook methods have demonstrated that the design and granularity of visual tokens play a crucial role in determining how effectively a single sequence model can adapt to vision tasks [5, 48]. Recent work goes further and seeks tokenizers that support both understanding and generation at the same time. OmniTokenizer [41] learns a joint image video tokenizer with a spatial-temporal decoupled transformer and reports strong reconstruction and synthesis across both domains. Very recent studies deepen this trend. UniFlow [49] proposes a unified pixel flow tokenizer that adapts a pre-trained encoder through layer-wise self-distillation and employs a lightweight patch-wise flow decoder, explicitly targeting the long-standing tension between

semantic abstraction and pixel-faithful reconstruction. Two concurrent works remove the traditional variational bottleneck entirely and build unified representation latents for diffusion transformers. Diffusion Transformers with Representation Autoencoders (RAE) [50] replace the usual reconstruction-only encoder with a pretrained representation encoder, *e.g.*, DINO or SigLIP, and a trained decoder, arguing that semantically rich latents accelerate convergence and improve generative fidelity. SVG [34] trains a diffusion model on DINO features with a small residual branch for details, reporting faster training, few-step sampling, and improved quality.

Our UAE aligns with this shift. It serves as a unified tokenizer that decouples continuous latent features explicitly corresponding to the underlying spectral structure by factorizing real-world contents into a low-frequency base and residual high-frequency bands. It anchors semantics in the core representation while relegating fine-grained details to residuals for progressive reconstruction.

Frequency and Multi-Resolution Modeling. Classical image synthesis adopted pyramids and wavelets to separate structure by scale, enabling coarse-to-fine generation and targeted refinement of detail. A typical example is the Laplacian pyramid of adversarial networks [4], which trains a generator per level and synthesizes images by successively adding higher-frequency residuals. Subsequent analyses of neural networks from a spectral perspective showed that standard architectures prioritize low frequencies and learn higher frequencies later, a phenomenon known as spectral bias [31]. Two lines of work respond to this bias. The first uses input or architecture design to improve access to high frequencies, *e.g.*, Fourier feature mappings and periodic activations that help multi-layer perceptrons represent fine detail [35,37]. The second introduces frequency-aware objectives and signal processing choices, such as focal frequency loss to emphasize hard frequencies and alias-free synthesis to avoid spurious high-frequency artifacts [14,15].

Modern generative models retain this multi-resolution view. Cascaded diffusion trains models at increasing resolutions, allowing each stage to learn the appropriate frequency band and its own error distribution [8]. Variants construct explicit feature pyramids or hierarchical patch schedules, enabling efficient diffusion on large images and video while preserving high-frequency detail [6,36]. Recent latent diffusion designs introduce cross-magnification spaces or zoomable pyramids that share information across scales and enable large-image reconstruction without retraining [46]. In autoregressive models, VAR [39] casts generation as predicting the next scale or resolution, showing strong ImageNet results and clean scaling trends. This explicit progression from global layout to fine detail emerges as a viable alternative to diffusion. Building on this idea, Next Visual Granularity (NVG) generation [43] produces sequences at a fixed resolution but with progressively finer token granularity, surpassing prior VAR baselines while maintaining structured control over detail. Similarly, NFIG [11] performs discrete next-frequency prediction, demonstrating strong generative performance on the ImageNet benchmark. In pixel space, DCTdiff [25] explicitly trun image into DCT space and perform generative modeling. Our approach aligns with

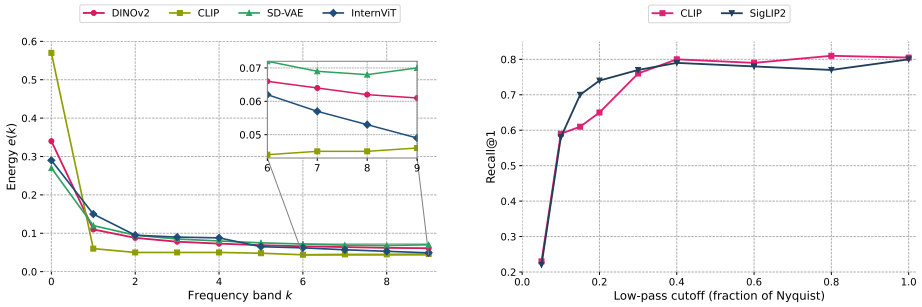

Fig. 2: Left: Frequency energy distribution. Normalized energy $e(k)$ across frequency bands for diverse tokenizers. DINOv2 and CLIP focus on low-frequency (semantic) content, while SD-VAE retains more high-frequency energy, capturing finer details. **Right: Retrieval results via frequency filtering.** Text–Image retrieval remains stable under low-pass filtering but degrades sharply under high-pass filtering, confirming that semantic alignment primarily resides in low-frequency components.

these trends but focuses on unified representation rather than the generator's training schedule.

3 Methodology

3.1 Preliminary Findings

The Prism Hypothesis. Natural inputs are regarded as projections of a common real-world signal onto a shared frequency spectrum. Semantic encoders emphasize a compact low-band that carries categories, attributes, and relations, while pixel encoders observe the same fundamental base together with higher bands that encode edges, textures, and fine appearance. The hypothesis indicates that cross-modal alignment depends primarily on the shared low band.
Formalization. Here, $\mathcal{F}$ and $\mathcal{F}^{\dagger}$ denote two-dimensional discrete Fourier transform and its inverse. For an image $I \in [0,1]^{3\times H\times W}$ and a smooth radial mask $\mathbf{M}_{\rho}^{\mathrm{LP}}$ that passes frequencies within normalized radius $\rho \in (0,1]$,

$$I_{\rho}^{\mathrm{LP}} \;=\; \mathcal{F}^{\dagger}\big(\mathbf{M}_{\rho}^{\mathrm{LP}} \odot \mathcal{F}(I)\big), \tag{1}$$

$$I_{\rho}^{\mathrm{HP}} \;=\; \mathcal{F}^{\dagger}\big(\mathbf{M}_{\rho}^{\mathrm{HP}} \odot \mathcal{F}(I)\big), \tag{2}$$

where $\mathbf{M}_{\rho}^{\mathrm{HP}}$ is complementary to $\mathbf{M}_{\rho}^{\mathrm{LP}}$, and both masks use cosine transitions to limit ringing artifacts. All filtering is performed in linear space prior to any model-specific normalization. With a frozen vision-language encoder E, we compute cosine similarities between text embeddings and image embeddings from I, I_{ρ}^{LP}, and I_{ρ}^{HP}. If semantic alignment is carried by the shared base, then the retrieval score $\mathcal{R}$ satisfies $\mathcal{R}_{\mathrm{LP}}(\rho)$ is nondecreasing in ρ, $\mathcal{R}_{\mathrm{HP}}(\rho)$ is nonincreasing in ρ, and $\mathcal{R}_{\mathrm{HP}}(\rho)$ approaches chance once the shared base is removed.

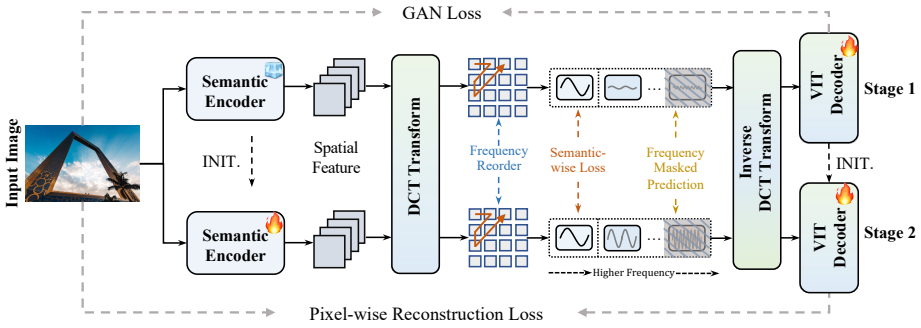

Fig. 3: Overall architecture of our proposed Unified Autoencoding (UAE). The input image is separately encoded by both a pretrained *Semantic Encoder* (e.g., DINOv2) and the trainable *Unified Encoder*. The unified encoder is initialized from the semantic encoder and optimized under two complementary objectives: a **semantic-wise loss** that aligns low-frequency components decomposed from the semantic encoder's representations, and a **pixel-wise reconstruction loss** that enforces visual fidelity via the *Pixel Decoder* by adaptively dilating the high-frequency components. The decoder employs spectral transform blocks to refine residual-frequency content and produce the reconstructed image. This joint optimization harmonizes semantic structure and pixel detail within a single latent space.

Empirical Verification. We empirically examine the Prism Hypothesis using two complementary analyses: (i) token-space frequency energy decomposition and (ii) robustness of text–image alignment under controlled spectral filtering.

Exp1 (Token-spectrum energy). For each encoder, we extract patch tokens from the input images, reshape them into a 2D token grid, and apply a channel-wise 2D DCT to obtain token-frequency coefficients. We then reorder the coefficients via zig-zag traversal (progressing from low to high frequencies), partition them into coarse bands, and compute the normalized band energy $e(k)$ as the average of squared magnitudes across channels and images. As shown in Fig. **??** (left), the resulting spectra are dominated by the lowest band, revealing a strong low-frequency bias across models (which is consistent with the natural images). Importantly, different encoders allocate different *relative* mass to higher bands: Semantic representation encoders concentrate more energy in low-frequency components, while the pixel-centric SD-VAE preserves relatively more energy in the mid- and high-frequency bands (inset), consistent with its stronger retention of fine-grained appearance details.

Exp2 (Low-pass retrieval). To directly investigate the frequency localization of cross-modal semantics, we evaluate text–image retrieval under progressive removal of high-frequency image components. Concretely, we apply a radial low-pass mask in the frequency domain, with a cutoff specified as a fraction of the Nyquist frequency, reconstruct the filtered images, and measure Recall@1 using frozen vision–language models (CLIP and SigLIP2) with fixed text prompts (*an image of ...*). Fig. **??** (right) shows that Recall@1 increases rapidly when only a small low-frequency portion is retained (e.g., $\sim$0.23 at 0.05 Nyquist vs. $\sim$0.58 at

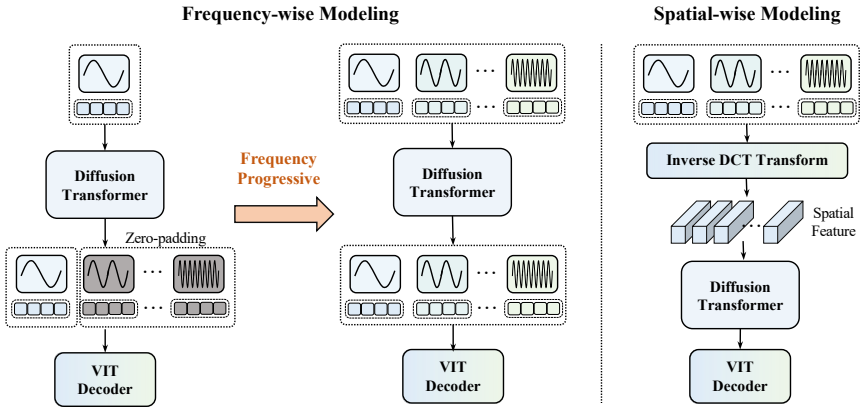

Fig. 4: Frequency-wise vs. spatial-wise modeling. *Left:* The diffusion transformer operates directly in the DCT domain and progressively models frequency tokens from low to high bands before decoding. *Right:* Frequency tokens are first transformed back to spatial features via inverse DCT, and diffusion is performed in the spatial domain prior to decoding.

0.10), and saturates near full-image performance by moderate cutoffs (around 0.3–0.4 Nyquist). This behavior indicates that a significant fraction of retrieval-relevant semantics is encoded in coarse, low-frequency image structure, whereas higher-frequency components are largely unrelated to semantic information.

Taken together, Exp1 and Exp2 support the central intuition of the Prism Hypothesis: cross-modal semantic alignment is primarily associated with a shared low-frequency base, whereas higher-frequency bands increasingly encode modality-specific, fine-grained visual detail.

3.2 Unified AutoEncoder

Using DINOv2 as an example, we initialize UAE from the pretrained encoder and discard the register tokens, retaining only the patch tokens with channel dimension C. Input images are center-cropped and resized to the encoder's native input resolution. The resulting patch-token sequence is then reshaped into a 2D latent grid $\mathbf{z} \in \mathbb{R}^{B\times C\times H\times W}$, which enables explicit frequency-domain processing.

Frequency transform and tokenization. Given $\mathbf{z}$, we apply a channel-wise 2D discrete cosine transform (DCT) $f(\cdot)$ to obtain a frequency-domain representation

$$\mathbf{h} = f(\mathbf{z}), \qquad \mathbf{h} \in \mathbb{R}^{B\times C\times H\times W}. \tag{3}$$

We then linearize the 2D DCT grid into a 1D token sequence via a frequency-preserving reordering scheme.

Token Reorder. In the 2D DCT grid, each coefficient $C(u, v)$ (not to be confused with the channel dimension C) corresponds to a spatial frequency indexed

by (u, v). A natural measure of frequency magnitude is the radial norm

$$r(u, v) = \sqrt{u^2 + v^2}. \tag{4}$$

As (u, v) moves away from the top-left DC component $(0, 0)$, $r(u, v)$ increases in discrete levels, yielding progressively higher spatial frequencies. This allows us to interpret the DCT lattice as a set of concentric "radial shells" centered at $(0, 0)$.

To convert the 2D frequency grid into a 1D token sequence while approximately preserving low-to-high frequency progression, we adopt zig-zag reordering. Zig-zag traversal roughly sorts coefficients by increasing $u + v$, which serves as a practical proxy for radial growth (since $u^2 + v^2$ typically increases with $u + v$ for small-to-moderate indices). This produces an ordered sequence

$$\{C_k\}_{k=0}^{N-1}, \qquad N = H \times W, \tag{5}$$

where the frequency magnitude increases approximately monotonically with k.

Semantic Regularization. To preserve the semantic priors of the pretrained teacher while expanding the representational bandwidth, we apply a *semantic-wise* alignment loss only on the lowest-frequency bands. Let $\mathbf{f}_s$ denote features from the frozen pretrained semantic encoder (teacher), and $\mathbf{f}_u$ denote features from the trainable unified encoder (student). After frequency decomposition, we obtain band-wise representations $\{\mathbf{f}_s^k\}_{k=0}^{K-1}$ and $\{\mathbf{f}_u^k\}_{k=0}^{K-1}$. We enforce alignment only for the first K_{base} low-frequency bands, which primarily encode global structure and category-level semantics, while leaving higher frequency less constrained:

$$\mathcal{L}_{\text{sem}} = \frac{\sum_{k=0}^{K_{\text{base}}-1} \|\mathbf{f}_u^k - \mathbf{f}_s^k\|_2^2}{K_{\text{base}}}. \tag{6}$$

This restricted supervision transfers the teacher's semantic organization to the low-frequency subspace, while leaving higher-frequency bands unconstrained so they can specialize in modality-specific, pixel-level detail. Empirically, this selective regularization stabilizes training and mitigates collapse toward purely pixel-driven features.

Frequency Masked Prediction. During training, we randomly mask a subset of high-frequency tokens to encourage robust inference of missing fine details. Let $\mathcal{L}$ and $\mathcal{H}$ denote the index sets of low- and high-frequency components, respectively. We construct a masked frequency representation $\tilde{\mathbf{F}}$ as

$$\tilde{\mathbf{F}}_i = \begin{cases} \mathbf{F}_i, & i \in \mathcal{L}, \\ 0, & i \in \mathcal{H}_{\text{mask}}, \end{cases} \tag{7}$$

where $\mathcal{H}_{\text{mask}} \subseteq \mathcal{H}$ is sampled independently at each iteration. That is, selected high-frequency coefficients are set to zero, while low-frequency components remain intact.

The masked frequency tokens are mapped back to the spatial domain via inverse DCT and fed to the decoder. The decoder is trained to reconstruct the full

Table 1: Reconstruction performance on ImageNet-1K and MS-COCO 2017. UAE consistently achieves strong PSNR/SSIM and competitive rFID compared with both continuous autoencoders and unified tokenizers. We additionally report linear probing accuracy to evaluate representation quality.

Method	Type	Ratio	Linear prob (Acc)↑	ImageNet-1K			MS-COCO 2017		
				PSNR↑	SSIM↑	rFID↓	PSNR↑	SSIM↑	rFID↓
SD-VAE-1.x [33]	Continuous	8	–	23.54	0.68	1.22	23.21	0.69	5.94
SD-VAE [33]	Continuous	8	–	25.68	0.72	0.75	25.43	0.73	0.76
SD-VAE-2.x [33]	Continuous	8	–	23.54	0.68	1.22	26.62	0.77	4.26
SD-VAE-XL [28]	Continuous	8	–	27.37	0.78	0.67	27.08	0.80	3.93
SD-VAE-3 [20]	Continuous	8	–	31.29	0.87	0.20	31.18	0.89	1.67
FLUX-VAE [17]	Continuous	8	–	**32.74**	**0.92**	**0.18**	**32.32**	**0.93**	**1.35**
VA-VAE	Continuous	16	–	27.96	0.79	0.28	27.50	0.81	2.71
SVG (DINOv3) [34]	Continuous	16	–	24.25	0.67	0.78	-	-	-
RAE (DINOv2-B) [50]	Continuous	14	–	18.05	0.5	2.04	18.36	0.47	6.01
UniFlow (DINOv2-L) [49]	Continuous	14	–	32.32	0.91	0.17	30.66	0.94	2.81
UAE (Siglip2)	Continuous	16	**80.1 (81.2)**	**31.00**	**0.91**	**0.43**	**30.20**	**0.89**	**2.91**
UAE (DINOv2-B)	Continuous	14	**83.0 (83.0)**	**32.17**	**0.92**	**0.35**	**31.19**	**0.91**	**2.01**
UAE (CLIP-L)	Continuous	14	**77.2 (79.4)**	**36.58**	**0.96**	**0.04**	**36.25**	**0.97**	**0.41**

high-fidelity image, including the masked high-frequency details, conditioned only on the preserved low-frequency structure and the remaining visible coefficients. This objective encourages the model to infer fine-grained textures and high-frequency signals from global semantic cues, improving robustness and generalization.

3.3 Generative Modeling

As illustrated in Figure 4, UAE latents admit two complementary parameterizations for generative modeling: *spatial-domain* features and *frequency-domain* coefficients. In **spatial-wise** modeling, we first apply the inverse DCT to map the frequency tokens back to a spatial latent grid, and then train a standard latent diffusion transformer on these spatial features, following prior representation-based tokenizers [50]. In **frequency-wise** modeling, we operate directly on the reordered DCT coefficients. Since our decoder can reconstruct high-quality images even when conditioned only on low-frequency tokens, we can adopt a coarse-to-fine generation strategy: the diffusion transformer is trained to model low-frequency components first and is then progressively extended to incorporate higher-frequency bands, yielding increasingly fine-grained details.

4 Experiments

4.1 Implementation Details

UAE Training. We train UAE with DINOv2-B, DINOv2-L, SigLIP2, and InternViT at 256×256 resolution. Training is performed in two stages. In **Stage 1**, we freeze the pretrained semantic encoder and train the decoder using a pixel-wise reconstruction loss and an adversarial loss, following the original RAE setting [50]. We use AdamW with an initial learning rate of 2×10^{-4}, linearly decayed

to 2×10^{-5}, and train for 16 epochs. The discriminator is optimized with AdamW using the same learning-rate schedule; discriminator training starts at epoch 6, and the GAN loss term on the generator is enabled from epoch 8. In **Stage 2**, we unfreeze the encoder and fine-tune the full model using the semantic-wise loss together with the reconstruction loss. We additionally apply frequency-masked prediction by randomly masking high-frequency components and training the model to reconstruct the full image under the GAN objective. The masking ratio is set to 75% for DINOv2 and 50% for the other encoders. We set K_{base} to 25% of the total tokens for DINOv2 and 50% for the remaining models.Unless otherwise specified, Stage 2 uses the same optimizer and learning-rate schedule as Stage 1, with high-frequency masking enabled from the beginning of Stage 2.
Generative Modeling. We train SiT on two UAE latent parameterizations: spatial latents ($768 \times 16 \times 16$) and frequency latents (768×256). In both cases, training is performed progressively over frequency tokens: we start from 64 tokens for the first 80 epochs, then linearly increase the token count to 256, updating the token budget every 20 epochs. All models are trained for a total of 1400 epochs to ensure stable convergence. We adopt the AdamW optimizer with an initial learning rate of 2×10^{-4}, which is linearly decayed to 2×10^{-5} over the course of training.

4.2 Visual Reconstruction

Quantitative Evaluation. In Table 1, We quantify reconstruction quality at 256×256 on ImageNet-1K and MS-COCO 2017. The compared baselines include widely-used generative tokenizers and variational decoders. For a fair comparison, we report the evaluation results for our UAE and the corresponding baseline that follows the same configurations of DINOv2-base. Here, we report PSNR and SSIM for fidelity, and rFID for perceptual quality.

Notably, our UAE delivers state-of-the-art reconstruction quality among unified tokenizers. On ImageNet-1K, UAE improves over the RAE baseline from 18.05 to 31.00 in PSNR and from 0.50 to 0.92 in SSIM, while reducing FID from 2.04 to 0.35. On MS-COCO, UAE raises PSNR from 18.36 to 31.19 and SSIM from 0.47 to 0.91, with FID decreasing from 6.01 to 2.01. These notable gains validate that moving masking out of the decoder and factorizing latents into a low-frequency base and residual bands preserve fine detail while maintaining semantic structure.

Beyond unified tokenizers, UAE is competitive with the best generative-only tokenizers. On ImageNet-1K, UAE(CLIP-L) attains an FID of 0.04, surpassing SD3 VAE and approaching Flux VAE, while maintaining high PSNR and SSIM. A similar pattern holds on MS-COCO, where UAE(CLIP-L) surpasses the strongest baselines in perceptual quality.
Qualitative Evaluation. Figure 5 compares source images with the resulting reconstructions from FLUX-VAE, RAE, and UAE. Note that our UAE can well preserve straight edges, fine textures, and small text, *e.g.*, street signs and printed documents. The improvement is consistent across natural photos and

Fig. 5: Qualitative comparison of reconstruction fidelity across autoencoding paradigms. We visualize reconstructed samples from representative methods, including Flux-VAE [17], RAE [50], and our proposed UAE. Each row corresponds to reconstructions from a fixed source set spanning text, human, object, and artistic domains. UAE produces the most consistent and semantically faithful reconstructions, preserving both high-frequency details (e.g., texture and edge sharpness) and global structure (e.g., layout and color harmony), while reducing the blurring and semantic drift observed in RAE.

illustrations. Moreover, we demonstrate that UAE is effective across different semantic encoders, including DINOv2 [26], CLIP [30], and SigLip2 [40].

4.3 Generative Modeling

We evaluate the generative capability of UAE by training SiT models on the proposed latent representations and performing class-conditional image generation on ImageNet 256×256. We report results under both unguided sampling and classifier-free guidance (CFG), following standard evaluation protocols with FID, Inception Score (IS), Precision, and Recall. As shown in Table 2, UAE achieves strong generative performance compared with prior latent tokenizers and diffusion-based approaches. In particular, UAE-Frequency (DINOv2-B) at-

Table 2: Class-conditional generation performance on ImageNet (256x256). We compare our proposed UAE with recent diffusion and autoregressive models using standard metrics.

Methods	gFID↓	IS↑	Prec↑	Rec↑
DiT [27]	9.62	121.5	0.67	0.67
SiT [23]	8.61	131.7	0.68	0.67
SVG [34]	3.36	181.2	-	-
UniFlow [49]	2.45	228.0	-	-
RAE [50]	**1.51**	**242.9**	0.79	**0.63**
UAE	1.52	234.5	**0.81**	0.62

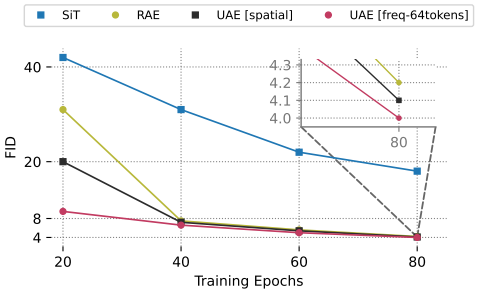

Fig. 6: UAE (spatial and frequency) converges significantly faster and achieves lower FID than SiT and RAE. The zoom-in highlights the final performance gap at 80 epochs.

tains the best overall FID under both unguided (gFID=1.37) and guided sampling (gFID=1.10), outperforming strong baselines such as RAE and DiT while maintaining competitive precision and recall. UAE-Spatial also achieves competitive results, demonstrating the effectiveness of the proposed unified latent representation. These results indicate that the frequency-aware latent structure of UAE provides a favorable representation space for generative modeling, enabling diffusion transformers to capture global structure and fine-grained details.

5 Discussion

5.1 The Impact of Generative Modeling

We further study the effect of generative modeling using UAE representations. As shown in Fig. 6, UAE significantly accelerates diffusion model convergence compared with SiT and RAE. Both spatial and frequency variants consistently achieve lower FID throughout training and reach high-quality generation within substantially fewer epochs. In particular, UAE attains strong performance as early as 40 epochs and continues to improve steadily, while baseline methods converge much more slowly. The zoom-in view highlights the final performance gap at 80 epochs, where UAE achieves the best FID. These results suggest that the frequency-aware representation of UAE provides a more generation-friendly latent space, enabling more efficient generative modeling.

5.2 Inference Efficiency

We compare the computational efficiency of different tokenizer and latent modeling configurations in Table 3. The proposed UAE tokenizer introduces negligible overhead in stage-1 reconstruction, exhibiting nearly identical GFLOPs and latency to the RAE baseline. In contrast, the efficiency gain becomes significant in stage-2 generative modeling. When the SiT backbone operates on the compact UAE 64 low-freq tokens, the computational cost is reduced from 313.84 GFLOPs

Table 3: Efficiency comparison across different representations. UAE matches RAE's computational cost, while *SiT (UAE-freq-64tokens)* achieves a 4× GFLOPs reduction and over 2× lower latency than the RAE baseline.

Model	GFLOPs	Latency
RAE	266.24	8.40
UAE-spatial	266.24	8.40
UAE-freq	266.27	8.76
SiT (RAE)	313.84	8.82
SiT (UAE-freq-64tokens)	**78.80**	**4.05**

Table 4: Ablations on semantic loss and mask prediction. (a) varies the fraction of base tokens K_{base} supervised by the semantic loss, reporting reconstruction PSNR and linear-probe accuracy. **(b)** varies the mask-prediction ratio, reporting gFID at 20 training epochs.

(a)

K_{base}	PSNR↑	ACC↑
0%	34.3	60.1
25%	32.2	83.0
50%	28.1	83.0
75%	20.1	83.0

(b)

Ratio	gFID@20 epoch↓
85%	9.7
15%	12.2
50%	27.2
25%	33.1

to 78.80 GFLOPs per image, resulting in a ∼4× reduction, while inference latency decreases from 8.82 ms to 4.05 ms per image.

5.3 The Impact of Semantic Loss

We ablate the strength of semantic regularization by varying K_{base} as shown in the Table 4 (a), the fraction of low-frequency tokens constrained by the semantic alignment loss using DINOv2 as the teacher. When semantic regularization is disabled ($K_{\text{base}} = 0$), the model achieves strong reconstruction but exhibits weak semantic alignment, as reconstruction gradients favor local appearance cues. Increasing K_{base} improves semantic performance by anchoring the low-frequency latent space to the pretrained semantic geometry, but excessive constraints reduce reconstruction fidelity. We find that aligning only 25% of frequency tokens already preserves most semantic capability while maintaining high reconstruction quality, supporting the design that semantics mainly reside in the low-frequency base while higher frequencies capture fine details.

5.4 The Impact of Mask Predictions

We further analyze frequency-masked prediction during UAE training by varying the masking ratio applied to high-frequency tokens. This strategy acts as a regularizer, forcing the decoder to reconstruct missing high-frequency details from the preserved low-frequency structure and the remaining visible components. As shown in Table 4(b), aggressive masking yields the best generative performance: masking 85% of high-frequency tokens achieves the lowest gFID (9.7) at 20 epochs. A small masking ratio provides limited regularization, leading the model to rely on deterministic high-frequency cues. Thus, like visual understanding, generative modeling relies heavily on low-frequency components.

5.5 The Impact of Unified Training

As shown in Fig. 7, we analyze the frequency energy distribution of vanilla encoders and their unified-trained counterparts under the same setting described

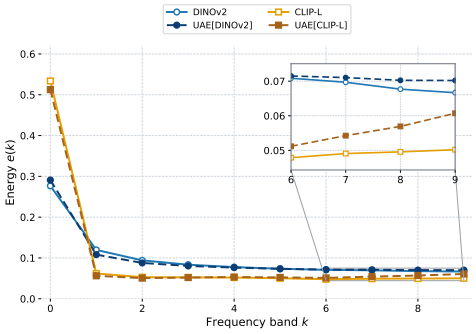

Fig. 7: Energy distribution across frequency bands before and after unified training. Unified training shifts energy from low- to mid- and high-frequency bands.

Table 5: Comparison of generative performance in pixel space. We evaluate JIT and its unified-trained variant UAE (JIT) under the same architecture, model size, and patch size. UAE consistently improves both FID (↓) and IS (↑) across training stages.

Model	Epoch	Size	Patch Size	FID↓	IS↑
JIT [18]	40	700M	32	31.81	60.13
JIT [18]	80	700M	32	16.06	123.30
UAE (JIT)	40	700M	32	**25.80**	**72.82**
UAE (JIT)	80	700M	32	**14.55**	**132.57**

in Sec. 3.1. Unified training induces a consistent redistribution of energy across frequency bands. Specifically, the UAE models exhibit reduced dominance in low-frequency components, accompanied by a noticeable increase in mid- and high-frequency energy. This shift suggests that unified training enhances the representation of fine-grained details while largely preserving global semantic structure. The zoom-in view further confirms that high-frequency components are better retained or even amplified after unification.

5.6 UAE in Pixel Space

We further evaluate the effectiveness of unified training in the pixel-space generative setting by applying UAE to the JIT framework. As shown in Table 5, we compare the original JIT model with its unified-trained counterpart under identical configurations. UAE consistently outperforms the vanilla JIT across all metrics and training stages. UAE achieves a notable improvement, consistently achieve lower fid score and high IS score. These results indicate that unified training extends beyond latent space and remains effective when applied directly in pixel space.

5.7 Transfer to Multimodal Understanding

(a) Zero-shot drop-in. Replacing LLaVA-1.5 [19]'s vision tower with UAE-CLIP-L at native resolution—keeping the released projector and language model frozen, with no training of any kind—recovers ∼97% of vanilla LLaVA-1.5's POPE accuracy (0.811 vs. 0.832) and matches BLEU-1/CIDEr to within ∼2%. Notably, UAE-CLIP-L is *more precise* than the original tower (POPE precision 0.953 vs. 0.896): even on a pipeline fitted for a different encoder, it grounds objects faithfully, and the residual gap reflects recall rather than hallucination. **(b) Model-agnostic Prism test.** On *vanilla* LLaVA-1.5 (no UAE involved),

Table 6: Transfer to multimodal understanding on POPE-popular (3K queries) and captioning. **(a)** Zero-shot drop-in: LLaVA-1.5's vision tower is replaced by UAE-CLIP-L with the released projector and LLM kept unchanged (no training of any kind). **(b)** Model-agnostic Prism test: on *vanilla* LLaVA-1.5, a post-hoc 2D-DCT low-pass keeps only the top-left $N \times N$ coefficient block before the unchanged projector.

Setting	Tokens	POPE Acc↑	POPE Prec↑	BLEU-1↑	CIDEr↑
LLaVA-1.5 (Stage-1+2, ref.)	576	0.832	0.896	0.67	0.87
(a) UAE-CLIP-L drop-in, zero-shot (no training)					
UAE-CLIP-L	64	0.704	0.970	0.58	0.58
UAE-CLIP-L	256	0.808	0.954	0.64	0.78
UAE-CLIP-L	576	0.811	0.953	0.66	0.84
(b) Vanilla LLaVA-1.5 + post-hoc DCT low-pass ($N \times N$ block)					
CLIP-L	16	0.659	0.915	0.52	0.41
CLIP-L	64	0.792	0.958	0.62	0.70
CLIP-L	144	0.832	0.954	0.65	0.80
CLIP-L	256	0.846	0.944	0.66	0.84

we apply a 2D-DCT to the projector inputs, zero all coefficients outside the top-left $N \times N$ block, invert the transform, and feed the result to the unchanged projector and LLM. The recovery curve saturates at 144 tokens—exactly the 25% low-frequency fraction (K_{base}=25%) used by UAE. At this point, accuracy already matches the full-grid reference, and higher-frequency tokens add little on this semantic task. This independently corroborates the Prism Hypothesis: semantic content concentrates in a compact low-frequency band.

6 Conclusion

We propose the Prism Hypothesis, which views diverse natural inputs as projections of a shared spectrum composed of a compact low-frequency semantic component and residual higher-frequency detail. Preliminary experiments about frequency-band distributions of various visual and textual encoders strongly enlighten this connection between feature spectra and representational function. Hence, we launch Unified AutoEncoding (UAE), which harmonizes semantic and pixel information within a single latent space via a frequency decomposition module. Extensive experiments confirm that it delivers greater generative capability over existing unified tokenizers, *e.g.*, RAE, SVG, and UniFlow, meanwhile preserving reconstruction quality on par with top-tier models like Flux-VAE. UAE offers a route toward unified tokenizers for understanding and generation.

7 Acknowledgment

This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOE-T2EP20223-0002). This research is also supported by cash and in-kind funding from NTU S-Lab and industry partner(s).

References

1. Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M.: Data2vec: A general framework for self-supervised learning in speech, vision and language. In: International conference on machine learning. pp. 1298–1312. PMLR (2022)
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
3. Chen, J., Zhu, D., Qian, G., Ghanem, B., Yan, Z., Zhu, C., Xiao, F., Culatana, S.C., Elhoseiny, M.: Exploring open-vocabulary semantic segmentation from clip vision encoder distillation only. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 699–710 (2023)
4. Denton, E.L., Chintala, S., Fergus, R., et al.: Deep generative image models using a laplacian pyramid of adversarial networks. Advances in neural information processing systems **28** (2015)
5. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
6. Fan, W.C., Chen, Y.C., Chen, D., Cheng, Y., Yuan, L., Wang, Y.C.F.: Frido: Feature pyramid diffusion for complex scene image synthesis. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 579–587 (2023)
7. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15180–15190 (2023)
8. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. Journal of Machine Learning Research **23**(47), 1–33 (2022)
9. Huang, R., Wang, C., Yang, J., Lu, G., Yuan, Y., Han, J., Hou, L., Zhang, W., Hong, L., Zhao, H., et al.: Illume+: Illuminating unified mllm with dual visual tokenization and diffusion refinement. arXiv preprint arXiv:2504.01934 (2025)
10. Huang, W., Peng, Z., Dong, L., Wei, F., Jiao, J., Ye, Q.: Generic-to-specific distillation of masked autoencoders. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15996–16005 (2023)
11. Huang, Z., Qiu, X., Ma, Y., Zhou, Y., Chen, J., Zhang, H., Zhang, C., Li, X.: Nfig: Multi-scale autoregressive image generation via frequency ordering. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
12. Jaegle, A., Borgeaud, S., Alayrac, J.B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., et al.: Perceiver io: A general architecture for structured inputs & outputs. arXiv preprint arXiv:2107.14795 (2021)
13. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
14. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Focal frequency loss for image reconstruction and synthesis. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13919–13929 (2021)
15. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. Advances in neural information processing systems **34**, 852–863 (2021)

16. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
17. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
18. Li, T., He, K.: Back to basics: Let denoising generative models denoise. arXiv preprint arXiv:2511.13720 (2025)
19. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
20. Lopez, J.: Stable diffusion 3: Research paper. `https://stability.ai/news/stable-diffusion-3-research-paper` (Mar 2025), stability AI
21. Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., Kembhavi, A.: Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26439–26455 (2024)
22. Lu, J., Clark, C., Zellers, R., Mottaghi, R., Kembhavi, A.: Unified-io: A unified model for vision, language, and multi-modal tasks. arXiv preprint arXiv:2206.08916 (2022)
23. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: European Conference on Computer Vision. pp. 23–40. Springer (2024)
24. Nguyen, D.K., Assran, M., Jain, U., Oswald, M.R., Snoek, C.G., Chen, X.: An image is worth more than 16x16 patches: Exploring transformers on individual pixels. arXiv preprint arXiv:2406.09415 (2024)
25. Ning, M., Li, M., Su, J., Jia, H., Liu, L., Beneš, M., Salah, A.A., Ertugrul, I.O.: Dctdiff: Intriguing properties of image generative modeling in the dct space. arXiv preprint arXiv:2412.15032 (2024)
26. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
27. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
28. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
29. Qu, L., Zhang, H., Liu, Y., Wang, X., Jiang, Y., Gao, Y., Ye, H., Du, D.K., Yuan, Z., Wu, X.: Tokenflow: Unified image tokenizer for multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2545–2555 (2025)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
31. Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., Courville, A.: On the spectral bias of neural networks. In: International conference on machine learning. pp. 5301–5310. PMLR (2019)
32. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems **32** (2019)

33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
34. Shi, M., Wang, H., Zheng, W., Yuan, Z., Wu, X., Wang, X., Wan, P., Zhou, J., Lu, J.: Latent diffusion model without variational autoencoder. arXiv preprint arXiv:2510.15301 (2025)
35. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. Advances in neural information processing systems **33**, 7462–7473 (2020)
36. Skorokhodov, I., Menapace, W., Siarohin, A., Tulyakov, S.: Hierarchical patch diffusion models for high-resolution video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7569–7579 (2024)
37. Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. Advances in neural information processing systems **33**, 7537–7547 (2020)
38. Tang, H., Xie, C., Bao, X., Weng, T., Li, P., Zheng, Y., Wang, L.: Unilip: Adapting clip for unified multimodal understanding, generation and editing. arXiv preprint arXiv:2507.23278 (2025)
39. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems **37**, 84839–84865 (2024)
40. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
41. Wang, J., Jiang, Y., Yuan, Z., Peng, B., Wu, Z., Jiang, Y.G.: Omnitokenizer: A joint image-video tokenizer for visual generation. Advances in Neural Information Processing Systems **37**, 28281–28295 (2024)
42. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al.: Image as a foreign language: Beit pretraining for vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19175–19186 (2023)
43. Wang, Y., Wang, Z., Wu, Z., Tao, Q., Liao, K., Loy, C.C.: Next visual granularity generation. arXiv preprint arXiv:2508.12811 (2025)
44. Wu, J.J., Chang, A.C.H., Chuang, C.Y., Chen, C.P., Liu, Y.L., Chen, M.H., Hu, H.N., Chuang, Y.Y., Lin, Y.Y.: Image-text co-decomposition for text-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26794–26803 (2024)
45. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18134–18144 (2022)
46. Yellapragada, S., Graikos, A., Triaridis, K., Prasanna, P., Gupta, R., Saltz, J., Samaras, D.: Zoomldm: Latent diffusion model for multi-scale image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23453–23463 (2025)
47. Yi, M., Cui, Q., Wu, H., Yang, C., Yoshie, O., Lu, H.: A simple framework for text-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7071–7080 (2023)

48. Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Birodkar, V., Gupta, A., Gu, X., et al.: Language model beats diffusion–tokenizer is key to visual generation. arXiv preprint arXiv:2310.05737 (2023)
49. Yue, Z., Zhang, H., Zeng, X., Chen, B., Wang, C., Zhuang, S., Dong, L., Du, K., Wang, Y., Wang, L., et al.: Uniflow: A unified pixel flow tokenizer for visual understanding and generation. arXiv preprint arXiv:2510.10575 (2025)
50. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. arXiv preprint arXiv:2510.11690 (2025)