Article

Not peer-reviewed version

Visual Tokenizers: A Comprehensive Survey

Qingyu Shi , Haobo Yuan , Yikang Zhou , Xiu Li , Jason Li , Size Wu , Haochen Wang , Ye Tian , Tao Zhang , Jinbin Bai , Yujing Wang , Lu Qi , Yunhai Tong *, Ming-Hsuan Yang *

Posted Date: 23 July 2026

doi: 10.20944/preprints202607.1756.v1

Keywords: visual tokenizer; multimodal understanding; multimodal generation

 doi:10.20944/preprints202607.1756.v1

Article

Visual Tokenizers: A Comprehensive Survey

Qingyu Shi [1], Haobo Yuan [2], Yikang Zhou [3], Xiu Li [4], Jason Li [5], Size Wu [5], Haochen Wang [6], Ye Tian [1], Tao Zhang [3], Jinbin Bai [7], Yujing Wang [1], Lu Qi [3], Yunhai Tong [1,*] and Ming-Hsuan Yang [2,*]

[1] National Key Laboratory of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University, Beijing, China
[2] UC Merced
[3] Wuhan University, Wuhan, China
[4] Tsinghua University, Beijing, China
[5] Nanyang Technological University, Singapore
[6] University of Chinese Academy of Sciences
[7] Nanjing University
* Correspondence: yhtong@pku.edu.cn (Y.T.); mhyang@ucmerced.edu (M.-H.Y.)

Abstract

Visual tokenization bridges the gap between high-dimensional visual data and sequence modeling by transforming images, videos, and 3D content into compact token sequences. Recent advances in Multimodal Large Language Models (LLMs) have further underscored the critical role of visual tokenizers. Despite this rapid progress, the literature in this area remains fragmented. This work addresses this gap by presenting a comprehensive taxonomy of visual tokenizers. We review the evolution of task-specific tokenizers along two complementary dimensions: high-level tokenizers that emphasize semantic representations and low-level tokenizers that focus on pixel reconstruction and compression. Motivated by the emergence of Multimodal Large Language Models capable of supporting both understanding and generation tasks, we highlight recent advances in unified tokenizers. These models aim to integrate semantic abstraction with fine-grained visual details within a single representation. We categorize unified tokenizers into single-encoder and dual-encoder architectures and further organize them according to their semantic preservation strategies and fusion mechanisms. We also summarize empirical results on downstream understanding and generation benchmarks, comparing unified tokenizers with both individual task-specific models and combinations of separate tokenizers. Finally, we discuss open challenges and future directions, including the trade-off between semantic abstraction and reconstruction fidelity, scalable token vocabularies, and extensions to video and 3D domains. We further examine the evolving role of visual tokens in next-generation multimodal systems. We maintain a curated list of related works at https://github.com/Shi-qingyu/Awesome-Visual-Tokenizer.

Keywords: visual tokenizer; multimodal understanding; multimodal generation

1. Introduction

The success of modern natural language processing relies on discrete token representations that enable scalable sequence modeling [1,2]. In contrast, visual data is continuous, high-dimensional, and spatially structured, which poses fundamental challenges for applying sequential modeling directly. Visual tokenizers [3,4] address this issue by transforming raw visual inputs, such as images, videos, and 3D data, into compact token sequences that can be processed by Transformers and related models. In image generation models, these tokenizers map the original image or video into a latent space, enabling generation in a lower-dimensional and more tractable representation space. In multimodal language models, tokenizers convert visual inputs into visual tokens that can be aligned and integrated

 doi:10.20944/preprints202607.1756.v1

with text tokens. Consequently, visual tokenization has become a foundational component of modern computer vision and multimodal systems.

Research on visual tokenization has evolved through several distinct stages. Early works [5–8] were typically driven by specific tasks and domains. On one hand, high-level tokenizers [5,8] are designed to extract semantically meaningful representations for visual recognition, retrieval, and dense prediction. On the other hand, low-level tokenizers [6,7] focus on compressing visual inputs for reconstruction and generation, prioritizing pixel-level fidelity through continuous or discrete latent representations. Historically, these two lines of research developed largely independently, following different objectives, architectures, and evaluation criteria.

The recent emergence of Large Language Models (LLMs) has significantly reshaped the role of visual tokenizers. Multimodal Large Language Models (MLLMs) leverage the strong capabilities of foundation LLMs, enabling a single model to handle multiple tasks, including both high-level understanding and low-level generation. This paradigm requires visual tokenizers to provide representations that are suitable for both types of tasks. Although some approaches [9–11] employ two separate tokenizers to obtain high-level and low-level features, a unified representation is ultimately desirable. Unified representations play a critical role in downstream training and facilitate effective interaction among different tasks [12,13]. Consequently, research on unified tokenizers has emerged rapidly in recent years.

Despite this rapid progress, the literature on visual tokenization remains fragmented. Existing methods differ substantially in architectural design, token representation constraints, training objectives, and scaling behaviors. Moreover, the term "visual tokenizer" is often used interchangeably with encoders, autoencoders, or representation learners. This inconsistency obscures the common structure underlying these approaches. The lack of a unified taxonomy and systematic comparison makes it difficult to understand design trade-offs, identify common patterns, or assess how recent unified methods relate to earlier task-specific tokenizers. Although some recent works summarize discrete tokenizers, these studies primarily focus on methods represented by VQ-VAE [7], while largely overlooking continuous tokenizers used for understanding [5] and generation tasks [6]. In this work, we also summarize unified tokenizers [12,14,15], providing insights that may facilitate the development of future unified models [11,16,17].

This work aims to provide a coherent and comprehensive overview of visual tokenizers. We first formalize the concept of visual tokenization and break down tokenizers into three core components: an encoder that extracts visual features, a bottleneck constraint mechanism that shapes the token representation, and a decoder that reconstructs or supervises visual content. And then we summarize the development of task-specific tokenizers. We categorize the techniques that enhance specific task-required abilities, distinguishing between high-level and low-level tokenizers. High-level tokenizers focus on training strategies and data scaling to enhance the semantic richness of token representations. Low-level tokenizers, on the other hand, focus on pixel reconstruction and compression. We also provide a summary of representative tokenizers' performance on well-known benchmarks to facilitate horizontal comparisons.

Building on the need to support both high-level semantics and low-level visual details, we further discuss recent advances in unified tokenizers [12–15]. These methods aim to tokenize semantic abstractions and fine-grained visual information within a single, shared feature space. We categorize existing approaches primarily by architectural design, and broadly group unified tokenizers into single-encoder [12] and dual-encoder paradigms [18]. For single-encoder designs [12,15,19–21], the central research challenge is how to obtain semantically rich representations while retaining sufficient visual fidelity. Prior work can be further organized into two lines: (i) enhancing semantics, which improves semantic content through tailored objectives, training strategies, or scaling; and (ii) preserving semantics, which leverages pretrained vision foundation models to inject or maintain strong semantic priors in the token space. In dual-encoder designs [13,14,18,22,23], the primary focus is on how to fuse complementary features from two encoders. Based on encoder organization, we further divide these

 doi:10.20944/preprints202607.1756.v1

methods into parallel and cascade architectures, depending on whether the two feature streams are integrated symmetrically or sequentially. Additionally, we summarize the performance of unified tokenizers [13,15,17] on downstream understanding and generation tasks, and compare them against (1) single task-specific tokenizers [24–27], and (2) unified systems that rely on separate task-specific tokenizers for different objectives [9–11].

Finally, we discuss several open challenges and future directions for visual tokenization. In particular, we examine the trade-off between semantic density and reconstruction quality, the efficient scaling of token vocabularies, and extensions to video and 3D modalities. We also discuss the evolving role of visual tokens in next-generation multimodal systems. By organizing this rapidly growing field within a unified framework, this work aims to serve as a useful reference for researchers and practitioners in computer vision and multimodal learning.

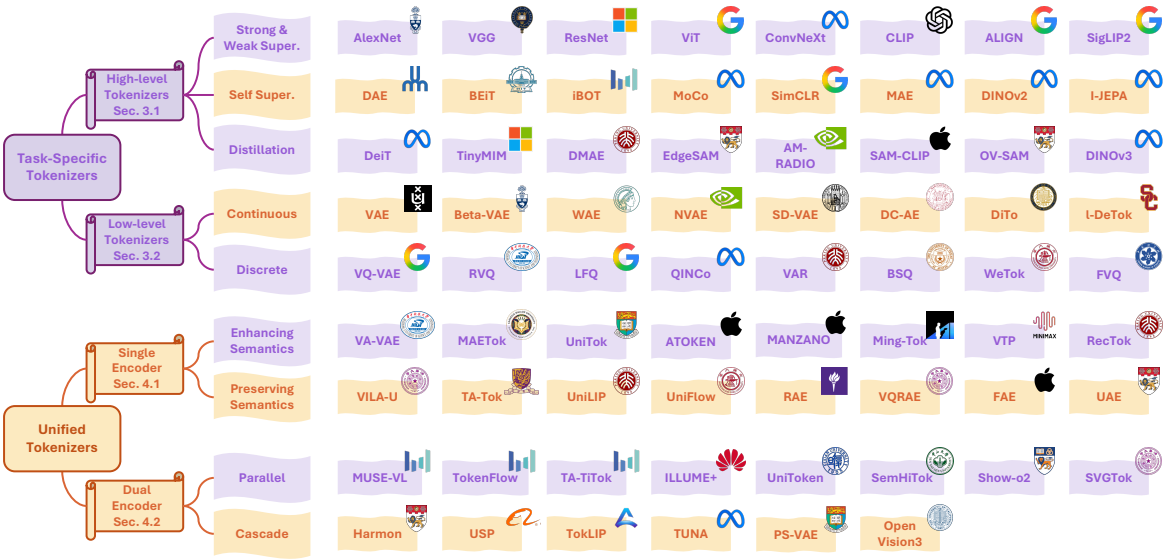

Figure 1. We categorize visual tokenizers by downstream usage into task-specific and unified tokenizers. Task-specific tokenizers target high-level tasks (e.g., classification, dense prediction) or low-level tasks (e.g., reconstruction, dehazing). Unified tokenizers, typically designed for multimodal LLMs, support both within a single representation. Architecturally, we group them into two families and distinguish them by how they capture semantics and fine-grained details.

2. Definition and Scope

A central lesson from large language models [1,28] is that *sequence-to-sequence modeling* [29] provides a universal computational paradigm. Transformers show that, with suitable tokenization, a single model class can support representation learning [8], reasoning [30], and generation [31] across tasks. This insight has influenced the design of visual representations. In this context, *tokenization* should be distinguished from classical *feature extraction*. A feature extractor (e.g., a convolutional backbone [32]) typically produces a structured feature field that is dense, grid-aligned, and closely coupled to the input domain. In contrast, tokenization [33] is a *structural transformation* that converts raw signals into discrete or semi-discrete units processed sequentially by a Transformer [29]. Conceptually, tokenization serves two primary roles:

- **Sequentialization**: imposing an ordering, indexing, or traversal over the input domain.
- **Transformation**: mapping raw inputs into latent representations suitable for attention-based processing.

While these roles are often intertwined in practice, distinguishing them clarifies the assumptions underlying a given visual tokenizer. This perspective raises a fundamental question:

What entities are treated as tokens, and what structural or group actions do these tokens admit?

 doi:10.20944/preprints202607.1756.v1

Different answers lead to different tokenization strategies, even when the downstream architecture remains unchanged. Visual tokenizers can be broadly characterized along two axes:

1. **Input primitives**: what constitutes a token (e.g., pixels, patches, points, voxels, simplices, or feature groups).
2. **Sequentialization mechanism**: how these primitives are ordered or indexed (e.g., grid order, positional encodings, curve traversal, or set-to-sequence projections).

A visual tokenizer maps a visual signal into a set of tokens together with an indexing structure that enables Transformer-based processing. Formally, given a visual input $\mathbf{x} \in \mathbb{R}^{T\times H\times W\times 3}$, where T denotes the temporal dimension and H, W denote the spatial dimensions (with $T = 1$ corresponding to a single image), a visual tokenizer defines a transformation

$$f_\theta : \mathbf{x} \mapsto \mathbf{z}, \qquad \mathbf{z} \in \mathbb{R}^{L\times D},$$

where L denotes the number of tokens and D the token dimension. In the discrete case, one may further have $\mathbf{z} \in \mathbb{N}^L$, corresponding to a sequence of codebook indices or symbolic token identities. Under this formulation, a tokenizer is defined not only by how it transforms raw visual inputs into latent units, but also by the choice of underlying primitives and the mechanism used to organize them into a sequence or indexable set. The resulting token representation $\mathbf{z}$ may therefore originate from pixels, patches, points, voxels, simplices, or higher-level feature groups, together with an ordering, positional scheme, or set-to-sequence projection that governs how attention operates over them.

2.1. Architecture

As shown in Figure 2. In numerous learned instantiations, this mapping is realized through an encoder, combined with a bottleneck, and paired with a decoder for different training supervision.

2.1.1. Encoder

1) Images. Images reside on a regular two-dimensional lattice, which makes tokenization comparatively straightforward. The dominant approach is *patchification*, where an image is partitioned into fixed-size patches and each patch is treated as a token. Sequentialization is achieved through positional encodings, including absolute embeddings, 2D sinusoidal encodings, or relative schemes such as rotary positional embeddings (RoPE).

From this perspective, convolutional neural networks can be interpreted as a special case of tokenization: local, translation-equivariant filters transform the image while the grid structure implicitly provides positional indexing. Vision Transformers make this separation explicit by projecting patches into a latent space, adding positional encodings, and applying global attention.

2) Video. Video extends image tokenization along the temporal dimension [34,35]. Spatial patches are extracted from each frame and stacked across time, producing tokens indexed by (x, y, t). Positional encodings capture both spatial and temporal structure, often through factorized or hierarchical designs. Despite the increased dimensionality, video still benefits from the rasterized grid, which provides a natural traversal of the signal domain.

3) 3D Data. Three-dimensional data departs from this setting. Unlike images or videos, 3D data does not possess a natural uniform grid. The ambient space is large, while meaningful information is typically sparse and irregularly distributed, as observed in point clouds, surfaces, or volumetric representations.

Three-dimensional representation learning has followed two primary paradigms. **Geometric (primitive-based) modeling** treats primitives such as points as tokens and processes them using permutation-invariant or equivariant encoders (e.g., PointNet-style models). **Voxelization-based modeling** discretizes space into a 3D grid, enabling 3D CNNs or sparse convolutions, albeit with memory and resolution trade-offs.

With the adoption of Transformers, recent approaches increasingly emphasize explicit sequence-based tokenization of 3D data. Examples include space traversal via one-dimensional curves, direct

 doi:10.20944/preprints202607.1756.v1

traversal of simplicial or mesh structures, and feature-set or set-to-sequence representations. A key distinction from image and video models is that 3D encoders typically require an *explicit compression stage*. Exhaustively traversing the ambient 3D space using (x, y, z) indices is computationally infeasible. Instead, redundancy and sparsity must be exploited to discard or aggregate information *before* applying a Transformer. This requirement reflects the mismatch between high-dimensional ambient spaces and the low-dimensional supports on which meaningful 3D signals often reside. Unlike grid-based modalities, there is no canonical one-dimensional traversal of 3D data. Any sequentialization therefore imposes additional structure and inductive bias, implicitly defining notions of locality, equivalence, and invariance. Consequently, different tokenization strategies correspond to different structural hypotheses about the data.

Although many of these techniques were originally developed for 3D shape modeling, the principles extend beyond geometric data. Any modality with high ambient dimensionality and sparsely distributed informative signals faces similar challenges. In such settings, tokenization is not merely an implementation detail but a central modeling decision that determines what structure is preserved, compressed, or treated as equivalent.

Visual tokenization thus reflects a shift from dense, domain-specific feature fields toward structured sequences of interacting units. Inspired by large language models, this paradigm elevates sequentialization and transformation to first-class design choices. While images and videos admit relatively canonical tokenizations due to their grid structure, 3D data highlights the deeper challenges—and opportunities—of tokenization in sparse, irregular, high-dimensional domains. In this sense, a visual tokenizer encodes an explicit structural hypothesis about perception itself.

Overall, visual tokenization shifts representation learning from dense, domain-specific feature fields toward structured collections of interacting units. For images and video, regular grids provide relatively canonical tokenization schemes. In contrast, for 3D data, token formation and serialization themselves become central modeling choices.

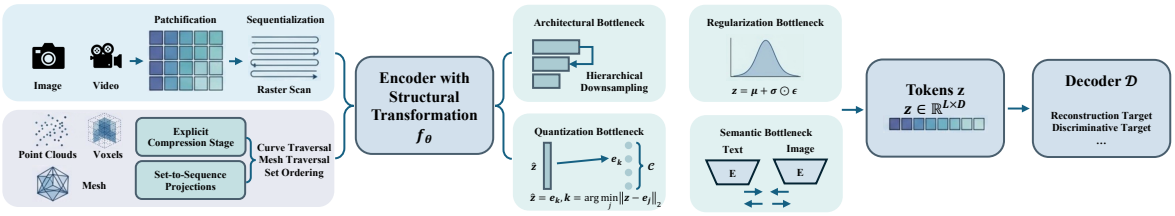

Figure 2. Overview of the visual tokenizer architecture. We conceptualize the architecture of a visual tokenizer as comprising three principal components: an encoder, an information bottleneck, and a decoder. The figure illustrates the end-to-end tokenization pipeline, proceeding from the initial preprocessing of visual data (i.e., patchification and sequentialization) to the extraction of latent tokens $\mathbf{z}$. The decoder is subsequently employed either for detokenization or to provide auxiliary supervision during training, such as via discriminative targets.

2.1.2. Bottleneck

Beyond token formation, a visual tokenizer is also defined by the bottleneck imposed on its token representation. This bottleneck determines which information is preserved, compressed, discarded, or aligned, and thus shapes the resulting token space. Depending on the model, the bottleneck may arise from architectural compression, continuous regularization, discrete quantization, or semantic supervision.

1) Architectural bottleneck. In many encoders, the bottleneck arises from the architecture through downsampling, pooling, strided convolutions, or hierarchical aggregation. Backbones such as ResNets [8] follow this design: the visual signal is progressively compressed into lower-resolution feature maps that can be viewed as continuous visual tokens. Although no explicit latent regularization is imposed, the architecture constrains the representation by limiting spatial resolution and promoting abstraction.

 doi:10.20944/preprints202607.1756.v1

2) Continuous bottleneck. A stronger bottleneck can be imposed by explicitly regularizing the latent space. In probabilistic models such as Variational Autoencoders (VAEs) [6], the latent variable is modeled as a continuous distribution $p(\mathbf{z} \mid \mathbf{x})$, typically parameterized by mean μ and variance σ^2. Sampling is made differentiable via the reparameterization trick,

$$\mathbf{z} = \mu + \sigma \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I}).$$

This encourages a smooth and structured latent space suitable for interpolation and generative modeling, though it may reduce reconstruction fidelity.

3) Discrete bottleneck. In discrete tokenizers [7], the encoder output $\hat{\mathbf{z}}$ is mapped to the nearest entry in a learnable codebook $\mathcal{C} = \{\mathbf{e}_k\}_{k=1}^{K}$:

$$\mathbf{z}_q = \mathbf{e}_k, \qquad k = \arg\min_j \|\hat{\mathbf{z}} - \mathbf{e}_j\|_2.$$

This quantization converts continuous features into indexable symbols, producing token sequences compatible with autoregressive or masked token modeling.

4) Semantic bottleneck. In vision-language models such as CLIP [33] and SigLIP [36], the bottleneck arises from semantic alignment rather than reconstruction or discretization. The encoder retains information predictive of language supervision while discarding nuisance variation, shaping the representation toward semantic alignment.

2.1.3. Decoder

The decoder maps token representations to the target space defined by the training objective. This space may correspond to the input domain, a task-specific label space, or a shared semantic embedding space. Thus, the decoder determines how token representations are read out and what information they preserve.

In reconstruction-based tokenizers, such as VAEs [6] and related compression models [37], the decoder $\mathcal{D}(\cdot)$ reconstructs the input $\mathbf{x}$ from $\mathbf{z}$, or its quantized form $\mathbf{z}_q$, such that $\hat{\mathbf{x}} = \mathcal{D}(\mathbf{z}) \approx \mathbf{x}$. These decoders are typically implemented with upsampling layers, transposed convolutions, or patch-based reconstruction modules, and are evaluated by PSNR, SSIM, and rFID.

In discriminative encoders such as ResNet-style [8] visual backbones, no explicit reconstruction decoder is required; the representation is instead read out through task-specific prediction heads. In contrastive vision-language models such as CLIP [33] and SigLIP [36], the decoder takes the form of a projection head into a shared semantic space aligned with text. Thus, while reconstruction-based models decode signals, discriminative and contrastive tokenizers decode predictions or semantic correspondences. We compare these tokenizers in Sec. 5.

3. Task Specific Visual Tokenizers

Prior to Large Language Models (LLMs) and AI-Generated Content (AIGC), visual tokenizers were mainly studied in visual representation learning and generation tasks. Tokenizer design was largely driven by task-specific supervision rather than a unified multimodal perspective. In this section, we review representative pre-LLM methods and organize them by the information encoded in the tokens: *high-level semantic tokens* and *low-level reconstruction tokens*. We focus on the evolution of technical approaches and design motivations.

3.1. High-Level Tokenizers

High-level tokenizers emphasize semantic abstraction rather than pixel-level fidelity. Before LLM, these works serve as backbone networks. This section reviews methods that strengthen semantic representations through training strategies and data scaling.

 doi:10.20944/preprints202607.1756.v1

3.1.1. Training Strategy

Summary. Training strategies differ by the source of supervision: *(1) Supervised learning* uses human annotations, ranging from clean curated labels (strong) to scalable but noisy web image–text pairs (weak). *(2) Self-supervised learning* replaces external labels with pretext signals mined from the data to learn transferable semantics. *(3) Distillation* uses a strong teacher to supervise a lightweight student tokenizer by mimicking its outputs or features.

Supervised Learning. *1) Strong Supervision:* Strong supervision typically refers to training on large-scale datasets with full ground-truth annotations, especially for image classification. Progress under this paradigm largely follows the evolution of deep architectures. AlexNet [38] first demonstrates GPU-trained CNNs on ImageNet. VGG [39] shows that deeper networks with small 3×3 convolutions improve performance. ResNet [8] introduces residual connections, enabling very deep networks. More recently, Vision Transformer (ViT) [4] treats images as patch sequences and shows that Transformers scale effectively. To incorporate useful inductive biases, Swin Transformer [40] proposes a hierarchical shifted-window design. Finally, ConvNeXt [41] revisits CNNs with ViT-inspired designs (e.g., larger kernels and LayerNorm), showing modernized CNNs remain competitive with Transformers.

2) Weak Supervision: Since dense human labels are costly, recent work trains on large-scale noisy web image–text pairs. CLIP [5] and ALIGN [42] show that contrastive learning on such data yields highly transferable representations. LiT [43] improves efficiency by freezing a strong image encoder while tuning the text encoder. SigLIP [44] replaces softmax with a pairwise sigmoid loss, enabling more stable large-scale training. SigLIP-2 [36] further improves multilingual and dense features via caption-based pretraining and self-distillation, while the Perception Encoder [45] shows that dense semantics often emerge in intermediate layers.

Strong supervision provides clean signals and reliable evaluation but is limited by annotation cost and fixed label spaces. Weak supervision scales cheaply and supports open-vocabulary transfer, though it suffers from noisy captions and dataset biases, making dense or fine-grained performance dependent on data curation and training recipes.

Self-Supervised Learning. While supervised learning scales with web text, it still depends on linguistic signals that may overlook fine-grained visual structure. Self-supervised learning (SSL) instead learns directly from visual data.

1) Predictive-Based: Predictive SSL mirrors masked language modeling [46] by predicting hidden visual content. BEiT [47] adapts BERT-style masking to images using discrete VAE tokens. MAE [48] simplifies this by masking many patches and reconstructing pixels, enabling scalable representation learning. I-JEPA [49] predicts missing regions in latent space rather than pixels. EVA [50] further connects predictive SSL with vision-language pretraining. TIPS [51] combines image–text contrastive learning with masked modeling to improve spatial awareness in frozen vision-language models.

2) Discriminative-Based: Discriminative SSL learns representations by distinguishing views of the same image. MoCo [52,53] uses a momentum encoder and queue for contrastive learning, while SimCLR [54] shows large batches and strong augmentations suffice without memory banks. BYOL [55] removes negatives using asymmetric networks with a predictor. DINO [56] extends this idea via self-distillation, revealing segmentation cues in ViTs. DINOv2 [57] scales this approach with curated data for general-purpose features. DINOv3 [58] further improves dense features with large-scale SSL and Gram Anchoring.

Predictive SSL typically learns spatially detailed features but may emphasize reconstruction and require decoders, whereas discriminative SSL yields more invariant semantics but relies heavily on training design.

Distillation. Beyond training from scratch, high-level tokenizers are often improved via distillation. Two directions dominate: (1) *Efficiency-driven distillation*, compressing foundation models into lightweight students. (2) *Capability-driven agglomeration*, consolidating multiple teachers into a single tokenizer.

 doi:10.20944/preprints202607.1756.v1

1) Efficient Models: Knowledge distillation (KD) [59] transfers representations from large teachers to compact models. DeiT [60] introduces a distillation token for efficient transformers. TinyMIM [61] and DMAE [62] transfer token-level relations for dense understanding. EdgeSAM [63] distills heavy segmentation backbones while preserving interactive performance. DINOv3 [58] further improves dense features using Gram-based teachers and multi-student distillation.

2) Agglomerative Strategies: Different teachers provide complementary capabilities. CLIP-like models capture global semantics, while DINO/SAM provide dense correspondence and structure. Agglomerative distillation combines these strengths into a single tokenizer. AM-RADIO [64] distills heterogeneous teachers (CLIP, DINO, SAM) into one model. RADIOv2.5 [65] addresses resolution shifts, teacher imbalance, and token-budget limits through multi-resolution training and improved loss balancing. C-RADIOv3 [66] further improves robustness via stronger distribution control. C-RADIOv4 [67] upgrades the teacher suite (SigLIP2/DINOv3/SAM3) and strengthens any-resolution modeling. Related work includes SAM-CLIP [68], which merges semantic and spatial models via replay distillation, and Open-Vocabulary SAM [69], which unifies segmentation with large-vocabulary recognition. Beyond perception, Theia [70] distills multiple VFMs into compact representations for robot learning.

Trend 3.1.1 High-level tokenizer training is moving toward unification: instead of separate experts (e.g., CLIP for semantics, DINO for density), recent methods train or distill a single tokenizer supporting both open-world semantics and high-resolution structure.

3.1.2. Data Scaling

Data scaling complements architectural advances: larger and more diverse corpora improve coverage and robustness, while dataset quality and composition determine the semantics encoded in tokens. We summarize progress from two perspectives—*data sources* and *data curation*.

Data Sources. *1) Real Datasets:* Real-world datasets have evolved from curated labeled collections to web-scale image–text corpora, where scale and filtering jointly determine representation quality. Early work relies on benchmarks such as ImageNet [71]. JFT-300M [72] shows that data scale can outweigh architectural changes when paired with effective curation. With weak supervision, open datasets such as LAION-5B [73] popularize CLIP-based filtering for quality control. DataComp-1B [74] formalizes dataset construction as an optimization problem and benchmarks curation strategies under fixed compute. At larger scales, WebLI-100B [75] extends multimodal pretraining to 100B examples. Overall, real datasets have progressed in both scale (up to 100B) and quality through improved filtering and curation.

2) Synthetic Data: Recent work explores synthetic data as a scalable alternative. StableRep [76] shows that diffusion models can generate diverse multi-view images for contrastive pretraining. CapsFusion [77] improves supervision by rewriting captions with LLMs to enrich semantic content. SynCLR [78] studies fully synthetic pipelines that generate both images and captions at scale, demonstrating that synthetic-only corpora can approach real-data performance. These results suggest synthetic data is not merely augmentation but a complementary route to scalable representation learning.

Data Curation. *1) Clustering and Balancing:* At web scale, structuring data becomes critical. Naive sampling over web crawls over-represents common content while under-sampling rare concepts. Clustering-based curation groups examples into semantic buckets (often using CLIP embeddings) and balances sampling across clusters. MetaCLIP [79] shows that metadata clustering improves coverage and reduces bias without dense annotation. Similar clustering-based selection also improves self-supervised learning under fixed compute [80].

2) Sampling Strategies: Beyond filtering, batch sampling also affects optimization. The *two-set training* strategy [81] mixes a frequently repeated subset with a stream of new samples to accelerate learning. In vision, DINOv3 [58] adopts mixed sampling by combining *homogeneous* ImageNet-1K

 doi:10.20944/preprints202607.1756.v1

batches for stable features with *heterogeneous* batches from larger datasets, improving stability while preserving generalization.

Trend 3.1.2 At web scale, data handling matters as much as volume: filtering, clustering, sampling, and synthesis are essential for improving performance.

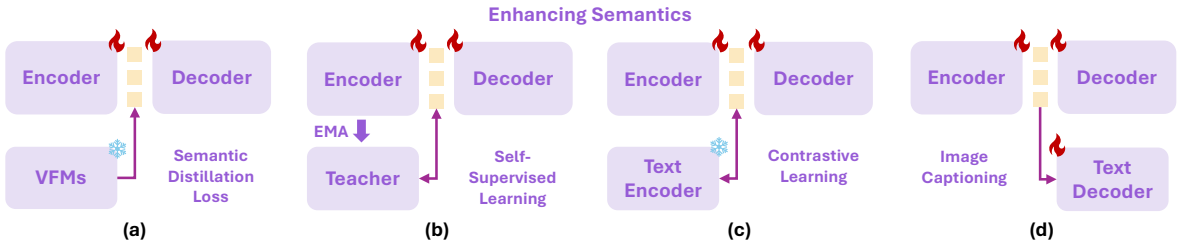

Figure 3. Overview of single-encoder unified tokenizers that enhance visual semantics: (a) distilling knowledge from vision foundation models (VFMs); (b) masked image modeling with EMA teacher alignment; (c) contrastive alignment between image and text encoders; and (d) image captioning as an auxiliary task for vision–language alignment.

3.2. Low-Level Tokenizers

Unlike semantic tokenization, which extracts high-level representations, low-level tokenization compresses visual signals into compact features while preserving details for reconstruction. Although some low-level tokenizers [6,82] are generative models, we focus here on tokenization rather than generation performance. We organize this section along three dimensions: (1) reconstruction fidelity, (2) representation and compression efficiency, and (3) representation disentanglement.

3.2.1. Maximizing Reconstruction Fidelity of Continuous Low-Level Tokenizers

Standard autoencoders can achieve low reconstruction error, but their unconstrained feature spaces often exhibit high variance and discontinuities, making them unsuitable for tokenization. Variational Autoencoders (VAEs) [6] address this by introducing Kullback–Leibler (KL) regularization to enforce a smooth latent manifold:

$$\mathcal{L}_{\text{VAE}} = \underbrace{\mathbb{E}_{q_\phi(z|x)}[-\log p_\theta(x|z)]}_{\mathcal{L}_{\text{rec}}} + \underbrace{\beta D_{\text{KL}}\big(q_\phi(z|x) \parallel p(z)\big)}_{\mathcal{L}_{\text{reg}}}, \tag{1}$$

where q_ϕ and p_θ denote the encoder and decoder. $\mathcal{L}_{\text{rec}}$ (typically MSE) preserves data fidelity, while $\mathcal{L}_{\text{reg}}$ regularizes the posterior toward the prior $p(z)$.

Despite this regularization, vanilla VAEs still exhibit lower reconstruction fidelity than standard autoencoders due to several limitations. First, **posterior collapse** occurs when $q_\phi(z|x)$ degenerates to the prior $p(z)$ (e.g., $\mathcal{N}(0, I)$), weakening the dependence between latent codes and inputs and causing loss of fine visual details. Second, the **simple Gaussian prior** forces complex visual distributions into an isotropic form, often producing overly smooth reconstructions with blurred textures. Third, the **information bottleneck** arises because reconstruction depends on low-dimensional latent codes, limiting achievable fidelity. Finally, **simple pixel-wise losses** (e.g., MSE) poorly capture perceptual similarity and local structure, further degrading reconstruction quality.

Addressing Posterior Collapse. Posterior collapse occurs when strong KL regularization forces $q_\phi(z|x)$ to match the prior, causing the decoder to ignore latent codes.

Wasserstein Autoencoders (WAE) [83] mitigate this by matching the *aggregated posterior* $q_\phi(z) = \mathbb{E}_{p_{\text{data}}(x)}[q_\phi(z|x)]$ to the prior:

$$\mathcal{L}_{\text{WAE}} = \underbrace{\mathbb{E}_{p_{\text{data}}(x)}\mathbb{E}_{q_\phi(z|x)}[c(x, p_\theta(z))]}_{\mathcal{L}_{\text{rec}}} + \underbrace{\beta D_Z(q_\phi(z), p(z))}_{\mathcal{L}_{\text{reg}}}. \tag{2}$$

 doi:10.20944/preprints202607.1756.v1

This relaxes per-sample constraints and better preserves high-frequency details. Sliced-WAE [84] further improves stability by projecting latent distributions into 1D and computing Wasserstein distance in closed form.

Another strategy weakens KL regularization directly. Latent Diffusion Models (LDM) [31] assign an extremely small KL weight (e.g., $\beta = 10^{-6}$), preventing over-regularization while maintaining high reconstruction quality.

A more radical solution is discrete tokenization. VQ-VAE [7] replaces the Gaussian prior with a learnable codebook $\mathcal{C} = \{e_k\}_{k=1}^{K}$:

$$z_q(x) = e_k, \quad k = \arg\min_j \|z_e(x) - e_j\|_2. \tag{3}$$

This discrete bottleneck eliminates KL collapse and enables high-fidelity modeling with autoregressive or diffusion priors [85–87].

Trend 3.2.1 Part 1 Posterior collapse is mitigated by relaxing KL constraints or replacing Gaussian priors with discrete codebooks.

Enriching Prior Distributions. The Gaussian prior often poorly matches the complex manifold of natural images. Several approaches therefore adopt more expressive priors.

Adversarial Autoencoders (AAE) [82,88] replace KL regularization with adversarial matching between the aggregated posterior and prior:

$$\min_G \max_D \mathbb{E}_{z\sim p(z)}[\log D(z)] + \mathbb{E}_{x\sim p_{\text{data}}}[\log(1 - D(G(x)))]. \tag{4}$$

IntroVAE [89] introduces self-adversarial training where the encoder acts as the discriminator, while Soft-IntroVAE [90] stabilizes training via ELBO-based energy functions.

Other work adopts explicit learnable priors. GMVAE [91] models $p(z)$ as a Gaussian mixture:

$$p(z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(z|\mu_k, \Sigma_k). \tag{5}$$

VampPrior [92] defines the prior as a mixture of variational posteriors conditioned on pseudo-inputs.

Two-stage methods instead decouple representation learning and prior estimation. VAIOP [93] learns an implicit optimal prior after encoder training, while contrastive priors [94] structure latent spaces via similarity learning. DiTo [95] replaces the prior with diffusion conditioning tokens for reconstruction. UNITE [96] unifies image tokenization and latent denoising through a shared Generative Encoder, enabling end-to-end single-stage training of latent diffusion models without a separately trained tokenizer.

Trend 3.2.1 Part 2 Expressive priors are obtained via adversarial matching, learnable mixtures, or two-stage prior estimation.

Addressing Information Bottleneck. Single-layer VAEs impose a severe information bottleneck. Hierarchical VAEs distribute information across multiple latent scales.

LVAE [97] introduces bottom-up and top-down inference paths with layer-wise KL terms:

$$\begin{aligned}\mathcal{L}_{\text{LVAE}} = \mathbb{E}_{q_\phi(\mathbf{z}|x)}[\log p_\theta(x|\mathbf{z})] \\ - \sum_{l=1}^{L} D_{\text{KL}}(q_\phi(z_l|z_{l-1}, x) \parallel p_\theta(z_l|z_{l+1}))\end{aligned}. \tag{6}$$

BIVA [98] strengthens bidirectional inference, while NVAE [99] scales hierarchical depth with residual parameterization. VDVAE [100] pushes this further with up to 78 stochastic layers.

 doi:10.20944/preprints202607.1756.v1

Another direction expands the encoder's receptive field. Latent Diffusion Models [31] add self-attention to convolutional encoders to capture global context and improve reconstruction.

Trend 3.2.1 Part 3 Solutions evolve from hierarchical latent modeling toward deeper architectures and global-context encoders.

Optimization Strategy. To overcome the limitations of pixel-wise reconstruction objectives, such as mean squared error (MSE), which often lead to over-smoothed and blurry outputs due to global averaging. VQGAN [85] pioneered a more powerful paradigm by introducing patch-based adversarial supervision and perceptual losses (e.g., LPIPS) into the vector-quantization framework. This design constrains reconstructed tokens to lie closer to the manifold of natural images, thereby preserving sharper structures and richer high-frequency textures.

Building on this line of work, SD-VAE [31] further improves reconstruction quality by incorporating attention blocks to capture long-range dependencies across spatial tokens. In addition, it leverages a series of practical engineering techniques to substantially enhance the reconstruction capability of the VAE, yielding significantly more faithful and visually appealing outputs. Subsequent work [101] further improves reconstruction quality by enriching perceptual supervision. In addition to conventional perceptual losses, they explicitly align the ConvNeXt [41] features of the input image x and the reconstructed image $\hat{x}$, which can be formulated as

$$\mathcal{L}_{\text{perc}}^{\text{ConvNeXt}} = \sum_{l} \left\| \phi_l^{\text{ConvNeXt}}(x) - \phi_l^{\text{ConvNeXt}}(\hat{x}) \right\|_2^2, \tag{7}$$

where $\phi_l^{\text{ConvNeXt}}(\cdot)$ denotes the feature representation at the l-th layer of a pretrained ConvNeXt model. Compared with pixel-space supervision, such feature-level alignment provides a stronger semantic and structural constraint, leading to more faithful reconstructions.

Meanwhile, some studies [19,102] replace or initialize the discriminator with stronger foundation models [57], endowing the adversarial branch with a more powerful prior over natural image statistics. Denoting such a discriminator by D_{FM}, the adversarial objective can be written as

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_x \left[\log D_{\text{FM}}(x) \right] + \mathbb{E}_{\hat{x}} \left[\log \left(1 - D_{\text{FM}}(\hat{x}) \right) \right]. \tag{8}$$

By leveraging the representational strength of pretrained foundation models, these approaches provide more informative supervision signals and further enhance perceptual fidelity and texture realism in the reconstructed images.

3.2.2. Maximizing Reconstruction Fidelity of Discrete Low-Level Tokenizers

Unlike continuous low-level tokenizers, discrete low-level tokenizers use quantization to represent visual data with a finite set of code entries. They were originally introduced to address the posterior collapse problem [7]. However, because discrete token representations are highly compatible with LLMs [1], discrete tokenizers have received broad attention, especially in terms of improving reconstruction quality.

As mentioned above, discrete tokenizers represent visual data using a limited number of code entries. Therefore, preserving fine-grained details usually requires a sufficiently large codebook [103]. Below, we summarize how earlier studies have effectively scaled up the codebook size.

Scaling the Codebook of Discrete Tokenizers. For discrete tokenizers,]the discretization paradigm was originally introduced to resolve the posterior collapse issue inherent in continuous VAEs. However, this introduces a new challenge: since the codebook is finite, the encoder is compelled to map diverse visual data to a limited set of discrete entries. This quantization process inevitably incurs information loss, manifesting as quantization error. A straightforward strategy to mitigate this loss is to increase the codebook size K. However, naively expanding K often leads to "codebook collapse," a phenomenon

 doi:10.20944/preprints202607.1756.v1

where the model utilizes only a small fraction of the available codes, rendering the majority of the codebook redundant.

1) Optimization Strategies. To maintain high utilization in large codebooks, VQ-VAE [7] employs Exponential Moving Average (EMA) updates to smooth the embedding learning. More aggressively, VQGAN-LC [103] introduces a frequency-based initialization strategy to revive "dead codes," enabling codebooks to scale effectively to 100,000 entries. SQ-VAE [104] replaces deterministic quantization with stochastic sampling, thereby encouraging the exploration of the codebook space. IBQ [105] facilitates gradient flow into the index selection process, correcting the mismatch typically caused by straight-through estimators. Furthermore, FVQ [106] proposes VQBridge, an efficient projector module inserted before the quantization step. Crucially, VQBridge ensures that even unused code entries receive gradient updates within the projection space, thereby achieving stable and effective codebook training. By combining VQBridge with a learning rate annealing schedule, FVQ achieves full (100%) codebook utilization across diverse configurations.

Trend 3.2.2 Part 1 Strategies for codebook optimization have evolved from focusing on handcrafted methods (e.g. EMA, Revive) to ensuring more effective gradient propagation to every code entry.

2) Addressing Quantization Cost. As the codebook size K increases, the quantization process, which necessitates calculating the distance between the encoder output and every codebook entry, incurs prohibitive computational and memory costs scaling linearly with $O(K)$. To circumvent this bottleneck, recent advancements [107–109] have proposed *implicit codebook quantization* methods that eliminate the need for explicit nearest-neighbor search.

Finite Scalar Quantization (FSQ) projects continuous representations onto a fixed, pre-defined integer grid, thereby removing the requirement for storing and learning explicit embedding vectors.

Lookup-free Quantization (LFQ), notably utilized in MAGVIT-v2 [107], leverages a binary latent token space where code indices are directly derived from the sign bits of each latent token embedding $z \in \mathbb{R}^d$ components:

$$k = \sum_{i=0}^{d-1} \mathbb{I}(z_i > 0) \cdot 2^i. \tag{9}$$

where z_i is the i-th component of z. This mechanism allows the effective vocabulary size to scale exponentially with the latent token dimension (e.g., 2^{18} for $d = 18$) while incurring zero storage and computation cost for the codebook. To encourage uniform usage of the binary codes, MAGVIT-v2 introduces an entropy penalty on the code distribution:

$$\mathcal{L}_{\text{entropy}} = \mathbb{E}_x[H(q(k|x)) - \lambda H(\bar{q}(k))], \tag{10}$$

where the first term minimizes the uncertainty of the code assignment for a given input (ensuring deterministic quantization), and the second term maximizes the entropy of the marginal code distribution $\bar{q}(k)$ to prevent codebook collapse. Binary Spherical Quantization (BSQ) [108] further refines this paradigm by projecting embeddings onto a high-dimensional hypersphere before binarization. By exploiting the geometric properties of the hypersphere, BSQ ensures that the quantization error remains bounded and that the binary codes are more semantically separable compared to standard Euclidean binarization. WeTok [109,110] introduces *Group-wise LFQ* to strike a balance between representation compactness and expressiveness. Instead of treating the entire latent vector as a single binary code, WeTok splits the token channels into multiple groups similar to PQ [111], applies LFQ within each group, and concatenates the resulting indices. This approach enables the model to capture multi-scale semantic information while maintaining the efficiency of lookup-free quantization.

Trend 3.2.2 Part 2 Various quantization methods have been introduced to reduce quantization costs and enhance training stability.

 doi:10.20944/preprints202607.1756.v1

3.2.3. Optimizing Compression Efficiency

To alleviate the computational burden on downstream generative tasks, maximizing the information density per token—thereby increasing the effective compression rate—is essential. We categorize recent efforts into per-token optimization and global compression strategies.

Per-token Perspective. Standard vector quantization often suffers from a trade-off between codebook size and reconstruction quality. Product Quantization (PQ) [111] addresses this by decomposing high-dimensional vectors into lower-dimensional subspaces and quantizing each independently, effectively expanding the representational capacity without increasing the codebook size. Residual Vector Quantization (RVQ) [112,113], widely adopted in models like RQ-VAE, employs a coarse-to-fine strategy. It iteratively quantizes the residual error from the previous step, approximating the latent vector z as a sum of discrete codes:

$$z \approx \sum_{m=1}^{M} e_{k_m}^{(m)}, \quad \text{where } e_{k_m}^{(m)} = \text{Quantize}\left(z - \sum_{j=1}^{m-1} e_{k_j}^{(j)}\right). \tag{11}$$

Complementary to this depth-wise residual approach, VAR [114] introduces a multi-scale quantization strategy tailored for visual autoregressive modeling. Instead of refining the residual at a fixed resolution, VAR quantizes the image feature maps into a multi-scale token pyramid (from 1×1 to $K \times K$). This hierarchical tokenization enables the downstream generative model to predict the "next-scale" token map from coarser scales, effectively decomposing the complex image generation task into a structured, coarse-to-fine prediction process. Additive Quantization (AQ) [115] generalizes this concept by approximating vectors as sums of codewords from multiple codebooks without enforcing a strict sequential order, offering even greater flexibility in representation.

Global Compression Perspective. While per-token methods optimize the information content of individual codes, global strategies focus on reducing the total number of tokens required to represent an image. DC-AE [37,116] targets the computational bottleneck in high-resolution synthesis by scaling the spatial downsampling factor to extreme levels ($f \in \{32, 64\}$). To mitigate the severe information loss associated with such aggressive compression, it introduces a Residual Autoencoding mechanism and Decoupled High-Resolution Adaptation, achieving significant inference acceleration while preserving visual fidelity. Taking a different approach, TiTok [101] departs from the traditional grid-based tokenization. It adopts a Transformer-based framework (specifically, a Q-Former architecture [117]), in which a set of learnable latent queries interacts with image features via cross-attention. This mechanism decouples the number of tokens from the image resolution, allowing the model to compress visual data into a compact, fixed-length sequence of 1D tokens, thereby facilitating highly efficient generation.

Trend 3.2.3 Approaches to enhancing compression efficiency include explicitly improving the expressive capacity of individual tokens and strengthening global information compression.

3.2.4. Enforcing Disentanglement

Information-Theoretic Disentanglement. To enhance the interpretability of token representations, β-VAE [118,119] introduces a hyperparameter β to modulate the information bottleneck. The objective is formulated as:

$$\mathcal{L}_{\beta\text{-VAE}} = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \beta\,\text{KL}\big(q_\phi(z|x) \,\|\, p(z)\big). \tag{12}$$

While $\beta > 1$ promotes disentanglement, it often necessitates a trade-off with reconstruction quality. To address this, ControlVAE [120] formulates KL regularization as a feedback control problem, dynamically adjusting β to maintain a target information capacity. Similarly, DynamicVAE [121] employs an adaptive schedule to balance latent utilization and reconstruction throughout training.

 doi:10.20944/preprints202607.1756.v1

However, penalizing the entire KL term indiscriminately suppresses useful mutual information. FactorVAE [122] and β-TCVAE [123] refine this by decomposing the KL term:

$$\mathbb{E}_{p(x)}[\mathrm{KL}(q(z|x)\|p(z))] = \underbrace{I(x;z)}_{\text{Mutual Info.}} + \underbrace{\mathrm{KL}(q(z)\|\prod_j q(z_j))}_{\text{Total Correlation}} + \underbrace{\sum_j \mathrm{KL}(q(z_j)\|p(z_j))}_{\text{Dimension-wise KL}}. \tag{13}$$

By specifically penalizing the Total Correlation (TC) term, these methods enforce statistical independence among latent dimensions without compromising the information content required for reconstruction. Alternatively, DIP-VAE achieves disentanglement via moment matching, constraining the aggregated posterior's covariance to be diagonal.

Trend 3.2.4 Approaches to enforcing disentanglement include modulating the information bottleneck parameter to balance disentanglement with reconstruction quality, and refining constraints by decomposing the KL divergence (e.g., penalizing Total Correlation) to specifically target latent independence.

4. Towards Unified Vision Tokenization

The advent of Large Language Models (LLMs) and AI-Generated Content (AIGC) has reshaped the requirements for visual tokenization, as shown in Figure 5. Traditional tokenizers emphasized either reconstruction fidelity (e.g., VQGAN) or discriminative semantics (e.g., CLIP), whereas modern systems require a *unified tokenizer*. Such tokenizers must simultaneously encode strong semantics for multimodal understanding and fine-grained details for high-fidelity generation. This section reviews recent efforts to bridge semantic understanding and pixel-level reconstruction.

4.1. Single Encoder

4.1.1. Enhancing Semantic Information

As illustrated in Figure 3, this paradigm begins with a reconstruction objective and introduces additional training signals to enrich semantic representations, improving both the reconstructability and semantic density of visual tokens.

Semantic Distillation. As shown in Figure 3(a), a common strategy distills knowledge from pretrained vision foundation models (VFMs) [5,36,57] into the tokenizer encoder during reconstruction training. This encourages representations that are both reconstructive and semantically meaningful.

Inspired by REPA [124], VA-VAE [125] introduces semantic distillation from VFMs into visual tokenizers, improving generation; AVFM [126] further shows these tokens are more discriminative than conventional VAEs [6]. MAETok [127] learns richer semantics via masked image modeling while compressing images into 1D tokens, and GigaTok [128] scales the tokenizer to 3B parameters with semantic distillation for stronger autoregressive generation. Ming-UniVision [12] advances unified tokenization by adding masked feature prediction and a causal semantic decoder, enabling a single tokenizer for both understanding and generation and outperforming dual-tokenizer designs (e.g., VAE+CLIP). SSDD [129]'s key design is a REPA-inspired pixel diffusion decoder that leverages transformer components and distillation to deliver GAN-free, single-step image tokenization with stronger reconstruction quality and faster sampling. RecTok [21] studies the reconstruction–semantics trade-off and proposes flow-based distillation to scale token dimensionality while improving reconstruction, generation, and discrimination.

Language Supervision. As illustrated in Figure 3(c)-(d), another direction aligns visual tokens with text through contrastive learning or caption supervision. ImageFolder [130] uses product quantization [111] to split tokens into low-level and high-level codes, supervising high-level codes with contrastive alignment to text. UniTok [131] adopts multi-codebook quantization and contrastive learning to

 doi:10.20944/preprints202607.1756.v1

support both generation and understanding. AToken [132] improves semantic quality using contrastive and image-text distillation losses while removing GAN supervision; it also extends unified tokenization to video and 3D. MANZANO [133] introduces image captioning as an auxiliary objective, using a single encoder with lightweight adapters and a caption decoder conditioned on visual tokens.

Table 1. Comprehensive Survey of Unified Tokenizers (Part 1/2). This section covers Single Encoder methods (Enhancing and Preserving Semantics). We list the datasets used for tokenizer training, the tokenizer architectures, and the pretrained models leveraged by each method, and we summarize the key highlights of each approach.

Tokenizer	Type	Data	Arch.	Pretrained Model		Highlight	Date
				High-level	Low-level		
Single Encoder (Enhancing Semantics)							
ImageFolder [130]	C. L.	ImageNet-1K	ViT	DINOv2	-	Decouples code entry into low-level and high-level via product quantization.	2024.10
VQ-KD [134]	Distill.	ImageNet-1K	CNN	CLIP	-	Aligning image tokenizers with semantic features, rather than just pixels, significantly improves autoregressive image generation.	2024.11
VA-VAE [125]	Distill.	ImageNet-1K	CNN	DINOv2	-	Using semantic information to mitigate dimension dilemma.	2025.01
MAETok [127]	Distill.	ImageNet-1K	ViT	DINOv2	-	Leverageing feature reconstruction to learn semantically rich 1D representations.	2025.02
UniTok [131]	C. L.	DataComp-1B	ViT+CNN	CLIP	-	Adopts Multi-Codebook Quantization; aligns text features via contrastive learning.	2025.02
GigaTok [128]	Distill.	ImageNet-1K	ViT+CNN	DINOv2	-	Scaling visual tokenizers to 3 billion parameters.	2025.04
ETT [135]	LM	SA-1B+OpenImages+LAION	CNN	Qwen2.5	IBQ	Jointly optimizing vision tokenizers with downstream tasks boosts multimodal performance.	2025.05
AToken [132]	C. L.	DFN+OpenImages+Internal Data+WebVid+TextVR+Panda70M+Objaverse	ViT	SigLIP2	-	Unified tokenization for image, video, and 3D data using a single tokenizer.	2025.09
MANZANO [133]	LM	CC3M&12M+COYO-700M+VeCap+Internal Data	ViT	CLIP	-	Generates continuous and discrete tokens via separate adapters.	2025.09
Ming-Tok [12]	Distill.	Internal Data	ViT	DINOv2	-	Training a unified tokenizer for autoregressive models via feature reconstruction.	2025.10
VTP [136]	S.L.+C.L.	277M subset of DataComp-1B	ViT	-	-	Utilizes CLIP, DINO, and reconstruction losses; validated scaling on large datasets.	2025.12
RecTok [21]	Distill.	ImageNet-1K	ViT	DINOv3	-	Implements reconstruction and alignment in forward flow to resolve dimensionality issues.	2025.12
GloTok [137]	Distill.	ImageNet-1K	CNN	DINOv2	-	GloTok leverages global relations to unify semantic distribution for superior image generation.	2025.12
PyraTok [138]	Distill.	Droplet-10M+OpenVid-1M+UltraVideo	ViT+CNN	Qwen2.5-VL+DINOv3	Wan2.2	PyraTok aligns pyramidal tokens with language for unified video understanding and generation.	2026.01
Single Encoder (Preserving Semantics)							
EMU2 [139]	Frozen	LAION-COCO+LAION-Aesthetics	ViT+CNN	CLIP	-	Uses frozen CLIP as tokenizer; employs trainable SDXL as pixel decoder.	2024.05
VILA-U [140]	C. L.	COYO-700M	ViT	SigLIP	-	Adapts to reconstruction by fine-tuning SigLIP; preserves semantics via contrastive loss.	2024.09
Divot [141]	Frozen	WebVid-10M+Panda-70M	ViT	ViT-H	DynamiCrafter	Utilizing diffusion to power video tokenizer for comprehension and generation.	2024.12
QLIP [142]	C. L.	DataComp-1B	ViT	CLIP	-	Two-stage: contrastive loss first, then freezes encoder to preserve semantics.	2025.02
DualToken [143]	Self-Distill.	CC12M	ViT	SigLIP	-	Uses SigLIP shallow features for recon, deep for semantics; aligns via concatenation.	2025.03
TA-Tok [144]	Self-Distill.	LAION	ViT	SigLIP2	-	Initializes codebook with LLM embeddings to align visual tokens with text features.	2025.07
VFMTok [145]	Frozen	ImageNet-1K	ViT	DINOv2	-	Enhances reconstruction with multi-level features; captures detail via deformable transformer.	2025.07
UniLIP [15]	Self-Distill.	BLIP3-o pretraining data 27M	ViT+CNN	InternViT	-	Two-stage training: stage one trains decoder only, stage two implies self-distillation loss.	2025.07
AVFM [126]	Self-Distill.	ImageNet-1K+LAION	ViT+CNN	DINOv2	-	Adapts DINOv2 to generation tasks via three-stage training; validated on LAION.	2025.09
UniFlow [146]	Self-Distill.	ImageNet-1K	ViT	InternViT	-	Performs self-distillation on multi-level features; uses flow matching pixel decoder.	2025.10
RAE [102]	Frozen	ImageNet-1K	ViT	DINOv2	-	SOTA generative performance with frozen DINOv2; realizes training in high-dim space.	2025.10
VFM-VAE [147]	Frozen	ImageNet-1K	ViT+CNN	SigLIP2	-	Obtains visual tokens via multi-level feature fusion; reconstructs using global/local features.	2025.11
DINO-Tok [148]	Frozen	ImageNet-1K	ViT	DINOv2	-	Explicitly decouples texture and semantic features; concatenates both in feature dimension.	2025.11
VQRAE [149]	Self-Distill.	BLIP3-o pretraining data 27M	ViT	InternViT	-	Uses continuous representations for understanding; post-quantization for generation.	2025.11
VGT [150]	Self-Distill.	BLIP3-o pretraining data 27M+CC12M+JourneyDB+200K HQ	ViT+CNN	QwenViT	DCAE	Empowers VLMs with efficient visual generation by aligning semantic encoders with pixel decoders.	2025.11
FAE [151]	Frozen	ImageNet-1K+CC12M	ViT	DINOv2	-	Trains an asymmetric VAE in high-dim semantic space; retains semantics while reducing dims.	2025.12
UAE [20]	Self-Distill.	ImageNet-1K	ViT	DINOv2	-	Freq domain: preserves low-freq high-level, amplifies high-freq low-level info.	2025.12
DINO-SAE [152]	Self-Distill.	ImageNet-1K	ViT+CNN	DINOv3	-	Bridges semantic and pixel reconstruction on spherical manifolds for high-fidelity generation.	2026.01

Visual Self-Supervised Learning. As illustrated in Figure 3(b), VTP [136] incorporates DINOv2-style self-supervised learning [57] to strengthen semantic representations. Unlike prior approaches that rely on pretrained text encoders, VTP trains the text encoder jointly and demonstrates strong scalability across tasks such as generation and classification.

 doi:10.20944/preprints202607.1756.v1

Trend 4.1.1 Early single-encoder tokenizers relied heavily on distillation from vision foundation models to enhance semantics, mainly targeting generative performance. Recent work increasingly explores self-supervised learning to improve scalability and reduce dependence on pretrained VFMs.

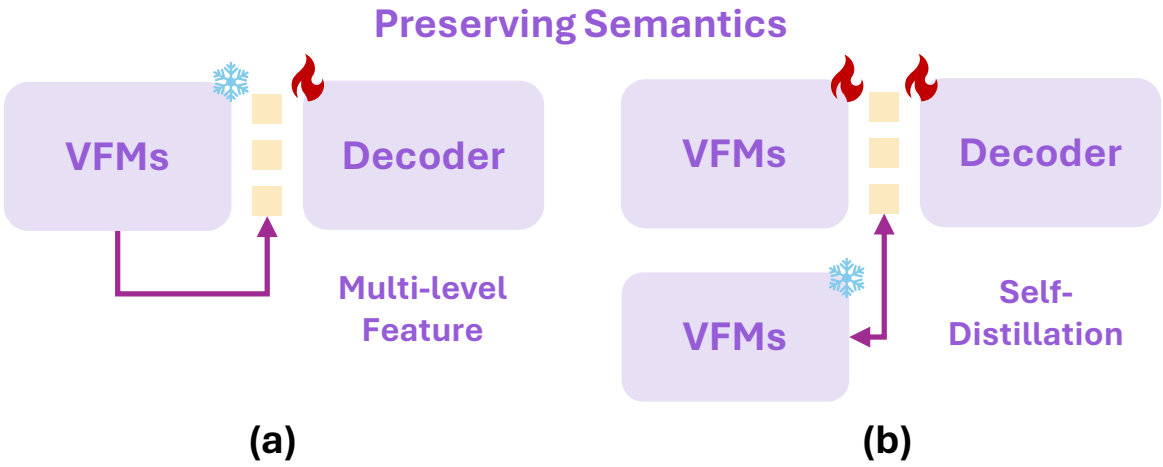

Figure 4. Overview of single-encoder unified tokenizers that preserve visual details: (a) freezing VFMs and reconstructing pixels from last-layer or multi-level features, typically requiring a heavy decoder; and (b) fine-tuning VFMs with self-distillation to retain pretrained semantics.

4.1.2. Preserving Semantic Information

This direction starts from strong high-level tokenizers—typically vision foundation models (VFMs)—and converts them into unified tokenizers by adding a pixel-level decoder. The key challenge is preserving semantic information while enabling high-fidelity reconstruction. As illustrated in Figure 4, existing approaches fall into two streams [13,14,18]: (1) freezing VFMs to preserve semantics while relying on powerful decoders or multi-level features for reconstruction, and (2) fine-tuning VFMs to improve reconstruction while preventing semantic degradation.

Frozen Vision Foundation Models. To preserve pretrained semantics, some methods freeze the VFM encoder and shift modeling capacity to the decoder. Emu2 [139] uses a frozen CLIP encoder and an SDXL U-Net decoder for reconstruction, but reconstruction fidelity is limited because CLIP features lack low-level details.

RAE [102] and FAE [151] show that powerful decoders (e.g., deep Vision Transformers) can reconstruct images from highly compressed semantic tokens, although pixel-level fidelity (PSNR/SSIM) still lags behind low-level tokenizers [6,85].

Other works exploit intermediate features in VFMs, which retain more visual detail than final-layer representations. VFMTok [145], DINO-Tok [148], and VFM-VAE [147] therefore reconstruct images from multi-level features while preserving semantic representations.

Finetuning Vision Foundation Models. While freezing preserves semantics, it limits reconstruction quality [15,146]. Another direction fine-tunes VFM encoders with careful regularization to avoid semantic drift. One strategy retains contrastive objectives during fine-tuning. VILA-U [140] and QLIP [142] combine reconstruction losses with image–text contrastive learning, maintaining language alignment while enabling reconstruction.

 doi:10.20944/preprints202607.1756.v1

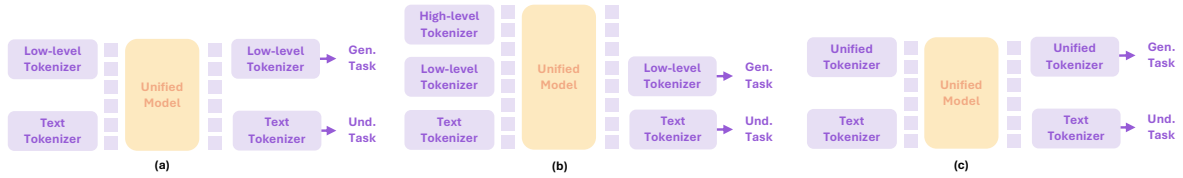

Figure 5. Overview of unified model architectures. (a) Early-stage models [24,27] utilize a single, low-level tokenizer for both understanding and generation tasks. However, these models often underperform on understanding benchmarks because their visual features lack rich semantic information. (b) Recent approaches [9–11] introduce two separate tokenizers tailored to different downstream tasks. Unfortunately, the discrepancy between these distinct feature spaces slows convergence and increases computational overhead [12,13], particularly for complex tasks like image editing [11]. (c) Unified tokenizers [12,15] overcome these limitations by enabling the model to employ a single, versatile tokenizer capable of simultaneously supporting both understanding and generation.

Table 2. Comprehensive Survey of Unified Tokenizers (Part 2/2). This section covers Dual Encoder methods (Parallel and Cascade Encoders).

Tokenizer	Type	Data	Arch.	Pretrained Model		Highlight	Date
				High-level	Low-level		
Dual Encoder (Parallel Encoders)							
MUSE-VL [22]	Parallel	ImageNet-1K+CC12M	ViT+CNN	SigLIP	-	Concatenates high-level and low-level features, followed by quantization.	2024.11
TokenFlow [18]	Parallel	LAION+COYO-700M	ViT	SigLIP	-	Shared mapping strategy; considers both high-level and low-level info during quantization.	2024.12
TA-TiTok [153]	Parallel	DataComp-1B	ViT	CLIP	-	Receives both image and text features during pixel decoding.	2025.01
ILLUME+ [154]	Parallel	ImageNet-1K	ViT+CNN	QwenViT	-	Fuses features only for pixel reconstruction; requires both for generation/understanding.	2025.04
UniToken [155]	Parallel	ShareGPT4V+LLaVA+ALLaVA	ViT+CNN	SigLIP	VQGAN	Concatenates high-level and low-level features before feeding into LLM.	2025.04
SemHiTok [23]	Parallel	ImageNet-1K+50M subset of COYO-700M	ViT	SigLIP	-	Hierarchical quantization: quantizes semantics first, then selects codebook for pixel features.	2025.03
Show-o2 [16]	Parallel	75M Internal Data	ViT+CNN	SigLIP	Wan2.1	Using spatial-temporal fusion to obtain the unified representation.	2025.06
SVGTok [156]	Parallel	ImageNet-1K	ViT+CNN	DINOv3	-	Introduces residual encoder for low-level info; fuses with DINOv3 features.	2025.10
EMMA [157]	Parallel	LLaVA+LAION+X2I2+OmniEdit+Synthetic	ViT+CNN	SigLIP2	DCAE	Fuses the high-level and low-level tokens through concatenation.	2025.12
Dual Encoder (Cascade Encoders)							
Harmon [17]	Cascade	103M Public Data	ViT+CNN	MAR	SD-VAE	Introducing MAR as the visual tokenizer, harmonize high-level and low-level information.	2025.03
USP [158]	Cascade	ImageNet-1K	ViT+CNN	-	SD-VAE	Performs mask reconstruction learning in the VAE feature space.	2025.03
TokLIP [14]	Cascade	CapsFusion+CC12M+LAION-High Resolution	ViT	SigLIP2	VQGAN	Trains high-level tokenizer in VQGAN feature space; adapts to AR model.	2025.05
TUNA [13]	Cascade	Internal Data	ViT+CNN	SigLIP2	Wan2.2	End-to-end training of high-level tokenizer on VAE.	2025.12
PS-VAE [19]	Cascade	ImageNet-1K	ViT+CNN	DINOv2	-	Trains VAE in high-level feature space to reduce dimensionality.	2025.12
OpenVision3 [159]	Cascade	DataComp-1B	ViT+CNN	-	FLUX.1-dev	Incorporating contrastive and caption loss for semantic learning, while VAE feature reconstruction preserves visual details.	2026.01

Another widely used strategy is self-distillation, where the fine-tuned model is constrained to match a frozen teacher. Representative methods include DualToken [143], TA-Tok [144], UniLIP [15], AVFM [126], UniFlow [146], VQRAE [149], and UAE [20]. DualToken [143] reconstructs from shallow features while distilling semantics from deeper layers. TA-Tok [144] aligns visual tokens with LLM embeddings. UniLIP [15] and AVFM [126] adopt multi-stage training to preserve semantics during reconstruction. UniFlow [146] applies multi-layer self-distillation with a flow-based decoder. UAE [20] decomposes tokens via FFT, preserving low-frequency semantics and reconstructing high-frequency details.

Trend 4.1.2 Semantic preservation methods generally follow two strategies: freezing VFMs with strong decoders, or fine-tuning them with contrastive learning or self-distillation to balance semantics and visual detail.

4.2. Dual Encoder

A single encoder often struggles to balance semantic richness and reconstruction fidelity. Dual-encoder designs therefore decouple high-level and low-level feature extraction and merge them into a unified token space.

 doi:10.20944/preprints202607.1756.v1

4.2.1. Parallel Architectures

As shown in Figure 6(a), parallel architectures use multiple visual encoders to capture low-level details and high-level semantics, then fuse them for multimodal tasks.

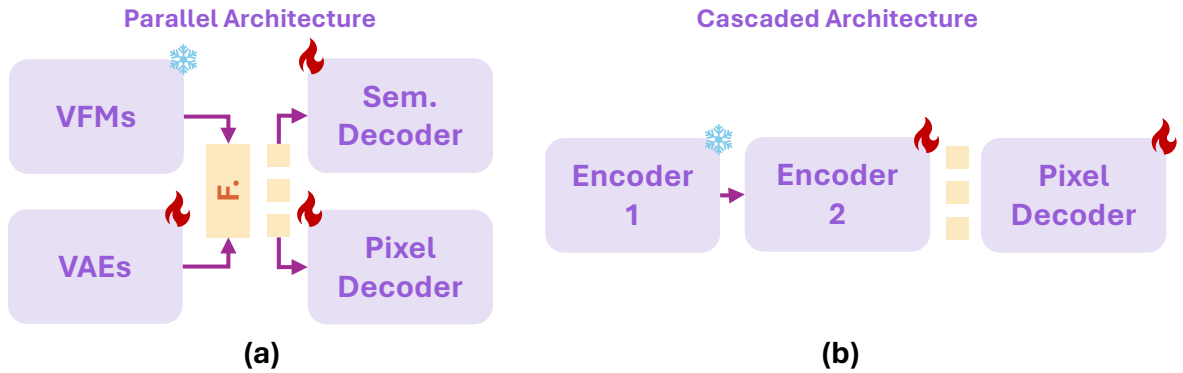

Figure 6. Dual-encoder unified tokenizers: (a) *parallel architectures*, which fuse features from two encoders; and (b) *cascade architectures*, which progressively extract detailed and semantic features.

Several models—Unified-IO 2 [9], Janus [10,160,161], UniFluid [162], and Bagel [11]—adopt two tokenizers, typically a VAE-style encoder for generation and a CLIP-style encoder for understanding, as shown in Figure 5 (b). However, the representations remain disjoint, forcing the backbone (e.g., an LLM) to bridge them, which increases training complexity and slows convergence [12,13]. Tasks such as image editing also require both semantics and detail, motivating unified representations [19].

Recent work focuses on feature fusion. MUSE-VL [22] and UniToken [155] concatenate semantic and pixel features, while TokenFlow [18] introduces weighted fusion. ILLUME+ [154] generates hierarchical tokens, and SemHiTok [23] performs semantic-guided hierarchical quantization. Other methods explore richer fusion strategies. Show-o2 [16] combines SigLIP [36] and VAE features through spatiotemporal fusion. TA-TiTok [153] introduces text-aware tokenization with a frozen text encoder. SVGTok [156] adds a lightweight residual encoder to recover high-frequency details, while EMMA [157] concatenates SigLIP and DCAE [37] features into unified tokens.

Trend 4.2.1 Parallel architectures follow a divide-and-conquer strategy: separate encoders extract semantic and detailed features, which are then fused into unified tokens.

4.2.2. Cascade Architecture

As illustrated in Figure 6(b), cascade architectures adopt progressive tokenization, gradually transforming task-specific features into unified representations.

USP [158] applies masked image modeling on the feature space of a frozen VAE, enabling semantic abstraction while preserving visual details. TokLIP [14] extends this paradigm to autoregressive models by placing a causal representation encoder on top of VQGAN features and aligning them with CLIP semantics. The AR backbone consumes these tokens for both understanding and generation while predicting VQGAN codes for reconstruction. In contrast, TUNA [13] adopts end-to-end training, jointly optimizing the tokenizer for understanding and generation without auxiliary objectives. Experiments show that a trainable unified tokenizer outperforms architectures relying on frozen visual encoders. In [19], PS-VAE introduces an additional autoencoder to compress high-level features, combining reconstruction with self-distillation to preserve semantics. On the other hand, OpenVision3 [159] builds a semantic encoder on top of a pretrained VAE [163], preserving details via VAE feature reconstruction while aligning features with text using contrastive and captioning losses. Overall, cascade architectures emphasize progressive abstraction, explicitly modeling the transition from low-level details to unified representations.

 doi:10.20944/preprints202607.1756.v1

Trend 4.2.2 Cascade architectures adopt progressive tokenization, hierarchically extracting features and introducing supervision in later stages to jointly preserve semantics and visual details.

5. Downstream Tasks and Benchmarks

Evaluating visual tokenizers requires multiple criteria, as different architectures emphasize different capabilities. While low-level tokenizers prioritize signal fidelity, high-level and unified models must also demonstrate semantic proficiency across discriminative and generative tasks. This section summarizes common evaluation protocols and benchmarks.

5.1. High-Level Semantic Evaluation

5.1.1. Global Semantic Capability

Global capability is mainly evaluated on ImageNet classification and cross-modal retrieval.

ImageNet Classification: Table 3 reports results under zero-shot (z.s.), linear probing (l.p.), and fine-tuning (Tune). Weakly supervised models such as SigLIP 2 and the Perception Encoder (PE) achieve strong zero-shot performance (up to 85.4%), while self-supervised models like DINOv3 (7B) reach the best linear probing accuracy of 88.4%.

Table 3. Comparison of High-Level Tokenizers. We categorize evaluation into Global (Classification and Retrieval) and Spatial (Detection and Segmentation) tasks. Metrics include ImageNet Top-1 Accuracy, Recall@1 (R@1) for retrieval, and mAP/mIoU for dense prediction. **In Spatial Tasks**, * is dense linear probing setting, which is much more challenging. **Abbreviations**: *z.s.*: zero-shot (VLM only); *l.p.*: linear probing; *Tune.*: tuned/trained on ImageNet; *b.*: box; *m.*: mask.

Method	Venue	Model Var.	Size	Global Tasks							Spatial Tasks		
				ImageNet			Flickr (R@1)		COCO (R@1)		COCO		ADE
				z.s.	l.p.	Tune.	I2T	T2I	I2T	T2I	b. mAP	m. mAP	mIoU
Strong Supervised Learning													
ViT [4]	ICLR'21	ViT-L/16	307M	-	-	87.8	-	-	-	-	-	-	-
ViT[4]	ICLR'21	ViT-H/14	632M	-	-	88.6	-	-	-	-	-	-	-
Swin [40]	ICCV'21	Swin-L	197M	-	-	87.3	-	-	-	-	58.0	50.4	53.5
ConvNeXt [41]	CVPR'22	ConvNeXt-L	198M	-	-	87.5	-	-	-	-	54.8	47.6	53.7
ConvNeXt [41]	CVPR'22	ConvNeXt-XL	350M	-	-	87.8	-	-	-	-	55.2	47.7	54.0
Weakly Supervised Learning													
CLIP [5]	ICML'21	ViT-L/14	307M	76.2	85.4	-	88.0	68.7	58.4	37.8	-	-	-
ALIGN [42]	ICML'21	EfficientNet-L2	480M	76.4	85.5	-	88.6	75.7	58.6	45.6	-	-	-
LiT [43]	CVPR'22	ViT-g/14	632M	85.2	-	-	-	-	59.3	41.9	-	-	-
SigLIP 2 [36]	arXiv'25	ViT-L/16	307M	83.1	-	-	95.2	85.0	71.4	55.3	-	-	-
SigLIP 2 [36]	arXiv'25	ViT-g/16	632M	85.0	-	-	95.4	86.0	72.8	56.1	-	-	-
PE [45]	arXiv'25	ViT-L/14	307M	83.5	-	-	96.6	85.5	75.9	57.1	-	-	-
PE [45]	arXiv'25	ViT-G/14	1.9B	85.4	-	-	96.2	85.7	75.4	58.1	57.0	49.8	41.5*
Self-Supervised Learning													
MoCov3 [53]	ICCV'21	ViT-L/14	307M	-	77.6	84.1	-	-	-	-	49.3	44.0	49.1
BEiT [47]	ICLR'22	ViT-L/14	307M	-	73.5	86.3	-	-	-	-	53.3	47.1	53.3
MAE [48]	CVPR'22	ViT-L/16	307M	-	75.8	85.9	-	-	-	-	53.3	47.2	53.6
MAE [48]	CVPR'22	ViT-H/14	632M	-	76.6	86.9	-	-	-	-	-	-	-
DINOv3 [58]	arXiv'25	ViT-7B/16	7B	-	88.4	-	-	-	-	-	65.6	-	55.9*
(+ dino.txt) [58]	arXiv'25	ViT-L/14	307M	82.3	-	-	-	-	63.7	45.6	-	-	24.7*
Mixed Training Methods													
EVA [50]	CVPR'23	ViT-1B	1B	78.5	86.5	89.7	-	-	-	-	64.5	55.0	62.3
TIPS [51]	ICLR'25	ViT-1.1B	1.1B	79.7	86.1	-	93.8	83.8	74.0	59.2	-	-	49.4*

Retrieval Performance: On Flickr and COCO [164], models are evaluated via text-to-image (T2I) and image-to-text (I2T) retrieval. SigLIP 2 and PE lead this category, indicating strong cross-modal alignment.

5.1.2. Spatial Dense Capability

Spatial capability measures the ability to capture fine-grained, localized information.

Detection and Instance Segmentation: On COCO [164], models are evaluated by box and mask mAP. Supervised backbones such as Swin-L and ConvNeXt-XL provide strong baselines, while self-

 doi:10.20944/preprints202607.1756.v1

supervised models like EVA (ViT-1B) and DINOv3 (7B) achieve competitive spatial performance, with DINOv3 reaching 65.6 box mAP.

Semantic Segmentation: Semantic segmentation is measured by mean IoU (mIoU) on ADE. A challenging setting is dense linear probing (marked with *), where encoder weights remain frozen. Recent models such as DINOv3 (55.9*), PE (41.5*), and TIPS (49.4*) show that their learned representations support dense prediction without task-specific fine-tuning.

5.1.3. Analysis

The benchmarks show a clear trend: high-level tokenizers are becoming more versatile, supporting both global classification and pixel-level localization. Scaling model size (from ∼300M to 7B parameters) further improves robustness and feature granularity.

5.2. Low-Level Reconstruction Evaluation

For low-level tokenizers, the goal is to compress visual data with minimal information loss while supporting downstream generation. Evaluation is typically conducted on standard datasets such as ImageNet-1K.

Table 4. Comparison of low-level tokenizers on reconstruction fidelity. Evaluated on ImageNet-1K (256×256); * indicates results on ImageNet-1K (128×128). For tokenizers with multiple variants, we report the version used in their main experiments.

Method	Type	ImageNet1K			Usage (%)
		rFID ↓	PSNR ↑	SSIM ↑	
Low-level Tokenzier (Discrete)					
SimVQ [165]	Discrete	1.99*	24.68*	0.803*	100*
VQGAN [85]	Discrete	4.98	20.00	0.629	5.9
LlamaGen [86]	Discrete	2.19	20.79	0.675	97
VQGAN-LC [103]	Discrete	2.62	-	-	99.9
MAGVITv2 [107]	Discrete	1.17	21.90	-	-
Open-MAGVIT2 [166]	Discrete	1.17	22.64	-	100
TiTok [101]	1D-Discrete	2.21	-	-	-
BSQ [108]	Discrete	0.45	28.14	0.814	100
IBQ [105]	Discrete	1.00	-	-	84
WeTok [109]	Discrete	0.19	29.69	-	100
FVQ [106]	Discrete	1.17	-	-	100
Low-level Tokenzier (Continuous)					
DC-AE [37]	Continuous	0.69	23.85	0.660	-
DC-AE-1.5 [116]	Continuous	0.26	-	-	-
SoftVQ [167]	1D-Continuous	0.61	22.97	0.739	-
SD-VAE [31]	Continuous	0.63	26.04	0.834	-
SD3-VAE [168]	Continuous	0.21	31.29	0.886	-
FLUX-VAE [163]	Continuous	0.17	32.86	0.917	-
WAN2.2-VAE	Continuous	0.75	31.25	0.878	-
l-DeTok [169]	Continuous	0.68	-	-	-
Unified Tokenzier (Single Encoder)					
GigaTok [128]	Discrete	0.81	-	-	-
VFMTok [145]	Discrete	1.13	19.91	0.488	100
DINO-Tok [148]	Discrete	1.15	23.98	0.741	-
VQRAE [148]	Discrete	1.39	22.88	0.784	-
VA-VAE [125]	Continuous	0.28	27.96	0.790	-
MAETok [127]	Continuous	0.48	23.61	0.763	-
AVFM [126]	Continuous	0.26	25.83	-	-
ATOKEN [132]	Continuous	0.38	27.14	0.801	-
RAE [102]	Continuous	0.49	-	-	-
FAE [151]	Continuous	0.66	-	-	-
UAE [20]	Continuous	0.19	29.65	0.880	-
RecTok [21]	Continuous	0.48	26.16	-	-
VTP [136]	Continuous	0.36	-	-	-
Unified Tokenzier (Dual Encoder)					
SDE [22]	Discrete	2.26	20.14	0.646	-
TokenFlow [18]	Discrete	1.37	21.41	0.687	-
SVGTok [156]	Continuous	0.65	23.89	0.650	-
PS-VAE [19]	Continuous	0.20	28.79	0.817	-

 doi:10.20944/preprints202607.1756.v1

Table 5. Comparison of tokenizers on class-conditional generation. Evaluated on ImageNet-1K (256×256) [38]. Metrics include gFID, IS, Precision, and Recall. For tokenizers with multiple variants, we report the version used in their main experiments.

Method	Generation Model			ImageNet-1K w/CFG			
	Backbone	Para.	Epochs	gFID ↓	IS ↑	Precision ↑	Recall ↑
Low-level Tokenzier (Discrete)							
VQGAN [85]	Discrete	-	-	5.20	280.3	0.629	-
LlamaGen [86]	LlamaGen-XXL	1.4B	300	3.09	253.6	0.83	0.53
Open-MAGVIT2 [166]	Open-MAGVIT2-AR-XL	1.5B	350	2.33	271.8	0.84	0.54
IBQ [105]	IBQ-XXL	2.1B	450	2.05	286.7	0.83	0.57
WeTok [109]	WeTok-AR-XL	1.5B	1000	2.31	276.6	0.84	0.55
FVQ [106]	LlamaGen-XL	775M	350	2.07	287.0	0.83	0.58
Low-level Tokenzier (Continuous)							
SD-VAE [31]	DDT	675M	800	1.26	310.6	0.79	0.65
l-DeTok [169]	MAR-L	437M	800	1.35	-	-	-
TiTok [101]	MaskGIT	177M	-	2.77	199.8	-	-
SoftVQ [167]	SiT-XL	675M	-	1.86	293.6	-	-
Unified Tokenzier (Single Encoder)							
VA-VAE [125]	LightningDiT-XL	675M	800	1.35	295.3	0.79	0.65
MAETok [127]	SiT-XL	675M	800	1.67	311.2	-	-
GigaTok [128]	LlamaGen-XXL	1.4B	300	1.98	256.76	0.81	0.62
VFMTok [145]	LlamaGen-XXL	1.4B	200	2.16	272.00	0.83	0.60
AVFM [126]	LightningDiT-XL	675M	800	1.37	293.6	0.79	0.65
ATOKEN [132]	LightningDiT-XL	675M	-	1.56	260.0	0.79	0.63
RAE [102]	DiT^{DH}-XL	839M	800	1.13	262.6	0.78	0.67
VFM-VAE [147]	LightningDiT-XL	675M	560	1.57	254.4	0.80	0.64
FAE [151]	LightningDiT-XL	675M	-	1.29	268.0	0.80	0.64
UAE [20]	-	-	-	1.68	301.6	0.77	0.61
VTP [136]	LightningDiT-XL	675M	80	-	-	-	-
RecTok [21]	DiT^{DH}-XL	839M	600	1.13	289.2	0.79	0.67
Unified Tokenzier (Dual Encoder)							
SVGTok [156]	SiT-XL	675M	1400	1.92	264.9	-	-

Table 6. Comparison of unified tokenizers on multimodal generation and understanding. Methods are grouped by tokenizer strategy. For models with multiple variants, we report the best version in their paper.

Method	LLM Backbone	Tokenizer Strategy	Generation			Multimodal Understanding										TextRich	
			GenEval	DPG-Bench	WISE	VQAv2	GQA	POPE	MME	MMB	MMMU	MMV	SEED	AI2D	MathV	MMS	OCRBench	TextVQA
Task-Specific Tokenizer (Single Tokenizer)																		
LWM [170]	LLaMA-2-7B	VQGAN	-	-	-	55.8	44.8	-	-	-	-	-	-	-	-	-	-	-
Transfusion [25]	LLaMA-2-7B	SD-VAE	0.63	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Show-o [27]	Phi-1.5-1.3B	MAGVITv2	0.53	67.27	0.28	69.4	58.0	80.0	1097.2	-	26.7	-	-	-	-	-	-	-
EMU3 [171]	8B from scratch	MoVQGAN	0.66	80.60	-	75.1	60.3	85.2	-	58.5	31.6	37.2	-	70.0	-	-	68.7	64.7
D-DiT [172]	SD3-2B	VAE	0.65	-	-	60.1	59.2	84.0	1124.7	-	-	-	-	-	-	-	-	-
Muddit [24]	Meissonic-1B	VQGAN	0.61	-	-	68.2	57.5	-	1107.4	28.4	27.6	-	-	-	-	-	-	-
MMaDA [26]	LLaDA-8B	MAGVITv2	0.63	69.97	0.67	76.7	61.3	86.1	1410.7	68.5	30.2	-	-	-	-	-	-	-
Lumina-DiMOO [173]	LLaDA-8B	VQGAN	0.88	86.04	-	-	-	87.4	1534.2	84.5	58.6	-	-	-	-	-	-	-
Task-Specific Tokenizer (Seperate Tokenizer)																		
Unified-IO 2 [9]	T5-XXL	VAE + ViT	-	-	-	77.0	-	-	-	-	-	-	-	-	-	-	-	-
Janus [10]	DeepSeek-LLM-1.3B	VQ + SigLIP	0.61	-	-	77.3	59.1	87.0	1338.0	69.4	30.5	34.3	63.7	-	-	-	-	-
JanusFlow [161]	DeepSeek-LLM-1.3B	VQ + SigLIP	0.63	80.09	-	79.8	60.3	88.0	1333.1	74.9	29.3	30.9	70.5	-	-	-	-	-
Janus-Pro [160]	DeepSeek-LLM-7B	VQ + SigLIP	0.80	84.19	-	-	62.0	87.4	1567.1	79.2	41.0	50.0	72.1	-	-	-	-	-
UniFluid [162]	Gemma-2	VAE + CLIP	0.59	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
MindOmni [174]	QwenVL-2.5-7B	VAE + QwenViT	0.83	-	0.71	-	-	-	-	83.2	51.6	-	-	-	-	-	-	-
Skywork UniPic [175]	Qwen-2.5-7B	MAR + SigLIP2	0.86	85.50	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BAGEL [11]	Qwen-2.5-7B	VAE + SigLIP2	0.82	-	0.70	-	-	-	2388.0	85.0	55.3	67.2	67.2	-	-	73.1	-	-
Unified Tokenizer (Single Encoder)																		
UniTok [131]	Llama-2-7B	C. L.	-	-	-	76.8	61.1	83.2	1448.0	-	-	33.9	-	-	-	-	-	-
ATOKEN [132]	SlowFast-LLaVA-1.5-7B	C. L.	0.65	-	-	-	-	-	-	-	48.7	-	-	81.2	61.2	-	74.5	-
Ming-Tok [12]	Ling-lite-16B-A3B	Distill.	0.85	-	-	-	-	-	2023.0	78.5	40.3	64.2	-	82.8	66.6	63.7	72.4	-
EMU3.5 [176]	-	Distill.	0.86	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
EMU2 [139]	Llama-2-33B	Frozen VFMs	-	-	-	84.9	65.1	-	1450.0	65.1	34.1	48.5	62.8	-	-	-	-	66.6
VILA-U [140]	Llama-2-7B	C. L.	-	-	-	75.3	61.1	83.9	1336.0	-	-	27.7	-	-	-	-	-	60.8
QLIP [142]	Vicuna-1.5-7B	C. L.	0.48	78.17	-	78.3	61.8	86.1	1498.3	-	-	-	-	-	-	-	-	55.2
DualToken [143]	Qwen-2.5-3B	Self-Distill.	-	-	-	77.8	-	86.0	1489.2	70.9	38.6	32.5	70.2	-	46.5	-	-	-
BLIP-3o [177]	QwenVL-2.5-7B	Frozen VFMs	0.84	81.60	0.62	83.1	-	-	2329.7	83.5	50.6	66.6	77.5	-	-	-	-	-
UniLIP [15]	InternVL-3-2B	Self-Distill.	0.90	-	0.63	-	-	-	1636.0	80.7	48.7	62.2	75.0	78.6	-	-	-	-
UniFlow [146]	Qwen-2.5-7B	Self-Distill.	-	-	-	-	65.90	89.8	2063.0	83.5	-	-	-	-	-	-	-	81.6
VQRAE [149]	Qwen-2.5-7B	Self-Distill.	0.96	89.78	-	-	-	90.5	1746.8	85.1	61.6	-	77.0	84.8	-	-	-	80.6
PS-VAE [19]	-	Self-Distill.	0.76	83.62	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Unified Tokenizer (Dual Encoder)																		
Muse-VL [22]	Qwen-2.5-32B	Parallel	0.53	-	-	-	-	-	-	81.8	50.1	-	71.0	79.9	55.9	56.7	-	-
TokenFlow [18]	Qwen-2.5-14B	Parallel	0.55	73.38	-	-	-	-	1922.2	76.8	43.2	48.2	72.6	75.8	-	-	-	-
Harmon [17]	Qwen-2.5-1.5B	MAR	0.76	-	0.41	-	58.9	87.6	1476.0	65.5	38.9	-	-	-	-	-	-	-
TA-TiTok [153]	-	Self-Distill.	0.55	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ILLUME+ [154]	Qwen-2.5-3B	Parallel	0.72	-	-	-	-	87.6	1414.0	80.8	44.3	40.3	73.3	74.2	-	-	67.2	69.9
UniToken [155]	Chameleon-7B	Parallel	0.63	-	-	-	-	-	-	71.1	32.8	-	69.9	68.7	38.5	46.1	75.7	-
SemHiTok [23]	Vicuna-1.5-7B	Parallel	-	-	-	61.7	85.5	1993.8	75.2	41.0	36.6	79.8	-	-	-	-	-	-
Show-o2 [16]	Qwen-2.5-7B	Parallel	0.76	86.14	-	-	63.1	-	1620.5	79.3	48.9	-	69.8	78.6	-	56.6	-	-
MANZANO [133]	Vicuna-13B	Parallel	-	-	-	-	-	-	-	83.4	57.8	-	76.0	86.0	73.3	-	86.3	84.4
TokLIP [14]	Qwen-2.5-7B	Cascade	-	-	-	-	57.3	85.2	1586.8	77.4	47.1	-	72.1	77.7	-	-	-	-
TUNA [13]	Qwen-2.5-7B	Cascade	0.90	86.76	-	-	63.9	-	1641.5	-	49.8	-	74.7	79.3	-	61.2	74.3	-
OpenVision3 [159]	Qwen-2.5-7B	Cascade	-	-	-	-	61.1	85.3	1679	-	-	-	66.0	-	-	-	-	-

 doi:10.20944/preprints202607.1756.v1

5.2.1. Reconstruction Fidelity

Reconstruction fidelity is evaluated using three metrics: rFID, PSNR, and SSIM [178]. rFID (reconstruction FID) measures the distributional divergence between original and reconstructed images in deep feature space and serves as the primary perceptual quality metric. PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) complement rFID by assessing pixel-level accuracy and structural similarity.

5.2.2. Codebook Analysis

For discrete tokenizers, codebook utilization is critical. Codebook Usage indicates whether the model avoids codebook collapse and effectively exploits the discrete vocabulary.

5.2.3. Analysis

Table 4 compares state-of-the-art low-level tokenizers on ImageNet-1K. Several key observations emerge.

Evolution of Discrete Tokenizers. Early models such as VQGAN [85] exhibited high reconstruction error (rFID 4.98) and low codebook utilization (5.9%), reflecting severe codebook collapse. Recent quantization advances have significantly improved performance. Methods such as BSQ [108] and WeTok [109] achieve nearly full codebook utilization and strong perceptual fidelity; notably, WeTok reaches an rFID of 0.19, showing that discrete tokenizers can rival or surpass continuous counterparts.
Continuous vs. Discrete Gap. Continuous tokenizers (e.g., VAEs) historically achieved better reconstruction due to unconstrained latent spaces. Models such as DC-AE [37] and SD3-VAE [168] remain strong (rFID ≈0.6–1.0). However, recent discrete methods—including Open-MAGVIT2 [166], BSQ [108], and WeTok [109]—have largely closed this gap while preserving the advantages of discrete representations for autoregressive modeling.
Impact of Foundation Models. Another trend is integrating pretrained vision foundation models into tokenization. Methods such as VA-VAE [125] and RAE [102], which leverage DINOv2 features, achieve strong reconstruction (rFID 0.28 and 0.49), suggesting that semantic priors from large-scale models enhance tokenizer capacity.
Trade-offs in 1D Tokenization. 1D tokenization approaches such as TiTok [101] enable extreme compression but still trail 2D spatial tokenizers in reconstruction quality (rFID 2.21), highlighting the challenge of balancing compression and fidelity.

5.3. Unified Multimodal Evaluation

With the emergence of unified tokenizers (e.g., VILA-U), evaluation has expanded to MLLMs, assessing both generation and understanding within a single framework.

5.3.1. Multimodal Understanding

We evaluate multimodal understanding on benchmarks covering: *VQA:* VQAv2 [179], GQA [180], VizWiz [181]; *comprehensive evaluation:* MMBench [182], MME [183], MMMU [184], SEED-Bench; *hallucination:* POPE [185], HallusionBench [186].

Task-specific tokenizers (e.g., Show-o, Harmon) benefit from specialized optimization, but unified tokenizers such as BLIP-3o [177] and VQRAE [149] achieve competitive results. On OCRBench, MANZANO [133] scores 86.3, surpassing EMU3 [171] (68.7), though some task-specific models (e.g., Lumina-DiMOO) still lead certain metrics (MMBench 84.5).

Among unified methods, single-encoder designs generally perform best. BLIP-3o reaches 2329.7 on MME, compared with 1993.8 for the best dual-encoder model (SemHiTok), suggesting tighter visual–text alignment in single-encoder architectures.

 doi:10.20944/preprints202607.1756.v1

5.3.2. Visual Generation

We evaluate two categories of generation: **class-conditional generation** and **multimodal generation**. Class-conditional generation synthesizes images from semantic labels and is evaluated on ImageNet-1K using gFID, IS, Precision, and Recall.

As shown in Table 5, continuous tokenizers generally achieve higher fidelity than discrete ones. For example, SD-VAE [31] achieves gFID 1.26, outperforming VQGAN [85] (5.20) and LlamaGen [86] (3.09), suggesting that quantization error may limit discrete tokenizers in this setting.

Unified tokenizers, especially single-encoder designs, further improve generation quality. RAE [102] and RecTok [21] achieve gFID 1.13, surpassing both continuous (SD-VAE 1.26) and discrete baselines (IBQ [105] 2.05). Unified tokenizers are also more parameter-efficient. Several models (e.g., VA-VAE, AVFM, FAE) achieve high fidelity (gFID $<$1.40) with $\sim$675M parameters, whereas discrete models often require larger backbones (1.4B–2.1B) yet yield worse results. This indicates that unified tokenization provides more efficient feature representation and cross-modal alignment.

For multimodal generation, we report results on GenEval [187], DPG-Bench [188], and WISE [189]. Unified tokenizers significantly outperform task-specific tokenizers: for example, reconstruction-oriented methods such as Show-o achieve only 0.53 on GenEval, whereas unified models with semantic supervision perform substantially better. Among unified models, single-encoder designs achieve the best performance, with VQRAE [149] and UniLIP [15] reaching 0.96 and 0.90 on GenEval. Dual-encoder approaches generally perform slightly worse, although cascade designs such as TUNA [13] remain competitive and outperform parallel architectures like TokenFlow [18].

6. Challenges and Future Directions

Scaling of Unified Tokenizers. A central challenge is balancing high-level semantic representations for multimodal understanding with the fine-grained representations required for high-fidelity generation. Recent unified tokenizers attempt to narrow this semantic–fidelity gap through techniques such as semantic distillation and frequency-domain decomposition, but their effectiveness at scale remains largely unexplored. Systematic studies on tokenizer scaling are scarce [136], and most existing vision–language and generative models are still limited to fewer than 10B parameters. As a result, the benefits of unified tokenization have not yet been fully validated in large-scale multimodal systems. Future work should therefore investigate how unified tokenizers behave as model size, dataset scale, and training compute increase. In particular, it is important to determine whether unified tokenizers can consistently improve performance relative to task-specific baselines while preserving low-level visual fidelity as semantic abstraction strengthens.

Towards Omni-modal Unified Tokenizers. The emergence of world models and omni-modal systems introduces new requirements for tokenization. Instead of compressing a single modality, tokenizers must support unified semantic grounding, consistent spatiotemporal alignment, and efficient cross-modal interaction across modalities such as images, video, audio, and language. Early work such as AToken [132] has begun exploring unified tokenization frameworks, but research in this direction remains limited. Future work should extend unified tokenization to a broader range of modalities within a shared latent space. Such omni-modal tokenizers could enable scalable world learners that capture coherent spatiotemporal dynamics and rich cross-modal dependencies, facilitating stronger knowledge sharing across tasks and modalities.

Tokenizer-free Understanding and Generation. Recent studies have begun exploring tokenizer-free paradigms for multimodal understanding [190–192]. By removing standalone visual encoders, these approaches encourage a vision-centric information flow that strengthens cross-modal interactions and improves data scalability. For example, visual tokens may dominate attention during inference, enabling stronger visual–language alignment compared with modular architectures. A similar shift is occurring in image generation: JiT [193] directly models pixel-space generation without relying on pretrained tokenizers, preserving intrinsic low-dimensional structures of visual data. These developments suggest a broader direction toward self-contained architectures that reduce reliance on external

 doi:10.20944/preprints202607.1756.v1

tokenizers. Extending tokenizer-free designs to richer multimodal settings may further enable unified end-to-end understanding and generation.

Tokenizer with 3D-Awareness. Another emerging challenge is integrating geometric reasoning into visual tokenization. Current 3D understanding models [194–196] often excel at semantic reasoning but lack precise geometric grounding, while reconstruction approaches [197–199] capture detailed geometry but operate largely independently from semantic cognition. This separation creates a disconnect between high-level reasoning and low-level geometric representation. Future research should aim to unify these perspectives by developing tokenization frameworks that jointly encode semantic reasoning and geometric structure, enabling more comprehensive 3D scene understanding and generation.

Dynamic and Adaptive Tokenization. Most existing tokenizers rely on fixed compression rates, treating all visual regions equally and often introducing computational redundancy. In contrast, human perception allocates more resources to salient regions while processing peripheral information at lower resolution. Incorporating such adaptive mechanisms could improve both efficiency and representation quality. Recent work on 1D tokenizers [101,117,127,153,200–204] provides flexibility but may lose spatial or temporal structure. Approaches such as DCS-LDM [205] partially mitigate this issue through patch-level compression. However, token allocation is still largely determined by heuristics. Developing learnable importance estimators that dynamically allocate tokens based on content may be crucial for scaling tokenization to high-resolution imagery and long-form video.

From Visual Representation to World Modeling. Current visual tokenizers [44,58,136] primarily encode static appearance and semantic information while remaining largely physics-blind. As research moves toward general world models, tokenizers must capture temporal dynamics and causal relationships. Recent work such as V-JEPA2 [206] models latent temporal dynamics to learn physical regularities, while LAPA [207] represents visual changes as action tokens for vision–language–action systems. Building on these ideas, future world models may employ unified tokenizers as shared state-space representations that integrate perception, generation, and action. Such unified token sequences could allow a single architecture to model complex world dynamics and move beyond modular multimodal systems.

7. Conclusions

Visual tokenization has emerged as a key abstraction for modern vision and multimodal systems, enabling high-dimensional visual signals to be represented and processed within token-based architectures. In this survey, we present a comprehensive overview of visual tokenizers and organize a rapidly growing body of work into a unified framework. By distinguishing between semantic-oriented and reconstruction-oriented paradigms and tracing their evolution, we highlight how different design choices reflect distinct objectives, assumptions, and evaluation strategies.

With the rise of unified multimodal models, visual tokenization is increasingly expected to support both understanding and generation within a shared representation. We therefore review recent progress in unified tokenizers and analyze their strengths and limitations across diverse tasks and benchmarks. Our analysis highlights several key challenges, including balancing semantic abstraction with visual fidelity, efficiently scaling token representations, and designing evaluation protocols suited to modern multimodal systems. Visual tokenization is becoming an increasingly central component of future multimodal architectures. Advances in unified and adaptive tokenization may enable more scalable and general-purpose AI systems capable of perception, reasoning, and generation within a shared representational framework. We hope this survey serves as a useful reference for future research in visual tokenization and multimodal representation learning.

 doi:10.20944/preprints202607.1756.v1

References

1. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* **2023**.
2. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the ACL, 2019.
3. Yu, L.; Cheng, Y.; Sohn, K.; Lezama, J.; Zhang, H.; Chang, H.; Hauptmann, A.G.; Yang, M.H.; Hao, Y.; Essa, I.; et al. Magvit: Masked generative video transformer. In Proceedings of the CVPR, 2023.
4. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the ICLR, 2021.
5. 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 Proceedings of the ICML, 2021.
6. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* **2013**.
7. van den Oord, A.; Vinyals, O.; Kavukcuoglu, K. Neural Discrete Representation Learning. In Proceedings of the NeurIPS, 2017.
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the CVPR, 2016.
9. 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 CVPR, 2024.
10. Wu, C.; Chen, X.; Wu, Z.; Ma, Y.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; Ruan, C.; et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848* **2024**.
11. Deng, C.; Zhu, D.; Li, K.; Gou, C.; Li, F.; Wang, Z.; Zhong, S.; Yu, W.; Nie, X.; Song, Z.; et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683* **2025**.
12. Huang, Z.; Zheng, D.; Zou, C.; Liu, R.; Wang, X.; Ji, K.; Chai, W.; Sun, J.; Wang, L.; Lv, Y.; et al. Ming-univision: Joint image understanding and generation with a unified continuous tokenizer. *arXiv preprint arXiv:2510.06590* **2025**.
13. Liu, Z.; Ren, W.; Liu, H.; Zhou, Z.; Chen, S.; Qiu, H.; Huang, X.; An, Z.; Yang, F.; Patel, A.; et al. TUNA: Taming Unified Visual Representations for Native Unified Multimodal Models. *arXiv preprint arXiv:2512.02014* **2025**.
14. Lin, H.; Wang, T.; Ge, Y.; Ge, Y.; Lu, Z.; Wei, Y.; Zhang, Q.; Sun, Z.; Shan, Y. Toklip: Marry visual tokens to clip for multimodal comprehension and generation. *arXiv preprint arXiv:2505.05422* **2025**.
15. 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**.
16. Xie, J.; Yang, Z.; Shou, M.Z. Show-o2: Improved Native Unified Multimodal Models. In Proceedings of the NeurIPS, 2025.
17. Wu, S.; Zhang, W.; Xu, L.; Jin, S.; Wu, Z.; Tao, Q.; Liu, W.; Li, W.; Loy, C.C. Harmonizing visual representations for unified multimodal understanding and generation. In Proceedings of the ICCV, 2025.
18. 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 CVPR, 2025.
19. Zhang, S.; Zhang, H.; Zhang, Z.; Ge, C.; Xue, S.; Liu, S.; Ren, M.; Kim, S.Y.; Zhou, Y.; Liu, Q.; et al. Both Semantics and Reconstruction Matter: Making Representation Encoders Ready for Text-to-Image Generation and Editing. *arXiv preprint arXiv:2512.17909* **2025**.
20. Fan, W.; Diao, H.; Wang, Q.; Lin, D.; Liu, Z. The Prism Hypothesis: Harmonizing Semantic and Pixel Representations via Unified Autoencoding. *arXiv preprint arXiv:2512.19693* **2025**.
21. Shi, Q.; Wu, S.; Bai, J.; Yu, K.; Wang, Y.; Tong, Y.; Li, X.; Li, X. RecTok: Reconstruction Distillation along Rectified Flow. *arXiv preprint arXiv:2512.13421* **2025**.
22. Xie, R.; Du, C.; Song, P.; Liu, C. Muse-vl: Modeling unified vlm through semantic discrete encoding. In Proceedings of the ICCV, 2025.
23. Chen, Z.; Wang, C.; Chen, X.; Xu, H.; Huang, R.; Zhou, J.; Han, J.; Xu, H.; Liang, X. Semhitok: A unified image tokenizer via semantic-guided hierarchical codebook for multimodal understanding and generation. *arXiv preprint arXiv:2503.06764* **2025**.

 doi:10.20944/preprints202607.1756.v1

24. Shi, Q.; Bai, J.; Zhao, Z.; Chai, W.; Yu, K.; Wu, J.; Song, S.; Tong, Y.; Li, X.; Li, X.; et al. Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. *arXiv preprint arXiv:2505.23606* **2025**.
25. Zhou, C.; Yu, L.; Babu, A.; Tirumala, K.; Yasunaga, M.; Shamis, L.; Kahn, J.; Ma, X.; Zettlemoyer, L.; Levy, O. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039* **2024**.
26. Yang, L.; Tian, Y.; Li, B.; Zhang, X.; Shen, K.; Tong, Y.; Wang, M. Mmada: Multimodal large diffusion language models. In Proceedings of the NeurIPS, 2025.
27. Xie, J.; Mao, W.; Bai, Z.; Zhang, D.J.; Wang, W.; Lin, K.Q.; Gu, Y.; Chen, Z.; Yang, Z.; Shou, M.Z. Show-o: One single transformer to unify multimodal understanding and generation. In Proceedings of the ICLR, 2025.
28. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *NeurIPS* **2020**.
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *NeurIPS* **2017**.
30. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. *NeurIPS* **2023**.
31. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the CVPR, 2022.
32. LeCun, Y.; Kavukcuoglu, K.; Farabet, C. Convolutional networks and applications in vision. In Proceedings of the ISCAS. IEEE, 2010.
33. Hegde, D.; Valanarasu, J.M.J.; Patel, V.M. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. *arXiv preprint arXiv:2303.11313* **2023**.
34. Wang, F.; Shi, Y.; Yang, C.; Guo, Q.; Sun, J.; Yuille, A.; Wang, P. VTok: A Unified Video Tokenizer with Decoupled Spatial-Temporal Latents. *arXiv preprint arXiv:2602.04202* **2026**.
35. Gwilliam, M.; Wang, X.; Hu, X.; Yang, Z. Implicit Neural Representation Facilitates Unified Universal Vision Encoding. *arXiv preprint arXiv:2601.14256* **2026**.
36. 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**.
37. Chen, J.; Cai, H.; Chen, J.; Xie, E.; Yang, S.; Tang, H.; Li, M.; Lu, Y.; Han, S. Deep compression autoencoder for efficient high-resolution diffusion models. *arXiv preprint arXiv:2410.10733* **2024**.
38. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the NeurIPS, 2012.
39. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the ICLR, 2015.
40. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. *ICCV* **2021**.
41. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the CVPR, 2022.
42. 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 Proceedings of the ICML, 2021.
43. Zhai, X.; Wang, X.; Mustafa, B.; Steiner, A.; Keysers, D.; Kolesnikov, A.; Beyer, L. Lit: Zero-shot transfer with locked-image text tuning. In Proceedings of the CVPR, 2022.
44. Zhai, X.; Mustafa, B.; Kolesnikov, A.; Beyer, L. Sigmoid loss for language image pre-training. In Proceedings of the ICCV, 2023.
45. Bolya, D.; Huang, P.Y.; Sun, P.; Cho, J.H.; Madotto, A.; Wei, C.; Ma, T.; Zhi, J.; Rajasegaran, J.; Rasheed, H.; et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181* **2025**.
46. Bai, J.; Lei, Y.; Wu, H.; Zhu, Y.; Li, S.; Xin, Y.; Li, X.; Tao, M.; Grover, A.; Yang, M.H. From Masks to Worlds: A Hitchhiker's Guide to World Models. *arXiv preprint arXiv:2510.20668* **2025**.
47. Bao, H.; Dong, L.; Piao, S.; Wei, F. Beit: Bert pre-training of image transformers. In Proceedings of the ICLR, 2022.
48. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the CVPR, 2022.

 doi:10.20944/preprints202607.1756.v1

49. Assran, M.; Duval, Q.; Misra, I.; Bojanowski, P.; Vincent, P.; Rabbat, M.; LeCun, Y.; Ballas, N. Self-supervised learning from images with a joint-embedding predictive architecture. In Proceedings of the CVPR, 2023.
50. Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; Cao, Y. Eva: Exploring the limits of masked visual representation learning at scale. In Proceedings of the CVPR, 2023.
51. Maninis, K.K.; Chen, K.; Ghosh, S.; Karpur, A.; Chen, K.; Xia, Y.; Cao, B.; Salz, D.; Han, G.; Dlabal, J.; et al. TIPS: Text-image pretraining with spatial awareness. In Proceedings of the ICLR, 2025.
52. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the CVPR, 2020.
53. Chen, X.; Xie, S.; He, K. An empirical study of training self-supervised vision transformers. In Proceedings of the ICCV, 2021.
54. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the ICML, 2020.
55. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. Bootstrap your own latent-a new approach to self-supervised learning. In Proceedings of the NeurIPS, 2020.
56. 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 ICCV, 2021.
57. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; HAZIZA, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning Robust Visual Features without Supervision. *TMLR* **2024**.
58. Siméoni, O.; Vo, H.V.; Seitzer, M.; Baldassarre, F.; Oquab, M.; Jose, C.; Khalidov, V.; Szafraniec, M.; Yi, S.; Ramamonjisoa, M.; et al. Dinov3. *arXiv preprint arXiv:2508.10104* **2025**.
59. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* **2015**.
60. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the ICML, 2021.
61. Ren, S.; Wei, F.; Zhang, Z.; Hu, H. Tinymim: An empirical study of distilling mim pre-trained models. In Proceedings of the CVPR, 2023.
62. Bai, Y.; Wang, Z.; Xiao, J.; Wei, C.; Wang, H.; Yuille, A.L.; Zhou, Y.; Xie, C. Masked autoencoders enable efficient knowledge distillers. In Proceedings of the CVPR, 2023.
63. Zhou, C.; Li, X.; Loy, C.C.; Dai, B. Edgesam: Prompt-in-the-loop distillation for on-device deployment of sam. *arXiv preprint* **2023**.
64. Ranzinger, M.; Heinrich, G.; Kautz, J.; Molchanov, P. AM-RADIO: Agglomerative vision foundation model reduce all domains into one. In Proceedings of the CVPR, 2024.
65. Heinrich, G.; Ranzinger, M.; Yin, H.; Lu, Y.; Kautz, J.; Tao, A.; Catanzaro, B.; Molchanov, P. RADIOv2.5: Improved baselines for agglomerative vision foundation models. In Proceedings of the CVPR, 2025.
66. NVIDIA. C-RADIOv3-L (Hugging Face model card). https://huggingface.co/nvidia/C-RADIOv3-L, 2025.
67. Ranzinger, M.; Heinrich, G.; McCarthy, C.; Kautz, J.; Tao, A.; Catanzaro, B.; Molchanov, P. C-RADIOv4 (Tech Report). *arXiv preprint arXiv:2601.17237* **2026**.
68. Wang, H.; Vasu, P.K.A.; Faghri, F.; Vemulapalli, R.; Farajtabar, M.; Mehta, S.; Rastegari, M.; Tuzel, O.; Pouransari, H. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. In Proceedings of the CVPR, 2024.
69. Yuan, H.; Li, X.; Zhou, C.; Li, Y.; Chen, K.; Loy, C.C. Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively. In Proceedings of the ECCV, 2024.
70. Shang, J.; Schmeckpeper, K.; May, B.B.; Minniti, M.V.; Kelestemur, T.; Watkins, D.; Herlant, L. Theia: Distilling Diverse Vision Foundation Models for Robot Learning. In Proceedings of the CoRL, 2024.
71. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *IJCV* **2015**.
72. Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the ICCV, 2017.
73. Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In Proceedings of the NeurIPS, 2022.
74. Gadre, S.Y.; Ilharco, G.; Fang, A.; Hayase, J.; Smyrnis, G.; Nguyen, T.; Marten, R.; Wortsman, M.; Ghosh, D.; Zhang, J.; et al. Datacomp: In search of the next generation of multimodal datasets. In Proceedings of the NeurIPS, 2023.

 doi:10.20944/preprints202607.1756.v1

75. Wang, X.; Alabdulmohsin, I.; Salz, D.; Li, Z.; Rong, K.; Zhai, X. Scaling pre-training to one hundred billion data for vision language models. *arXiv preprint arXiv:2502.07617* **2025**.
76. Tian, Y.; Fan, L.; Isola, P.; Chang, H.; Krishnan, D. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. In Proceedings of the NeurIPS, 2023.
77. Yu, Q.; Sun, Q.; Zhang, X.; Cui, Y.; Zhang, F.; Cao, Y.; Wang, X.; Liu, J. Capsfusion: Rethinking image-text data at scale. In Proceedings of the CVPR, 2024.
78. Tian, Y.; Fan, L.; Chen, K.; Katabi, D.; Krishnan, D.; Isola, P. Learning vision from models rivals learning vision from data. In Proceedings of the CVPR, 2024.
79. Xu, H.; Xie, S.; Tan, X.; Huang, P.Y.; Howes, R.; Sharma, V.; Li, S.W.; Ghosh, G.; Zettlemoyer, L.; Feichtenhofer, C. Demystifying CLIP Data. In Proceedings of the ICLR, 2024.
80. Vo, H.V.; Khalidov, V.; Darcet, T.; Moutakanni, T.; Smetanin, N.; Szafraniec, M.; Touvron, H.; camille couprie.; Oquab, M.; Joulin, A.; et al. Automatic Data Curation for Self-Supervised Learning: A Clustering-Based Approach. *TMLR* **2024**.
81. Charton, F.; Kempe, J. Emergent properties with repeated examples. *arXiv preprint arXiv:2410.07041* **2024**.
82. Zhao, J.; Kim, Y.; Zhang, K.; Rush, A.; LeCun, Y. Adversarially regularized autoencoders. In Proceedings of the ICML, 2018.
83. Tolstikhin, I.; Bousquet, O.; Gelly, S.; Schoelkopf, B. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558* **2017**.
84. Kolouri, S.; Pope, P.E.; Martin, C.E.; Rohde, G.K. Sliced Wasserstein Auto-Encoders. In Proceedings of the ICLR, 2019.
85. Esser, P.; Rombach, R.; Ommer, B. Taming transformers for high-resolution image synthesis. In Proceedings of the CVPR, 2021.
86. Sun, P.; Jiang, Y.; Chen, S.; Zhang, S.; Peng, B.; Luo, P.; Yuan, Z. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525* **2024**.
87. Bai, J.; Ye, T.; Chow, W.; Song, E.; Chen, Q.G.; Li, X.; Dong, Z.; Zhu, L.; Yan, S. Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2024.
88. Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; Frey, B. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644* **2015**.
89. Huang, H.; He, R.; Sun, Z.; Tan, T.; et al. Introvae: Introspective variational autoencoders for photographic image synthesis. *Advances in neural information processing systems* **2018**, *31*.
90. Daniel, T.; Tamar, A. Soft-introvae: Analyzing and improving the introspective variational autoencoder. In Proceedings of the CVPR, 2021.
91. Dilokthanakul, N.; Mediano, P.A.; Garnelo, M.; Lee, M.C.; Salimbeni, H.; Arulkumaran, K.; Shanahan, M. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648* **2016**.
92. Tomczak, J.; Welling, M. VAE with a VampPrior. In Proceedings of the AISTATS, 2018.
93. Takahashi, H.; Iwata, T.; Yamanaka, Y.; Yamada, M.; Yagi, S. Variational autoencoder with implicit optimal priors. In Proceedings of the AAAI, 2019.
94. Aneja, J.; Schwing, A.; Kautz, J.; Vahdat, A. A contrastive learning approach for training variational autoencoder priors. *NeurIPS* **2021**.
95. Chen, Y.; Girdhar, R.; Wang, X.; Rambhatla, S.S.; Misra, I. Diffusion autoencoders are scalable image tokenizers. *arXiv preprint arXiv:2501.18593* **2025**.
96. Duggal, S.; Bai, X.; Wu, Z.; Zhang, R.; Shechtman, E.; Torralba, A.; Isola, P.; Freeman, W.T. End-to-End Training for Unified Tokenization and Latent Denoising. *arXiv preprint arXiv:2603.22283* **2026**.
97. Sønderby, C.K.; Raiko, T.; Maaløe, L.; Sønderby, S.K.; Winther, O. Ladder variational autoencoders. *NeurIPS* **2016**.
98. Maaløe, L.; Fraccaro, M.; Liévin, V.; Winther, O. Biva: A very deep hierarchy of latent variables for generative modeling. *NeurIPS* **2019**.
99. Vahdat, A.; Kautz, J. NVAE: A deep hierarchical variational autoencoder. *NeurIPS* **2020**.
100. Child, R. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650* **2020**.
101. Yu, Q.; Weber, M.; Deng, X.; Shen, X.; Cremers, D.; Chen, L.C. An image is worth 32 tokens for reconstruction and generation. *NeurIPS* **2024**.

 doi:10.20944/preprints202607.1756.v1

102. Zheng, B.; Ma, N.; Tong, S.; Xie, S. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690* **2025**.
103. Zhu, L.; Wei, F.; Lu, Y.; Chen, D. Scaling the codebook size of VQ-GAN to 100,000 with a utilization rate of 99%. *NeurIPS* **2024**.
104. Takida, Y.; Shibuya, T.; Liao, W.; Lai, C.H.; Ohmura, J.; Uesaka, T.; Murata, N.; Takahashi, S.; Kumakura, T.; Mitsufuji, Y. Sq-vae: Variational bayes on discrete representation with self-annealed stochastic quantization. *arXiv preprint arXiv:2205.07547* **2022**.
105. Shi, F.; Luo, Z.; Ge, Y.; Yang, Y.; Shan, Y.; Wang, L. Scalable image tokenization with index backpropagation quantization. In Proceedings of the ICCV, 2025.
106. Chang, Y.; Qin, J.; Qiao, L.; Wang, X.; Zhu, Z.; Ma, L.; Wang, X. Scalable training for vector-quantized networks with 100% codebook utilization. *arXiv preprint arXiv:2509.10140* **2025**.
107. 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**.
108. Zhao, Y.; Xiong, Y.; Krähenbühl, P. Image and video tokenization with binary spherical quantization. *arXiv preprint arXiv:2406.07548* **2024**.
109. Zhuang, S.; Guo, Y.; Fu, C.; Huang, Z.; Tian, Z.; Wang, F.; Zhang, Y.; Li, C.; Wang, Y. Wetok: Powerful discrete tokenization for high-fidelity visual reconstruction. *arXiv preprint arXiv:2508.05599* **2025**.
110. Zhuang, S.; Ai, Y.; Han, J.; Mao, W.; Li, X.; Wang, F.; Wang, X.; Li, Y.; Lin, S.; Xu, K.; et al. UniWeTok: An Unified Binary Tokenizer with Codebook Size 2^{128} for Unified Multimodal Large Language Model. *arXiv preprint arXiv:2602.14178* **2026**.
111. Jégou, H.; Douze, M.; Schmid, C. Product Quantization for Nearest Neighbor Search. In Proceedings of the IEEE TPAMI, 2010.
112. Chen, Y.; Guan, T.; Wang, C. Approximate nearest neighbor search by residual vector quantization. *Sensors* **2010**.
113. Lee, D.; Kim, C.; Kim, S.; Cho, M.; Han, W.S. Autoregressive image generation using residual quantization. In Proceedings of the CVPR, 2022.
114. Tian, K.; Jiang, Y.; Yuan, Z.; Peng, B.; Wang, L. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *NeurIPS* **2024**.
115. Babenko, A.; Lempitsky, V. Additive Quantization for Extreme Vector Compression. In Proceedings of the CVPR, 2014.
116. Chen, J.; Zou, D.; He, W.; Chen, J.; Xie, E.; Han, S.; Cai, H. Dc-ae 1.5: Accelerating diffusion model convergence with structured latent space. In Proceedings of the ICCV, 2025.
117. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In Proceedings of the ICML, 2023.
118. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. β-VAE: Learning basic visual concepts with a constrained variational framework. In Proceedings of the ICLR, 2017.
119. Burgess, C.P.; Higgins, I.; Pal, A.; Matthey, L.; Watters, N.; Desjardins, G.; Lerchner, A. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599* **2018**.
120. Shao, H.; Yao, S.; Sun, D.; Zhang, A.; Liu, S.; Liu, D.; Wang, J.; Abdelzaher, T. Controlvae: Controllable variational autoencoder. In Proceedings of the ICML, 2020.
121. Shao, H.; Lin, H.; Yang, Q.; Yao, S.; Zhao, H.; Abdelzaher, T. Dynamicvae: Decoupling reconstruction error and disentangled representation learning. *arXiv preprint arXiv:2009.06795* **2020**.
122. Kim, H.; Mnih, A. Disentangling by factorising. In Proceedings of the ICML. PMLR, 2018.
123. Chen, R.T.; Li, X.; Grosse, R.B.; Duvenaud, D.K. Isolating sources of disentanglement in variational autoencoders. *NeurIPS* **2018**.
124. Yu, S.; Kwak, S.; Jang, H.; Jeong, J.; Huang, J.; Shin, J.; Xie, S. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940* **2024**.
125. Yao, J.; Yang, B.; Wang, X. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In Proceedings of the CVPR, 2025.
126. Chen, B.; Bi, S.; Tan, H.; Zhang, H.; Zhang, T.; Li, Z.; Xiong, Y.; Zhang, J.; Zhang, K. Aligning visual foundation encoders to tokenizers for diffusion models. *arXiv preprint arXiv:2509.25162* **2025**.
127. Chen, H.; Han, Y.; Chen, F.; Li, X.; Wang, Y.; Wang, J.; Wang, Z.; Liu, Z.; Zou, D.; Raj, B. Masked autoencoders are effective tokenizers for diffusion models. In Proceedings of the ICML, 2025.

 doi:10.20944/preprints202607.1756.v1

128. Xiong, T.; Liew, J.H.; Huang, Z.; Feng, J.; Liu, X. Gigatok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation. *arXiv preprint arXiv:2504.08736* **2025**.
129. Vallaeys, T.; Verbeek, J.; Cord, M. Ssdd: Single-step diffusion decoder for efficient image tokenization. *arXiv preprint arXiv:2510.04961* **2025**.
130. Li, X.; Qiu, K.; Chen, H.; Kuen, J.; Gu, J.; Raj, B.; Lin, Z. Imagefolder: Autoregressive image generation with folded tokens. *arXiv preprint arXiv:2410.01756* **2024**.
131. Ma, C.; Jiang, Y.; Wu, J.; Yang, J.; Yu, X.; Yuan, Z.; Peng, B.; Qi, X. Unitok: A unified tokenizer for visual generation and understanding. *arXiv preprint arXiv:2502.20321* **2025**.
132. Lu, J.; Song, L.; Xu, M.; Ahn, B.; Wang, Y.; Chen, C.; Dehghan, A.; Yang, Y. Atoken: A unified tokenizer for vision. *arXiv preprint arXiv:2509.14476* **2025**.
133. Li, Y.; Qian, R.; Pan, B.; Zhang, H.; Huang, H.; Zhang, B.; Tong, J.; You, H.; Du, X.; Gan, Z.; et al. MANZANO: A Simple and Scalable Unified Multimodal Model with a Hybrid Vision Tokenizer. *arXiv preprint arXiv:2509.16197* **2025**.
134. Wang, L.; Zhao, Y.; Zhang, Z.; Feng, J.; Liu, S.; Kang, B. Image understanding makes for a good tokenizer for image generation. *NeurIPS* **2024**.
135. Wang, W.; Zhang, F.; Cui, Y.; Diao, H.; Luo, Z.; Lu, H.; Liu, J.; Wang, X. End-to-end vision tokenizer tuning. *arXiv preprint arXiv:2505.10562* **2025**.
136. Yao, J.; Song, Y.; Zhou, Y.; Wang, X. Towards Scalable Pre-training of Visual Tokenizers for Generation. *arXiv preprint arXiv:2512.13687* **2025**.
137. Zhao, X.; Zhang, Z.; Huang, Y.; Mi, Y.; Mu, G.; Ding, S.; Wang, J.; Guo, R.; Zhou, S. GloTok: Global Perspective Tokenizer for Image Reconstruction and Generation. *arXiv preprint arXiv:2511.14184* **2025**.
138. Susladkar, O.; Prakash, T.; Juvekar, A.; Nguyen, K.A.; Jang, D.H.; Dhillon, I.S.; Lourentzou, I. PyraTok: Language-Aligned Pyramidal Tokenizer for Video Understanding and Generation. *arXiv preprint arXiv:2601.16210* **2026**.
139. Sun, Q.; Cui, Y.; Zhang, X.; Zhang, F.; Yu, Q.; Wang, Y.; Rao, Y.; Liu, J.; Huang, T.; Wang, X. Generative multimodal models are in-context learners. In Proceedings of the CVPR, 2024.
140. Wu, Y.; Zhang, Z.; Chen, J.; Tang, H.; Li, D.; Fang, Y.; Zhu, L.; Xie, E.; Yin, H.; Yi, L.; et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429* **2024**.
141. Ge, Y.; Li, Y.; Ge, Y.; Shan, Y. Divot: Diffusion Powers Video Tokenizer for Comprehension and Generation. *arXiv preprint arXiv:2412.04432* **2024**.
142. Zhao, Y.; Xue, F.; Reed, S.; Fan, L.; Zhu, Y.; Kautz, J.; Yu, Z.; Krähenbühl, P.; Huang, D.A. Qlip: Text-aligned visual tokenization unifies auto-regressive multimodal understanding and generation. *arXiv preprint arXiv:2502.05178* **2025**.
143. Song, W.; Wang, Y.; Song, Z.; Li, Y.; Sun, H.; Chen, W.; Zhou, Z.; Xu, J.; Wang, J.; Yu, K. Dualtoken: Towards unifying visual understanding and generation with dual visual vocabularies. *arXiv preprint arXiv:2503.14324* **2025**.
144. Han, J.; Chen, H.; Zhao, Y.; Wang, H.; Zhao, Q.; Yang, Z.; He, H.; Yue, X.; Jiang, L. Vision as a Dialect: Unifying Visual Understanding and Generation via Text-Aligned Representations. *arXiv preprint arXiv:2506.18898* **2025**.
145. Zheng, A.; Wen, X.; Zhang, X.; Ma, C.; Wang, T.; Yu, G.; Zhang, X.; Qi, X. Vision foundation models as effective visual tokenizers for autoregressive image generation. *arXiv preprint arXiv:2507.08441* **2025**.
146. 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**.
147. Bi, T.; Zhang, X.; Lu, Y.; Zheng, N. Vision Foundation Models Can Be Good Tokenizers for Latent Diffusion Models. *arXiv preprint arXiv:2510.18457* **2025**.
148. Jia, M.; Li, M.; Fan, L.; Shi, T.; Guo, J.; Li, Z.; Guo, X.; Long, X.X.; Zhang, Q.; Tan, P.; et al. DINO-Tok: Adapting DINO for Visual Tokenizers. *arXiv preprint arXiv:2511.20565* **2025**.
149. Du, S.; Guo, J.; Li, B.; Cui, S.; Xu, Z.; Luo, Y.; Wei, Y.; Gai, K.; Wang, X.; Wu, K.; et al. VQRAE: Representation Quantization Autoencoders for Multimodal Understanding, Generation and Reconstruction. *arXiv preprint arXiv:2511.23386* **2025**.
150. Guo, J.; Du, S.; Yao, J.; Liu, W.; Li, B.; Cao, H.; Gai, K.; Yuan, C.; Wu, K.; Wang, X. Visual Generation Tuning. *arXiv preprint arXiv:2511.23469* **2025**.
151. Gao, Y.; Chen, C.; Chen, T.; Gu, J. One Layer Is Enough: Adapting Pretrained Visual Encoders for Image Generation. *arXiv preprint arXiv:2512.07829* **2025**.

 doi:10.20944/preprints202607.1756.v1

152. Chang, H.; Cha, B.; Ye, J.C. DINO-SAE: DINO Spherical Autoencoder for High-Fidelity Image Reconstruction and Generation, 2026, [arXiv:cs.CV/2601.22904].
153. Kim, D.; He, J.; Yu, Q.; Yang, C.; Shen, X.; Kwak, S.; Chen, L.C. Democratizing text-to-image masked generative models with compact text-aware one-dimensional tokens. *arXiv preprint arXiv:2501.07730* **2025**.
154. 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**.
155. Jiao, Y.; Qiu, H.; Jie, Z.; Chen, S.; Chen, J.; Ma, L.; Jiang, Y.G. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. In Proceedings of the CVPR, 2025.
156. 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**.
157. He, X.; Wei, L.; Ouyang, J.; Liao, M.; Xie, L.; Tian, Q. EMMA: Efficient Multimodal Understanding, Generation, and Editing with a Unified Architecture. *arXiv preprint arXiv:2512.04810* **2025**.
158. Chu, X.; Li, R.; Wang, Y. Usp: Unified self-supervised pretraining for image generation and understanding. *arXiv preprint arXiv:2503.06132* **2025**.
159. Zhang, L.; Ren, S.; Liu, Y.; Li, X.; Wang, Z.; Zhou, Y.; Yao, H.; Zheng, Z.; Nie, W.; Liu, G.; et al. OpenVision 3: A Family of Unified Visual Encoder for Both Understanding and Generation. *arXiv preprint arXiv:2601.15369* **2026**.
160. Chen, X.; Wu, Z.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; Ruan, C. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling. *arXiv preprint arXiv:2501.17811* **2025**.
161. Ma, Y.; Liu, X.; Chen, X.; Liu, W.; Wu, C.; Wu, Z.; Pan, Z.; Xie, Z.; Zhang, H.; yu, X.; et al. JanusFlow: Harmonizing Autoregression and Rectified Flow for Unified Multimodal Understanding and Generation, 2024.
162. Fan, L.; Tang, L.; Qin, S.; Li, T.; Yang, X.; Qiao, S.; Steiner, A.; Sun, C.; Li, Y.; Zhu, T.; et al. Unified autoregressive visual generation and understanding with continuous tokens. *arXiv preprint arXiv:2503.13436* **2025**.
163. Labs, B.F. FLUX. https://github.com/black-forest-labs/flux, 2024.
164. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the ECCV, 2014.
165. Zhu, Y.; Li, B.; Xin, Y.; Xia, Z.; Xu, L. Addressing representation collapse in vector quantized models with one linear layer. In Proceedings of the ICCV, 2025.
166. Luo, Z.; Shi, F.; Ge, Y.; Yang, Y.; Wang, L.; Shan, Y. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *arXiv preprint arXiv:2409.04410* **2024**.
167. Chen, H.; Wang, Z.; Li, X.; Sun, X.; Chen, F.; Liu, J.; Wang, J.; Raj, B.; Liu, Z.; Barsoum, E. Softvq-vae: Efficient 1-dimensional continuous tokenizer. In Proceedings of the CVPR, 2025.
168. Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. Scaling rectified flow transformers for high-resolution image synthesis. In Proceedings of the ICML, 2024.
169. Yang, J.; Li, T.; Fan, L.; Tian, Y.; Wang, Y. Latent denoising makes good visual tokenizers. *arXiv preprint arXiv:2507.15856* **2025**.
170. Liu, H.; Yan, W.; Zaharia, M.; Abbeel, P. World model on million-length video and language with blockwise ringattention. *arXiv preprint arXiv:2402.08268* **2024**.
171. Cui, Y.; Chen, H.; Deng, H.; Huang, X.; Li, X.; Liu, J.; Liu, Y.; Luo, Z.; Wang, J.; Wang, W.; et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583* **2025**.
172. Li, Z.; Li, H.; Shi, Y.; Farimani, A.B.; Kluger, Y.; Yang, L.; Wang, P. Dual diffusion for unified image generation and understanding. In Proceedings of the CVPR, 2025.
173. Xin, Y.; Qin, Q.; Luo, S.; Zhu, K.; Yan, J.; Tai, Y.; Lei, J.; Cao, Y.; Wang, K.; Wang, Y.; et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308* **2025**.
174. Xiao, Y.; Song, L.; Chen, Y.; Luo, Y.; Chen, Y.; Gan, Y.; Huang, W.; Li, X.; Qi, X.; Shan, Y. Mindomni: Unleashing reasoning generation in vision language models with rgpo. *arXiv preprint arXiv:2505.13031* **2025**.
175. Wang, P.; Peng, Y.; Gan, Y.; Hu, L.; Xie, T.; Wang, X.; Wei, Y.; Tang, C.; Zhu, B.; Li, C.; et al. Skywork unipic: Unified autoregressive modeling for visual understanding and generation. *arXiv preprint arXiv:2508.03320* **2025**.

 doi:10.20944/preprints202607.1756.v1

176. Cui, Y.; Chen, H.; Deng, H.; Huang, X.; Li, X.; Liu, J.; Liu, Y.; Luo, Z.; Wang, J.; Wang, W.; et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583* **2025**.
177. Chen, J.; Xu, Z.; Pan, X.; Hu, Y.; Qin, C.; Goldstein, T.; Huang, L.; Zhou, T.; Xie, S.; Savarese, S.; et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568* **2025**.
178. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: from error visibility to structural similarity. In Proceedings of the TIP, 2004.
179. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the CVPR, 2017.
180. Hudson, D.A.; Manning, C.D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the CVPR, 2019.
181. Gurari, D.; Li, Q.; Stangl, A.J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; Bigham, J.P. Vizwiz grand challenge: Answering visual questions from blind people. In Proceedings of the CVPR, 2018.
182. Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. Mmbench: Is your multi-modal model an all-around player? In Proceedings of the ECCV. Springer, 2024.
183. Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394* **2023**.
184. Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In Proceedings of the CVPR, 2024.
185. Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W.X.; Wen, J.R. Evaluating object hallucination in large vision-language models. In Proceedings of the EMNLP, 2023.
186. Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In Proceedings of the CVPR, 2024.
187. Ghosh, D.; Hajishirzi, H.; Schmidt, L. Geneval: An object-focused framework for evaluating text-to-image alignment. *NeurIPS* **2023**.
188. Hu, X.; Wang, R.; Fang, Y.; Fu, B.; Cheng, P.; Yu, G. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135* **2024**.
189. Niu, Y.; Ning, M.; Zheng, M.; Jin, W.; Lin, B.; Jin, P.; Liao, J.; Ning, K.; Feng, C.; Zhu, B.; et al. WISE: A World Knowledge-Informed Semantic Evaluation for Text-to-Image Generation. *arXiv preprint arXiv:2503.07265* **2025**.
190. Lei, W.; Wang, J.; Wang, H.; Li, X.; Liew, J.H.; Feng, J.; Huang, Z. The scalability of simplicity: Empirical analysis of vision-language learning with a single transformer. *arXiv preprint arXiv:2504.10462* **2025**.
191. Diao, H.; Cui, Y.; Li, X.; Wang, Y.; Lu, H.; Wang, X. Unveiling encoder-free vision-language models. *Advances in Neural Information Processing Systems* **2024**, *37*, 52545–52567.
192. Diao, H.; Li, M.; Wu, S.; Dai, L.; Wang, X.; Deng, H.; Lu, L.; Lin, D.; Liu, Z. From Pixels to Words–Towards Native Vision-Language Primitives at Scale. *arXiv preprint arXiv:2510.14979* **2025**.
193. Li, T.; He, K. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720* **2025**.
194. Yang, S.; Yang, J.; Huang, P.; Brown, E.; Yang, Z.; Yu, Y.; Tong, S.; Zheng, Z.; Xu, Y.; Wang, M.; et al. Cambrian-s: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670* **2025**.
195. Wang, H.; Zhao, Y.; Wang, T.; Fan, H.; Zhang, X.; Zhang, Z. Ross3d: Reconstructive visual instruction tuning with 3d-awareness. *arXiv preprint arXiv:2504.01901* **2025**.
196. Yang, J.; Yang, S.; Gupta, A.W.; Han, R.; Fei-Fei, L.; Xie, S. Thinking in space: How multimodal large language models see, remember, and recall spaces. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 10632–10643.
197. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* **2021**, *65*, 99–106.
198. Kerbl, B.; Kopanas, G.; Leimkühler, T.; Drettakis, G. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **2023**, *42*, 139–1.
199. Wang, J.; Chen, M.; Karaev, N.; Vedaldi, A.; Rupprecht, C.; Novotny, D. Vggt: Visual geometry grounded transformer. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 5294–5306.

 doi:10.20944/preprints202607.1756.v1

200. Li, J.; Li, D.; Xiong, C.; Hoi, S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In Proceedings of the ICML, 2022.
201. Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P.N.; Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems* **2023**, *36*, 49250–49267.
202. Zhou, Y.; Zhang, T.; Gong, D.; Wu, Y.; Tian, Y.; Wang, H.; Yuan, H.; Wang, J.; Qi, L.; Fei, H.; et al. SAMTok: Representing Any Mask with Two Words. *arXiv preprint arXiv:2601.16093* **2026**.
203. Guo, P.; Wang, J.; Xing, Z.; Liu, C.; Dong, D.; Qian, X.; Wu, Z. DeRA: Decoupled Representation Alignment for Video Tokenization. *arXiv preprint arXiv:2512.04483* **2025**.
204. Duggal, S.; Isola, P.; Torralba, A.; Freeman, W.T. Adaptive length image tokenization via recurrent allocation. In Proceedings of the First Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models, 2024.
205. Zhong, T.; Tian, X.; Wang, X.; Jiang, B.; Tao, X.; Wan, P. Decoupling Complexity from Scale in Latent Diffusion Model. *arXiv preprint arXiv:2511.16117* **2025**.
206. Assran, M.; Bardes, A.; Fan, D.; Garrido, Q.; Howes, R.; Muckley, M.; Rizvi, A.; Roberts, C.; Sinha, K.; Zholus, A.; et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985* **2025**.
207. Ye, S.; Jang, J.; Jeon, B.; Joo, S.; Yang, J.; Peng, B.; Mandlekar, A.; Tan, R.; Chao, Y.W.; Lin, B.Y.; et al. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758* **2024**.