ATOKEN: A UNIFIED TOKENIZER FOR VISION

Anonymous authors
Paper under double-blind review

ABSTRACT

We present ATOKEN, the first unified visual tokenizer that achieves both high-fidelity reconstruction and semantic understanding across images, videos, and 3D assets. Unlike existing tokenizers that specialize in either reconstruction or understanding for single modalities, ATOKEN encodes these diverse visual inputs into a shared 4D latent space, unifying both tasks and modalities in a single framework. Specifically, we introduce a pure transformer architecture with 4D rotary position embeddings to process visual inputs of arbitrary resolutions and temporal durations. To ensure stable training, we introduce an adversarial-free training objective that combines perceptual and Gram matrix losses, achieving state-of-the-art reconstruction quality. By employing a progressive training curriculum, ATOKEN gradually expands from single images, videos, and 3D, and supports both continuous and discrete latent tokens. ATOKEN achieves 0.21 rFID with 82.2% ImageNet accuracy for images, 3.01 rFVD with 40.2% MSRVTT retrieval for videos, and 28.19 PSNR with 90.9% classification accuracy for 3D. In downstream applications, ATOKEN enables both visual generation tasks (*e.g.*, image generation with continuous and discrete tokens, text-to-video generation, image-to-3D synthesis) and understanding tasks (*e.g.*, multimodal LLMs), achieving competitive performance across all benchmarks. These results shed light on the next-generation multimodal AI systems built upon unified visual tokenization.

1 INTRODUCTION

Large Language Models (LLMs) (Chowdhery et al., 2023; Achiam et al., 2023; Touvron et al., 2023; Team et al., 2023; Guo et al., 2025) have achieved unprecedented generalization, with single models handling coding, reasoning, translation, and numerous other tasks that previously required specialized systems. This versatility largely stems from transformer architectures and simple tokenizers, such as BPE (Sennrich et al., 2015), which convert all text types – code, documents, tables, and multiple languages – into a unified token space. This shared representation enables efficient scaling and seamless knowledge transfer across language tasks.

In contrast, visual representations remain fragmented due to inherent complexities. Unlike text's discrete symbolic nature, visual tasks demand distinct levels of abstraction: generation requires tokenizers that preserve low-level visual details for reconstruction, while understanding requires encoders that extract high-level semantic features through text alignment. Moreover, visual data exists in disparate formats: 2D grids for images, temporal sequences for videos, and varied 3D representations (*e.g.*, meshes, voxels, and Gaussian splats) (Mescheder et al., 2019; Achlioptas et al., 2018; Mildenhall et al., 2021). Without a shared representation, vision systems remain limited, unable to achieve the generalization and transfer learning that characterizes modern language models.

Despite recent progress, unified visual tokenizers face three fundamental challenges. First, existing approaches optimize for either reconstruction or understanding, but not both: visual encoders (Radford et al., 2021; Zhai et al., 2023; Bolya et al., 2025) achieve semantic alignment but lack pixel-level detail, while VAE-based tokenizers (Esser et al., 2020; Rombach et al., 2022; Polyak et al., 2024; Yu et al., 2022b) preserve visual details but lack semantic understanding. Second, architectural choices create different limitations: convolutional tokenizers exhibit diminishing returns when scaling model parameters (Xiong et al., 2025), while transformer tokenizers (Yu et al., 2021; Wang et al., 2024b; Hansen-Estruch et al., 2025) achieve better scaling but suffer from severe adver-

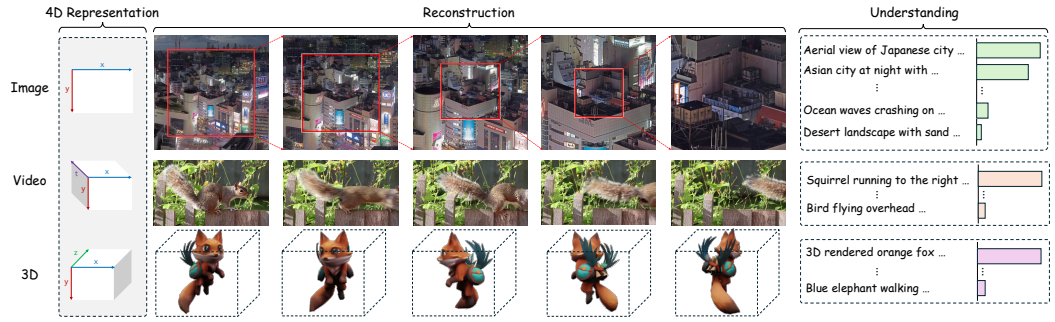


Figure 1: **ATokenon images, videos, and 3D.** Our method uses a shared 4D latent space (left) to produce high-fidelity reconstructions (middle: zoomed regions for images, temporal frames for videos, viewpoints for 3D) while preserving semantic understanding (right: zero-shot text retrieval).

sarial training instabilities. Third, recent unification efforts remain limited to images (Deng et al., 2025; Wu et al., 2024c; Ma et al., 2025), while video and 3D modalities remain unexplored.

We present ATOKEN, a general-purpose visual tokenizer that achieves *high-fidelity reconstruction* and *rich semantic understanding* across *images*, *videos*, and *3D*. Our model learns a unified representation that captures both fine-grained visual details and high-level semantics, accessible through progressive encoding: semantic embeddings for understanding, low-dimensional continuous latents for generation, and discrete tokens via quantization. This design enables the next generation of multimodal systems that seamlessly handle both understanding and generation across visual modalities.

To address format discrepancies across visual modalities, we introduce a sparse 4D representation where each modality naturally occupies different subspaces: images as 2D slices, videos as temporal stacks, and 3D assets as surface voxels extracted from multi-view renderings (Xiang et al., 2024). We implement this through a pure transformer architecture with space-time patch embeddings and 4D Rotary Position Embeddings (RoPE), enabling efficient scaling and joint modeling across all modalities while maintaining native resolution and temporal length processing.

To overcome training instabilities that affect transformer-based visual tokenizers, we develop an adversarial-free loss combining perceptual and Gram matrix terms. This approach achieves state-of-the-art reconstruction quality while maintaining stable, scalable training. We introduce a progressive curriculum that builds capabilities incrementally: starting from a pretrained vision encoder, jointly optimizing reconstruction and understanding for images, extending to videos and 3D data, with optional quantization for discrete tokens. This curriculum reveals that multimodal training can enhance rather than compromise single-modality performance – our final model achieves better image reconstruction than earlier image-only stages while maintaining strong semantic understanding.

ATOKEN demonstrates significant advances in both scalability and performance. The model natively processes arbitrary resolutions and time durations, and accelerates inference through KV-caching mechanisms. To validate its effectiveness, we conduct comprehensive evaluations across three dimensions: reconstruction quality, semantic understanding, and downstream applications. These experiments confirm that ATOKEN achieves competitive or state-of-the-art performance across all modalities while maintaining computational efficiency.

2 BACKGROUND

Visual tokenization transforms raw visual data into compact representations for understanding and generation tasks. However, existing approaches remain fragmented across modalities and objectives, lacking the versatility of language models. To address space constraints, we provide a comprehensive overview of visual tokenization approaches organized along three critical dimensions in Table 6 (Appendix), while extensive related work discussion can be found in Section A (Appendix).

Task Specialization. Current visual tokenizers fall into two distinct categories: reconstruction methods (SD-VAE (Rombach et al., 2022), VQGAN (Esser et al., 2020), GigaTok (Xiong et al., 2025), Cosmos (Agarwal et al., 2025)) excel at compression for generation but cannot extract se-

mantic features; understanding encoders (CLIP (Radford et al., 2021), SigLIP2 (Tschannen et al., 2025), VideoPrism (Zhao et al., 2024)) produce rich semantics but cannot reconstruct content. Only VILA-U (Wu et al., 2024c) and UniTok (Ma et al., 2025) attempt both, limited to images. This divide prevents models that excel at both generation and understanding.

Modality Fragmentation. Beyond task specialization, visual tokenizers are limited to specific modalities. While most video tokenizers naturally handle images as single-frame videos (*e.g.*, TAE (Polyak et al., 2024), Hunyuan (Kong et al., 2024)), they cannot process 3D data. Conversely, 3D tokenizers like Trellis-SLAT (Xiang et al., 2024) are restricted to 3D-only data, unable to leverage the massive image and video data for pretraining. Understanding tasks face similar constraints: image encoders process videos frame-by-frame without temporal compression, while dedicated video encoders (Zhao et al., 2024; Wang et al., 2022b) lack image-specific optimizations.

Architectural Trade-offs. Key design trade-offs emerge across methods: *(1) Architecture:* Understanding encoders use transformers while reconstruction tokenizers favor convolutions (SD-VAE (Rombach et al., 2022)), with recent hybrid (GigaTok (Xiong et al., 2025)) and pure transformer (ViTok (Hansen-Estruch et al., 2025)) approaches, the latter suffering from adversarial training instabilities. *(2) Token representation:* Methods choose discrete tokens for LLM compatibility (VQGAN (Esser et al., 2020)) or continuous tokens for reconstruction quality (TAE (Polyak et al., 2024)), with few supporting both. *(3) Resolution handling:* Convolutions naturally handle arbitrary resolutions, while only SigLIP2 (Tschannen et al., 2025) among transformers supports native resolution. *(4) Training objectives:* GAN-based training dominates reconstruction tokenizers despite instabilities.

3 Model

This section describes AToken's architecture and training. We present our unified 4D representation for all modalities (Section 3.1), the transformer architecture processing these representations (Section 3.2), adversarial-free training objectives (Section 3.3), and a progressive curriculum for multimodal learning (Section 3.4). Detailed training recipes and implementation are in Section B.

3.1 Unified Latent Representation

Unified Modalities – Image, Video and 3D. Our central insight is that all visual modalities can be represented within a shared 4D space. As illustrated in Figure 2, we process each modality through space-time patchification to produce sets of feature-coordinate pairs:

$$\boldsymbol{z} = \{(\boldsymbol{z}_i, \boldsymbol{p}_i)\}_{i=1}^{L}, \quad \boldsymbol{z}_i \in \mathbb{R}^C, \quad \boldsymbol{p}_i \in \{0, 1, \ldots, N-1\}^4 \tag{1}$$

where $\boldsymbol{z}_i$ represents the latent feature at position $\boldsymbol{p}_i = [\mathrm{t}, \mathrm{x}, \mathrm{y}, \mathrm{z}]$ in 4D space (temporal and spatial coordinates), with N defining the resolution along each axis and L the number of active locations.

This sparse representation unifies all modalities by activating only their relevant dimensions: images occupy the (x, y) plane at $\mathrm{t} = \mathrm{z} = 0$, videos extend along the temporal axis with $\mathrm{z} = 0$, and 3D assets as surface voxels in $(\mathrm{x}, \mathrm{y}, \mathrm{z})$ space with $\mathrm{t} = 0$. For 3D assets, we adapt Trellis-SLAT (Xiang et al., 2024) by rendering multi-view images from spherically sampled cameras, applying our unified patchification, then aggregating features into voxel space (detailed in Section 3.2). This approach enables a single encoder $\mathcal{E}$ to process all modalities without architectural modifications.

Unified Tasks – Reconstruction and Understanding. From the unified structured latents $\boldsymbol{z} = \{(\boldsymbol{z}_i, \boldsymbol{p}_i)\}$, we extract representations for both reconstruction and understanding through complementary projections. For reconstruction, we project each latent to a lower-dimensional space $\boldsymbol{z}^r = \boldsymbol{W}_r(\boldsymbol{z})$ with KL regularization (Rombach et al., 2022), optionally applying FSQ (Mentzer et al., 2023) for discrete codes $\tilde{\boldsymbol{z}}^r = \mathrm{FSQ}(\boldsymbol{z}^r)$. The decoder $\mathcal{D}_\theta$ then reconstructs the input from these latents. For understanding, we aggregate latents via attention pooling (Radford et al., 2021; Tschannen et al., 2025) into a global representation $\bar{\boldsymbol{z}}$, which is projected to $\boldsymbol{z}^s = \boldsymbol{W}_s(\bar{\boldsymbol{z}})$ for alignment with text embeddings. This dual projection design allows joint optimization without architectural duplication – the same encoded features $\boldsymbol{z}$ support both pixel-level reconstruction through individual latents and semantic understanding through their aggregation.

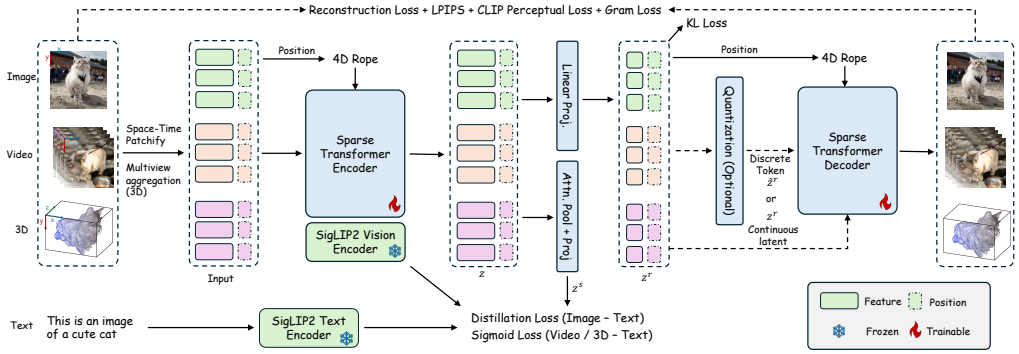


Figure 2: **Overview of our method.** All modalities undergo unified space-time patchification and encoding into sparse 4D latents, which support both reconstruction through modality-specific decoders and understanding through attention pooling and text alignment. The architecture jointly optimizes reconstruction and understanding losses, maintaining sparse structured representations throughout for efficient multimodal processing.

3.2 Transformer based Architecture

Unified Space-Time Patch Embedding. We employ a unified patchification scheme that enables all modalities to share the same encoder. Given an input $\boldsymbol{x} \in \mathbb{R}^{T\times H\times W\times 3}$, we partition it into non-overlapping space-time patches of size $t \times p \times p$. For images ($T = 1$), we apply temporal zero-padding to create t-frame patches, ensuring consistent dimensions across modalities. Videos are directly partitioned along both spatial and temporal dimensions.

For 3D assets, we adapt Trellis-SLAT (Xiang et al., 2024) to our unified pipeline. As shown in Figure 6 in the Appendix, we render multi-view images from spherically sampled cameras and apply our standard space-time patchification. Each voxel in a 64^3 grid is back-projected to gather and average patch features from relevant views. Unlike Xiang et al. (2024), which uses DINOv2 features, we achieve comparable quality using our unified patch representation.

Sparse Transformer Encoder and Decoder. We employ a unified transformer architecture for both encoder and decoder, as illustrated in Figure 2. Both components process sparse structured representations – sets of feature-position pairs rather than dense grids – enabling efficient handling of all modalities with native support for arbitrary resolutions and temporal lengths.

Our encoder $\mathcal{E}$ extends the pretrained SigLIP2 vision tower (Tschannen et al., 2025) from 2D images to 4D representations through two modifications. First, we generalize patch embedding to space-time blocks of size $t \times p \times p$, with zero-initialized temporal weights preserving the original image features. Second, we augment SigLIP2's learnable 2D position embeddings with 4D RoPE (Lu et al., 2024a) applied in every attention layer, providing relative position awareness across $(\mathrm{t}, \mathrm{x}, \mathrm{y}, \mathrm{z})$ dimensions. This design maintains SigLIP2's semantic priors and resolution flexibility while enabling unified processing across modalities.

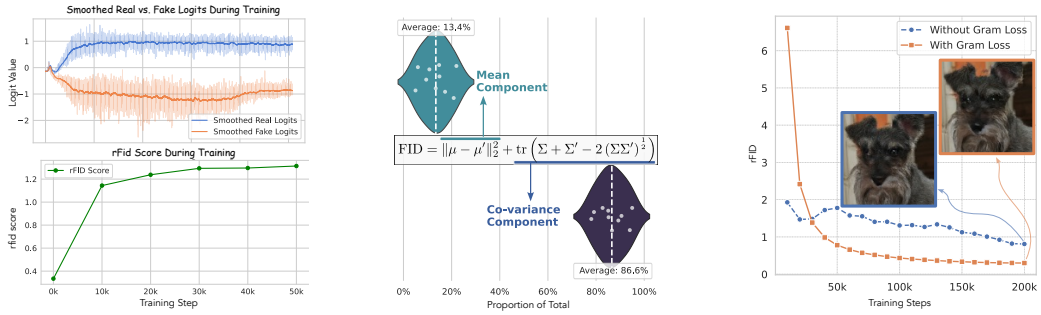
The decoder $\mathcal{D}$ shares the encoder's transformer architecture but is trained from scratch for reconstruction. It maps structured latents back to visual outputs through task-specific heads. For images and videos, we decode directly to pixel space:

$$\mathcal{D}_{\mathrm{P}} : \{(\boldsymbol{z}_i, \boldsymbol{p}_i)\}_{i=1}^{L} \to \boldsymbol{x} \in \mathbb{R}^{T\times H\times W\times 3} \tag{2}$$

treating images as single-frame videos ($T = 1$) and discarding temporal padding following Polyak et al. (2024). For 3D assets, we first decode to pixel-space features, then apply an additional layer to generate Gaussian splatting parameters for efficient rendering:

$$\mathcal{D}_{\mathrm{GS}} : \{(\boldsymbol{z}_i, \boldsymbol{p}_i)\}_{i=1}^{L} \to \{\{(\boldsymbol{o}_i^k, \boldsymbol{c}_i^k, \boldsymbol{s}_i^k, \alpha_i^k, \boldsymbol{r}_i^k)\}_{k=1}^{K}\}_{i=1}^{L} \tag{3}$$

where each location generates K Gaussians with parameters: position offset $\boldsymbol{o}$, color $\boldsymbol{c}$, scale $\boldsymbol{s}$, opacity α, and rotation $\boldsymbol{r}$. Following Xiang et al. (2024), we constrain Gaussian positions to remain near their source voxels using $\boldsymbol{x}_i^k = \boldsymbol{p}_i + \tanh(\boldsymbol{o}_i^k)$, ensuring local feature coherence.



(a) GAN training instability (b) Decomposition of rFID. (c) Gram loss efficiency

Figure 3: **Adversarial-free training with Gram loss achieves stable, high-fidelity reconstruction.** (a) GAN training fails as the discriminator overpowers the generator, degrading rFID. (b) rFID decomposition shows $\approx 86.6\%$ of error stems from covariance (texture/style) vs. $\approx 13.4\%$ from mean. (c) Gram loss directly optimizes second-order statistics without adversarial training, achieving superior and stable rFID.

3.3 Training Objectives

We jointly optimize for reconstruction and understanding through an adversarial-free training loss:

$$\mathcal{L} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}}, \tag{4}$$

where $\mathcal{L}_{\text{KL}}$ is the KL regularization on projected reconstruction latents $\boldsymbol{z}^r$, with weights λ_{rec}, λ_{sem}, λ_{KL}. We achieve state-of-the-art reconstruction without adversarial training, which is unstable at scale (Wu et al., 2025a) and incompatible with sparse 3D representations.

Reconstruction Loss. While GANs (Goodfellow et al., 2014) are standard for visual tokenizers, we found them unsuitable for our transformer architecture. Figure 3(a) shows the discriminator rapidly dominates the generator, causing mode collapse and degraded reconstruction quality. To develop an alternative, we analyzed the reconstruction error by decomposing rFID into mean and covariance components (Figure 3(b)). The covariance component – capturing second-order statistics like texture and style – dominates at $\approx 86.6\%$, while the mean contributes only 13.4%. This motivated adopting Gram matrix loss (Gatys et al., 2016), which directly optimizes feature covariance without adversarial training by computing the Gram matrix $G(F) = FF^\top$ for feature maps from different layers. As shown in Figure 3(c), this achieves superior performance throughout training.

For images, we combine four complementary loss components:

$$\mathcal{L}^{\text{I}}_{\text{rec}} = \lambda_1\mathcal{L}_1 + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}} + \lambda_{\text{GRAM}}\mathcal{L}_{\text{GRAM}} + \lambda_{\text{CLIP}}\mathcal{L}_{\text{CLIP}}, \tag{5}$$

where $\mathcal{L}_1 = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_1$ provides pixel supervision, $\mathcal{L}_{\text{LPIPS}}$ (Zhang et al., 2018) measures perceptual similarity, $\mathcal{L}_{\text{GRAM}}$ captures texture, and $\mathcal{L}_{\text{CLIP}}$ enforces semantic consistency. For video and 3D assets, we use $\mathcal{L}^{\text{V/3D}}_{\text{rec}} = \mathcal{L}_1$ for efficiency, relying on cross-modal transfer from images for details:

$$\mathcal{L} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}}, \tag{6}$$

where $\mathcal{L}_{\text{KL}}$ is the KL regularization term applied to the projected reconstruction latents $\boldsymbol{z}^r$, with λ_{rec}, λ_{sem} and λ_{KL} balancing components. Notably, we achieve state-of-the-art reconstruction quality without adversarial training, which has been observed to be unstable when scaling (Wu et al., 2025a) and incompatible with our sparse 3D representations.

Semantic Loss. We align visual representations $\boldsymbol{z}^s$ with text embeddings through modality-specific objectives. For images, we distill knowledge from the frozen SigLIP2 vision encoder by minimizing the KL divergence between temperature-scaled vision-text similarity distributions:

$$\mathcal{L}^{\text{I}}_{\text{sem}} = \text{KL}\left(\text{softmax}(\tau^{-1}\boldsymbol{s}^{\text{teacher}}) \,\|\, \text{softmax}(\tau^{-1}\boldsymbol{s}^{\text{student}})\right), \tag{7}$$

where $\boldsymbol{s}^{\text{teacher}}$ and $\boldsymbol{s}^{\text{student}}$ are vision-text similarity scores from frozen SigLIP2 and our model respectively, both paired with the same frozen text encoder, and τ is the temperature parameter. For videos and 3D, we directly optimize alignment using the sigmoid loss from SigLIP (Zhai et al., 2023), which proves more stable for the smaller batch sizes typical in these domains. This dual strategy preserves pretrained image semantics while enabling efficient learning for new modalities.

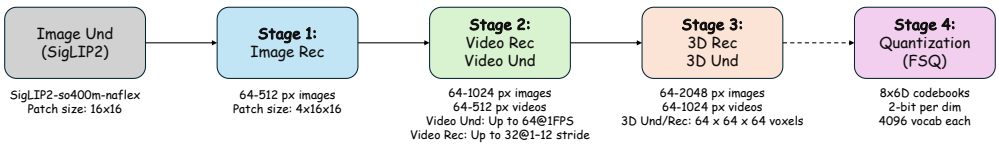


Figure 4: **Progressive training curriculum of AToken.** Our model starts from SigLIP2 image understanding and progressively adds: (1) image reconstruction, (2) video capabilities with temporal modeling, (3) 3D understanding with expanded resolutions, and optionally (4) discrete tokenization via FSQ. Each box shows the new capabilities introduced at that stage, along with supported resolutions, patch sizes, and sampling strategies.

3.4 Training Strategy

Our training employs a four-stage curriculum (Figure 4) that builds from image foundations to video dynamics to 3D geometry, with optional discrete quantization. Starting from the pretrained SigLIP2 vision encoder, we gradually introduce more complex objectives and modalities while using gradient accumulation to balance image-text distillation with reconstruction, video-text alignment, and 3D-text alignment across all stages. This ensures semantic alignment is preserved as reconstruction capabilities expand through round-robin sampling.

Stage 1: Image Foundation. Starting from pretrained SigLIP2, we establish core visual representations by adding image reconstruction capabilities with 32 latent dimensions (Yao & Wang, 2025). Training uses variable resolution sampling from 64 to 512 pixels.

Stage 2: Video Dynamics. We extend to temporal sequences, expanding latent dimensions to 48 for motion complexity (Seawead et al., 2025). Resolution increases to 1024 for images and 512 for videos. We employ temporal tiling with adaptive sampling and KV-caching (Figure 7 in Appendix) to eliminate redundant computation.

Stage 3: 3D Geometry. We incorporate 3D assets as 64^3 voxel grids, using Gaussian splatting for reconstruction and attention pooling for understanding. Resolution further increases to 2048 for images and 1024 for videos. Joint optimization across modalities prevents catastrophic forgetting while leveraging cross-modal learning.

Stage 4: Discrete Tokenization. Optionally, we add FSQ quantization (Mentzer et al., 2023), partitioning 48-dimensional latents into 8 discrete tokens from 4096-entry codebooks, enabling compatibility with discrete generative models across all visual domains.

See Section B.3 for complete training configurations and Section B.4 for implementation details.

4 Results

We evaluate AToken as the first visual tokenizer to achieve reconstruction and understanding across images, videos, and 3D assets. Our evaluation demonstrates that unified tokenization achieves competitive performance across all modalities (Section 4.1), seamlessly integrates into existing understanding pipelines (Section 4.3), enables high-quality generation without architectural changes (Section 4.4), and reveals critical insights about model scaling and cross-modal benefits (Section 4.2). In the Appendix, Section C reveals progressive improvements with detailed per-modality evaluations, and Section D validates versatility in video generation and 3D synthesis applications.

4.1 Unified Tokenizer Evaluation

Table 1 compares visual tokenizers across modalities using ImageNet (Deng et al., 2009) (reconstruction: rFID; understanding: zero-shot accuracy), TokenBench (Agarwal et al., 2025), MSR-VTT (Xu et al., 2016) for video, and Toys4k (Stojanov et al., 2021a) for 3D. Existing approaches fall into three limited categories: reconstruction-only tokenizers (SD-VAE (Rombach et al., 2022), Hunyuan (Kong et al., 2024), Trellis-SLAT (Xiang et al., 2024)) excel at generation but lack semantics; understanding-only encoders (SigLIP2 (Tschannen et al., 2025), VideoPrism (Zhao et al., 2024)) provide semantics but cannot reconstruct; recent unified attempts (SeTok (Wu et al., 2024b), UniTok (Ma et al., 2025)) combine both but remain image-only.

Table 1: **Performance comparison of visual tokenizers across modalities.** We evaluate on ImageNet for image reconstruction and zero-shot classification, TokenBench for video reconstruction with MSR-VTT, and Toys4k for 3D reconstruction and classification. Discrete tokenizers are indicated with gray shading.

Method	Comp. Ratio	Latent Channels	Token Type	Image			Video			3D		
				PSNR↑	rFID↓	Acc.↑	PSNR↑	rFVD↓	R@1↑	PSNR↑	LPIPS↓	Acc.↑
Reconstruction Only												
SD-VAE	(1, 8, 8)	4	VAE	26.26	0.61	-	-	-	-	-	-	-
FLUX.1 [dev]	(1, 8, 8)	16	VAE	32.86	**0.18**	-	-	-	-	-	-	-
VA-VAE	(1, 16, 16)	32	VAE	27.70	0.28	-	-	-	-	-	-	-
GigaTok-XL-XXL	(1, 16, 16)	8	VQ	22.42	0.80	-	-	-	-	-	-	-
Cosmos-0.1-CV8×8	(4, 8, 8)	16	AE	30.11	7.55	-	34.33	8.34	-	-	-	-
OmniTokenizer†	(4, 8, 8)	8	VAE	26.74	1.02	-	19.39	173.48	-	-	-	-
Hunyuan	(4, 8, 8)	16	VAE	**33.32**	0.67	-	36.37	3.78	-	-	-	-
Wan2.2	(4, 16, 16)	48	VAE	31.25	0.75	-	**36.39**	3.19	-	-	-	-
OmniTokenizer†	(4, 8, 8)	8	VQ	24.69	1.41	-	19.89	202.46	-	-	-	-
Cosmos-0.1-DV8×8	(4, 8, 8)	6	FSQ	26.34	7.86	-	31.42	25.94	-	-	-	-
Trellis-SLAT	-	8	VAE	-	-	-	-	-	-	26.97	**0.054**	-
Understanding Only												
SigLIP2-So/16	(1, 16, 16)	-	-	-	-	83.4	-	-	41.9	-	-	-
$PE_{core}L$	(1, 14, 14)	-	-	-	-	**83.5**	-	-	**50.3**	-	-	-
Reconstruction & Understanding												
VILA-U	(1, 16, 16)	16	RQ	22.24	4.23	78.0	-	-	-	-	-	-
UniTok	(1, 16, 16)	64	MCQ	25.34	**0.36**	78.6	-	-	-	-	-	-
ATOKEN-So/D	(4, 16, 16)	48	FSQ	**27.00**	0.38	82.2	**33.12**	**22.16**	40.3	**28.17**	**0.063**	**91.3**
ATOKEN-So/C	(4, 16, 16)	48	VAE	29.72	0.21	82.2	36.07	**3.01**	40.2	28.28	0.062	90.9

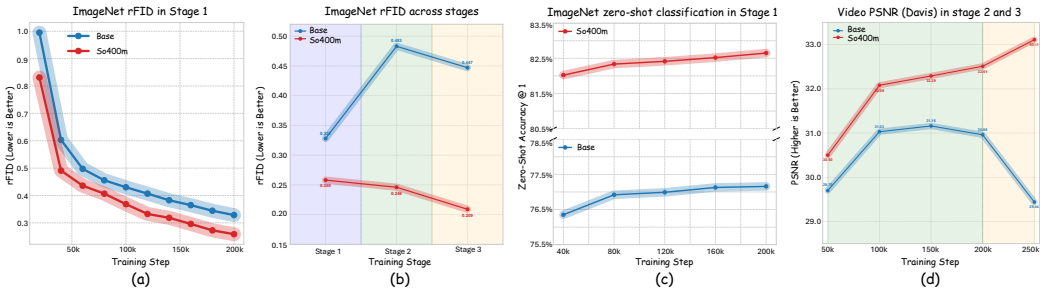


Figure 5: **Architectural scaling comparison: Base vs. So400m models.** (a) ImageNet rFID during Stage 1 training. (b) ImageNet rFID across training stages. (c) ImageNet zero-shot classification accuracy in Stage 1. (d) Video PSNR on DAVIS in Stages 2 and 3. The So400m model maintains or improves performance across all stages, while the Base model shows significant degradation when extending beyond single-modality training, indicating that sufficient model capacity is critical for successful multimodal visual tokenization.

ATOKEN-So/C breaks these boundaries as the first unified tokenizer across all modalities, achieving 0.21 rFID with 82.2% ImageNet accuracy (vs. UniTok's 0.36 rFID and 78.6%), while extending to video (3.01 rFVD, 40.2% R@1) and 3D (28.28 PSNR, 90.9% accuracy), matching specialized methods like Wan2.2 (Wan et al., 2025) and Trellis-SLAT. Our discrete variant (ATOKEN-So/D) maintains competitive performance, pioneering discrete tokenization across all modalities. Detailed evaluations with comprehensive baselines are in Appendix C: image reconstruction and zero-shot benchmarks (Section C.1), video reconstruction and retrieval (Section C.2), 3D reconstruction and classification (Section C.3), with qualitative visualizations (Figures 9 to 11).

4.2 Scaling and Cross-Modal Benefits

To investigate the scaling property of the visual tokenizer, we compare our So400m model with a smaller Base variant following identical training procedures. The Base model initializes from SigLIP-Base-patch16-naflex (Tschannen et al., 2025), comprising 12 transformer blocks with hidden dimension $d = 768$ and 12 attention heads for both encoder and decoder, yielding approximately 192M parameters compared to So400m's 800M.

Table 2: **Image understanding comparison across multimodal LLMs.** Evaluation of SlowFast-LLaVA-1.5 with frozen ATOKEN-So/C vision encoder versus Oryx-ViT and other state-of-the-art MLLMs.

			General & Knowledge					TextRich	
Multimodal LLM	**Vision Encoder**	**# Input Pixels**	**RW-QA (test)**	**AI2D (test)**	**SQA (test)**	**MMMU (val)**	**MathV (testmini)**	**OCRBench (test)**	**TextVQA (val)**
1B Model Comparison									
MolmoE-1B	MetaCLIP	4.10M	60.4	86.4	-	34.9	34.0	-	78.8
SlowFast-LLaVA-1.5-1B	Oryx-ViT	2.36M	59.2	72.8	87.7	40.5	51.0	70.0	71.3
SlowFast-LLaVA-1.5-1B	ATOKEN-So/C	2.36M	60.1	74.2	88.7	40.6	52.5	67.6	72.5
3B Model Comparison									
MM1.5-3B	CLIP	4.52M	56.9	65.7	85.8	37.1	44.4	65.7	76.5
SlowFast-LLaVA-1.5-3B	Oryx-ViT	2.36M	63.4	77.0	90.3	44.7	58.6	73.4	73.0
SlowFast-LLaVA-1.5-3B	ATOKEN-So/C	2.36M	64.3	79.1	89.7	45.7	58.4	73.3	72.8
7B Model Comparison									
Oryx1.5-7B	Oryx-ViT	2.36M	-	79.7	-	47.1	-	71.3	75.7
InternVL2.5-8B	InternViT	9.63M	70.1	84.5	-	56.0	64.4	-	79.1
Qwen2-VL-7B	DFN	-	70.1	83.0	-	54.1	58.2	-	84.3
SlowFast-LLaVA-1.5-7B	Oryx-ViT	2.36M	67.5	80.4	91.1	49.0	62.5	76.4	76.4
SlowFast-LLaVA-1.5-7B	ATOKEN-So/C	2.36M	68.8	81.2	92.1	48.7	61.2	74.5	77.7

Table 3: **Video understanding performance on multimodal LLMs.** Evaluation of SlowFast-LLaVA-1.5 with frozen ATOKEN-So/C vision encoder versus Oryx-ViT and other video MLLMs.

			General VideoQA			Long-Form Video Understanding		
Multimodal LLM	**Vision Encoder**	**# Input Tokens**	**VideoMME (w/o sub)**	**PercepTest (val)**	**NExT-QA (test)**	**LongVideoBench (val)**	**MLVU (m-avg)**	**LVBench (avg)**
1B Model Comparison								
Qwen2-VL-2B	DFN	16K	55.6	53.9	77.2	48.7	62.7	39.4
SlowFast-LLaVA-1.5-1B	Oryx-ViT	9K	56.6	61.9	76.7	54.3	64.3	39.7
SlowFast-LLaVA-1.5-1B	ATOKEN-So/C	9K	56.7	63.9	74.8	55.1	64.7	41.1
3B Model Comparison								
Apollo-3B	SigLIP	3K	58.4	65.0	-	55.1	68.7	-
SF-LLaVA-1.5-3B	Oryx-ViT	9K	60.8	65.8	80.8	57.2	68.8	43.3
SF-LLaVA-1.5-3B	ATOKEN-So/C	9K	60.4	66.0	80.8	57.2	66.7	41.3
7B Model Comparison								
InternVL2.5-8B	InternViT	16K	64.2	-	85.0	60.0	69.0	43.2
Qwen2-VL-7B	DFN	16K	63.3	62.3	81.2	55.6	69.8	44.7
SlowFast-LLaVA-1.5-7B	Oryx-ViT	9K	63.9	69.6	83.3	62.5	71.5	45.3
SlowFast-LLaVA-1.5-7B	ATOKEN-So/C	9K	64.5	70.3	83.7	60.6	69.8	44.8

As shown in Figure 5, both models achieve reasonable single-modal performance in Stage 1, with So400m outperforming Base (0.258 vs 0.323 rFID, 82.7% vs 77.2% accuracy). However, the Base model suffers severe degradation when expanding to videos, with ImageNet rFID degrading 49% (0.323→0.483) and video PSNR declining across stages. In contrast, So400m improves continuously – ImageNet rFID enhances 19% (0.258→0.209) while video PSNR rises from 32.51 to 33.11. This scaling analysis reveals that multimodal tokenization has a capacity requirement: small models suffer from interference while large models benefit from cross-modal learning. Additional ablations and extensive reconstruction visualizations across all modalities are provided in Section C.4.

4.3 MULTIMODAL LLMS

To validate ATOKEN's effectiveness for vision-language understanding, we integrate it into SlowFast-LLaVA-1.5 (Xu et al., 2025), replacing the Oryx-ViT (Liu et al., 2024b) vision encoder with ATOKEN-So/C while freezing ATOKEN parameters during training.

Image Understanding. Table 2 shows results on 7 standard benchmarks including RW-QA, AI2D (Kembhavi et al., 2016), SQA (Lu et al., 2022b), MMMU (Yue et al., 2024), MathVISTA (Lu et al., 2024b), OCRBench (Liu et al., 2024a), and TextVQA (Singh et al., 2019). SlowFast-LLaVA-1.5 with ATOKEN outperforms Oryx-ViT across model scales, with the 7B model achieving gains of 1.3% on RW-QA, 1.0% on SQA, and 1.3% on TextVQA. The 3B model achieves superior results on almost all benchmarks, demonstrating strong generalization ability.

Video Understanding. Table 3 covers video tasks including Video-MME (Fu et al., 2024), PercepTest (Pătrăucean et al., 2023), NExT-QA (Xiao et al., 2021), and long-video benchmarks LongVideoBench (Wu et al., 2025b), MLVU (Zhou et al., 2024b), and LVBench (Wang et al.,

Table 4: **Continuous tokenizers** on ImageNet.

Tokenizer	CFG	gFID↓	IS↑	Pre.↑	Rec.↑
DiT	1.5	2.27	278.2	**0.83**	0.57
REPA	1.35	1.42	**305.7**	0.80	**0.65**
VAVAE	6.7†	**1.35**	295.3	0.79	**0.65**
ATOKEN-So/C					
Stage 1	1.5	1.62	253.3	0.78	0.63
Stage 2	1.65	1.88	231.1	0.80	0.60
Stage 3	1.65	1.56	260.0	0.79	0.63

Table 5: **Discrete tokenizers** on ImageNet.

Tokenizer	CFG	gFID↓	IS↑	Pre.↑
LFQ	-	1.91	**324.3**	-
TikTok-L	-	6.18	182.1	0.80
VQGAN	1.75	2.34	253.9	0.81
UniTok	1	2.51	216.7	**0.82**
TokenBridge	3.1	**1.76**	294.8	0.80
ATOKEN-So/D	3.1	2.23	274.5	0.79

2024c). ATOKEN excels at smaller scales, with the 1.5B model achieving state-of-the-art performance on most benchmarks. It provides strong gains on general video QA, achieving 64.5% on VideoMME and 70.3% on PercepTest with 7B models. While Oryx-ViT shows advantages on long-form understanding (particularly MLVU), likely due to its video-specific design, ATOKEN demonstrates competitive unified performance across modalities.

4.4 IMAGE GENERATION WITH CONTINUOUS & DISCRETE TOKENS

Continuous Tokens. We evaluate continuous token generation using Lightning-DiT (Yao & Wang, 2025), comparing against diffusion methods (DiT (Peebles & Xie, 2022), SiT (Ma et al., 2024a)) and reconstruction-specialized approaches (REPA (Yu et al., 2024b), VAVAE (Yao & Wang, 2025)). For fair comparison with VAVAE, we use identical training code, adapting only for ATOKEN's 48-dimensional latents (vs. 32). Following Lightning-DiT protocols, we apply CFG scale 1.65 across all channels. As shown in Table 4, ATOKEN-So/C achieves 1.56 gFID, competitive with VAVAE (1.35) and REPA (1.42) despite optimizing for multiple modalities. The So model improves from Stage 2 to Stage 3 (1.88→1.56 gFID), suggesting multimodal training can enhance generation quality when given sufficient capacity.

Discrete Tokens. We integrate ATOKEN-So/D into the TokenBridge (Wang et al., 2025) autoregressive framework, replacing only the tokenizer. Unlike TokenBridge's 16 dimensions with 8-level vocabularies, ATOKEN-So/D uses 8 dimensions with 4096-level vocabularies – a more challenging configuration requiring modeling of larger discrete spaces. As shown in Table 5, ATOKEN-So/D achieves 2.23 gFID, outperforming UniTok (2.51 gFID), the only other unified visual tokenizer. While TokenBridge achieves a lower gFID (1.76), this gap is expected given our larger vocabulary size (4096 vs. 8), demonstrating that multimodal capabilities need not compromise generation quality.

Extended Generative Applications. Our unified tokens enable diverse downstream tasks beyond images. For text-to-video (Section D.1), ATOKEN-So/C achieves 78.46% VBench score, matching specialized tokenizers (Wan2.1: 78.60%, Hunyuan 78.02%). For image-to-3D synthesis (Section D.2), we successfully generate 3D assets from single images, though our 48-dimensional latents require further optimization versus task-specific 8-channel approaches. These results validate unified tokenization as a foundation for multimodal generation (samples: Figure 12–Figure 14).

5 DISCUSSION AND CONCLUSION

The effectiveness of ATOKEN across diverse modalities and tasks suggests new opportunities: visual tokenization can achieve the same unification that transformed language modeling. Our single framework achieves both high-fidelity reconstruction and semantic understanding across images, videos, and 3D assets. This integration became possible through the combination of our sparse 4D representation, transformer-based architecture, adversarial-free training strategy, and progressive multimodal curriculum. Due to limited computational resources, we could only test ATOKENon separate downstream tasks. Building the comprehensive omnimodel that would demonstrate ATOKEN's full potential remains as future work. Looking forward, ATOKEN opens paths for visual foundation models to follow language modeling's trajectory toward true generalization. We hope this work sheds light on the next-generation multimodal AI systems built upon unified visual tokenization.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv:2303.08774*, 2023.

Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pp. 40–49. PMLR, 2018.

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv:2501.03575*, 2025.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv:2106.08254*, 2021.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv:2504.13181*, 2025.

João Carreira, Dilara Gokay, Michael King, Chuhan Zhang, Ignacio Rocco, Aravindh Mahendran, Thomas Albert Keck, Joseph Heyward, Skanda Koppula, Etienne Pot, et al. Scaling 4d representations. *arXiv:2412.15212*, 2024.

David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pp. 190–200, 2011.

Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Y. Qiao, Tong Lu, and Limin Wang. Videollm: Modeling video sequence with large language models. *ArXiv*, abs/2305.13292, 2023.

Hao Chen, Ze Wang, Xiang Li, Ximeng Sun, Fangyi Chen, Jiang Liu, Jindong Wang, Bhiksha Raj, Zicheng Liu, and Emad Barsoum. Softvq-vae: Efficient 1-dimensional continuous tokenizer. *ArXiv*, abs/2412.10958, 2024a.

Hao Chen, Yujin Han, Fangyi Chen, Xiang Li, Yidong Wang, Jindong Wang, Ze Wang, Zicheng Liu, Difan Zou, and Bhiksha Raj. Masked autoencoders are effective tokenizers for diffusion models. *ArXiv*, abs/2502.03444, 2025.

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

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13320–13331, 2024b.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36:2252–2274, 2023.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv:2505.14683*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020.

Patrick Esser, Sumith Kulal, A. Blattmann, Rahim Entezari, Jonas Muller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. *ArXiv*, abs/2403.03206, 2024.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *ArXiv*, abs/2309.17425, 2023.

Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *ArXiv*, abs/2408.14023, 2024.

Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3299–3309, 2021.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv:2405.21075*, 2024.

Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet : End-to-end video-language transformers with masked visual-token modeling. *ArXiv*, abs/2111.12681, 2021.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016.

Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, pp. 102–118. Springer, 2022.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *NeurIPS*, 2023.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv:2501.12948*, 2025.

Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv:2303.05371*, 2023.

Philippe Hansen-Estruch, David Yan, Ching-Yao Chung, Orr Zohar, Jialiang Wang, Tingbo Hou, Tao Xu, Sriram Vishwanath, Peter Vajda, and Xinlei Chen. Learnings from scaling visual tokenizers for reconstruction and generation. *arXiv:2501.09755*, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Xianglong He, Junyi Chen, Sida Peng, Di Huang, Yangguang Li, Xiaoshui Huang, Chun Yuan, Wanli Ouyang, and Tong He. Gvgen: Text-to-3d generation with volumetric representation. In *European Conference on Computer Vision*, pp. 463–479. Springer, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv:2104.08718*, 2021.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv:1207.0580*, 2012.

Jingjia Huang, Yinan Li, Jiashi Feng, Xiaoshuai Sun, and Rongrong Ji. Clover: Towards a unified video-language alignment and fusion model. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14856–14866, 2022.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024.

Ka-Hei Hui, Ruihui Li, Jingyu Hu, and Chi-Wing Fu. Neural wavelet-domain diffusion for 3d shape generation. In *SIGGRAPH Asia 2022 conference papers*, pp. 1–9, 2022.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Yang Jiao, Haibo Qiu, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3600–3610, 2025.

Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv:2305.02463*, 2023.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.

Dongwon Kim, Ju He, Qihang Yu, Chenglin Yang, Xiaohui Shen, Suha Kwak, and Liang-Chieh Chen. Democratizing text-to-image masked generative models with compact text-aware one-dimensional tokens. *ArXiv*, abs/2501.07730, 2025.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, 2013.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *NeurIPS*, 2023.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv:2412.03603*, 2024.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

Yushi Lan, Fangzhou Hong, Shuai Yang, Shangchen Zhou, Xuyi Meng, Bo Dai, Xingang Pan, and Chen Change Loy. Ln3diff: Scalable latent neural fields diffusion for speedy 3d generation. In *European Conference on Computer Vision*, pp. 112–130. Springer, 2024.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11513–11522, 2022.

Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23119–23129, 2022.

Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Jiuxiang Gu, Bhiksha Raj, and Zhe Lin. Imagefolder: Autoregressive image generation with folded tokens. *ArXiv*, abs/2410.01756, 2024a.

Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Jiuxiang Gu, Jindong Wang, Zhe Lin, and Bhiksha Raj. Xq-gan: An open-source image tokenization framework for autoregressive generation. *ArXiv*, abs/2412.01762, 2024b.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, 2023.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 2024a.

Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv:2409.12961*, 2024b.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *ArXiv*, abs/2206.08916, 2022a.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26439–26455, 2024a.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022b.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024b.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of clip for end to end video clip retrieval. *arXiv:2104.08860*, 2021.

Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2837–2845, 2021.

Tiange Luo, Justin Johnson, and Honglak Lee. View selection for 3d captioning via diffusion ranking. In *European Conference on Computer Vision*, pp. 180–197. Springer, 2024a.

Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *ArXiv*, abs/2409.04410, 2024b.

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *arXiv:2502.20321*, 2025.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024a.

Yiyang Ma, Xingchao Liu, Xi aokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Computer Vision and Pattern Recognition*, 2024b.

Fabian Mentzer, David C. Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *ArXiv*, abs/2309.15505, 2023.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4460–4470, 2019.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106, 2021.

David Mizrahi, Roman Bachmann, Ouguzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *ArXiv*, abs/2312.06647, 2023.

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv:2212.08751*, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, et al. Open-sora 2.0: Training a commercial-level video generation model in $200 k. *arXiv:2503.09642*, 2025.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv:2410.13720*, 2024.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *ArXiv*, abs/1704.00675, 2017.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *NeurIPS*, 2023.

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6964–6974, 2021.

Rui Qian, Yeqing Li, Liangzhe Yuan, Boqing Gong, Ting Liu, Matthew Brown, Serge J. Belongie, Ming-Hsuan Yang, Hartwig Adam, and Yin Cui. On temporal granularity in self-supervised video representation learning. In *British Machine Vision Conference*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Adrià Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Patraucean, Florent Altch'e, Michael Valko, Jean-Bastien Grill, Aäron van den Oord, and Andrew Zisserman. Broaden your views for self-supervised video learning. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1235–1245, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Team Seawead, Ceyuan Yang, Zhijie Lin, Yang Zhao, Shanchuan Lin, Zhibei Ma, Haoyuan Guo, Hao Chen, Lu Qi, Sen Wang, et al. Seaweed-7b: Cost-effective training of video generation foundation model. *arXiv:2504.08685*, 2025.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv:1508.07909*, 2015.

J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20875–20886, 2023.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tianbo Ye, Yang Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232, 2023.

Stefan Stojanov, Anh Thai, and James M Rehg. Using shape to categorize: Low-shot learning with an explicit shape bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1798–1808, 2021a.

Stefan Stojanov, Anh Thai, and James M. Rehg. Using shape to categorize: Low-shot learning with an explicit shape bias. 2021b.

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

Chameleon Team and Jacob Kahn. Chameleon: Mixed-modal early-fusion foundation models. *ArXiv*, abs/2405.09818, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv:2312.11805*, 2023.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv:2403.05530*, 2024.

Rui Tian, Mingfei Gao, Mingze Xu, Jiaming Hu, Jiasen Lu, Zuxuan Wu, Yinfei Yang, and Afshin Dehghan. Unigen: Enhanced training & test-time strategies for unified multimodal understanding and generation. *arXiv:2505.14682*, 2025.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *ArXiv*, abs/2203.12602, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023.

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

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv:2210.02399*, 2022.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv:2503.20314*, 2025.

Hanyu Wang, Saksham Suri, Yixuan Ren, Hao Chen, and Abhinav Shrivastava. Larp: Tokenizing videos with a learned autoregressive generative prior. *arXiv:2410.21264*, 2024a.

Junke Wang, Yi Jiang, Zehuan Yuan, Bingyue Peng, Zuxuan Wu, and Yu-Gang Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. *Advances in Neural Information Processing Systems*, 37: 28281–28295, 2024b.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, 2022a.

Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4563–4573, 2023.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv:2406.08035*, 2024c.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *ArXiv*, abs/2212.03191, 2022b.

Yuqing Wang, Zhijie Lin, Yao Teng, Yuanzhi Zhu, Shuhuai Ren, Jiashi Feng, and Xihui Liu. Bridging continuous and discrete tokens for autoregressive visual generation. *arXiv:2503.16430*, 2025.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv:2108.10904*, 2021.

Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report, 2025a.

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

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *NeurIPS*, 2025b.

Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv:2406.05127*, 2024b.

Weijia Wu, Yuzhong Zhao, Zhuang Li, Jiahong Li, Hong Zhou, Mike Zheng Shou, and Xiang Bai. A large cross-modal video retrieval dataset with reading comprehension. *Pattern Recognition*, 157:110818, 2025c.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv:2409.04429*, 2024c.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv:2412.01506*, 2024.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021.

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

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *ArXiv*, abs/2506.15564, 2025.

Bojun Xiong, Si-Tong Wei, Xin-Yang Zheng, Yan-Pei Cao, Zhouhui Lian, and Peng-Shuai Wang. Octfusion: Octree-based diffusion models for 3d shape generation. *arXiv:2408.14732*, 2024.

Tianwei Xiong, Jun Hao Liew, Zilong Huang, Jiashi Feng, and Xihui Liu. Gigatok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation. *arXiv:2504.08736*, 2025.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao (Bernie) Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke S. Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *ArXiv*, abs/2309.16671, 2023.

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

Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models. *arXiv:2407.15841*, 2024.

Mingze Xu, Mingfei Gao, Shiyu Li, Jiasen Lu, Zhe Gan, Zhengfeng Lai, Meng Cao, Kai Kang, Yinfei Yang, and Afshin Dehghan. Slowfast-llava-1.5: A family of token-efficient video large language models for long-form video understanding. *arXiv:2503.18943*, 2025.

Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos. *ArXiv*, abs/2408.10188, 2024.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv:2104.10157*, 2021.

Wilson Yan, Matei Zaharia, Volodymyr Mnih, Pieter Abbeel, Aleksandra Faust, and Hao Liu. Elastictok: Adaptive tokenization for image and video. *ArXiv*, abs/2410.08368, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. *ArXiv*, abs/2408.06072, 2024.

Jingfeng Yao and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. *ArXiv*, abs/2501.01423, 2025.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv:2110.04627*, 2021.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv:2205.01917*, 2022a.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. Magvit: Masked generative video transformer. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10459–10469, 2022b.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10459–10469, 2023.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37: 128940–128966, 2024a.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024b.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16354–16366, 2022.

Kaiwen Zha, Lijun Yu, Alireza Fathi, David A. Ross, Cordelia Schmid, Dina Katabi, and Xiuye Gu. Language-guided image tokenization for generation. *ArXiv*, abs/2412.05796, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Long Zhao, Nitesh B Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, et al. Videoprism: A foundational visual encoder for video understanding. 2024.

Chuanxia Zheng, Long Tung Vuong, Jianfei Cai, and Dinh Q. Phung. Movq: Modulating quantized vectors for high-fidelity image generation. *ArXiv*, abs/2209.09002, 2022.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke S. Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *ArXiv*, abs/2408.11039, 2024a.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv:2406.04264*, 2024b.