AToken: A Unified Tokenizer for Vision

Jiasen Lu* **Liangchen Song*** **Mingze Xu** **Byeongjoo Ahn**
Yanjun Wang **Chen Chen** **Afshin Dehghan** **Yinfei Yang**
Apple

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 *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 28.3 PSNR with 90.9% accuracy for 3D. In downstream applications,* AToken *enables visual generation tasks and understanding tasks, achieving competitive performance across all benchmarks. These results shed light on next-generation multimodal AI systems built upon unified visual tokenization.*

1. Introduction

Large Language Models (LLMs) [1, 7, 17, 42, 43] 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 [39], 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.

*Leading authors, equal contribution.

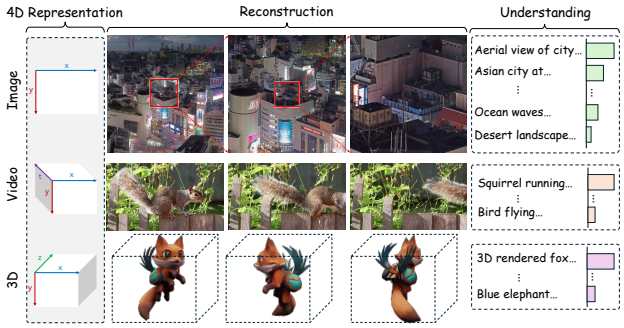

Figure 1. **AToken on images, videos, and 3D.** Our method uses a shared 4D latent space (left) to produce high-fidelity reconstructions (middle: zoomed regions for images, temporal frames for videos, viewpoints for 3D) while preserving semantic understanding (right: zero-shot text retrieval).

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) [2, 31, 32]. 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 [5, 36, 67] achieve semantic alignment but lack pixel-level detail, while VAE-based tokenizers [11, 34, 37, 63] preserve visual details but lack semantic understanding. Second, architectural choices create different limitations: convolutional tokenizers plateau when scaling model parameters [58], while transformer tokenizers [18, 46, 62] achieve better scaling but suffer from severe adversarial training instabilities. Third, recent unification

efforts remain limited to images [9, 28, 55], 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 [56]. 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, but existing approaches remain fragmented across modalities and objectives, lacking language models' versatility. Due to space limitations, we defer comprehensive discussion to the Appendix.

Task Specialization. Current visual tokenizers fall into two distinct categories: reconstruction methods (SD-VAE [37], VQGAN [11], GigaTok [58], Cosmos [3]) excel at compression for generation but cannot extract semantic features; understanding encoders (CLIP [36], SigLIP2 [44], VideoPrism [69]) produce rich semantics but cannot reconstruct content. Only VILA-U [55] and UniTok [28] 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 [34], Hunyuan [20]), they cannot process 3D data. Conversely, 3D tokenizers like Trellis-SLAT [56] 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 [49, 69] 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 [37]), with recent hybrid (GigaTok [58]) and pure transformer (ViTok [18]) approaches, the latter suffering from adversarial training instabilities. *(2) Token representation:* Methods choose discrete tokens for LLM compatibility (VQGAN [11]) or continuous tokens for reconstruction quality (TAE [34]), with few supporting both. *(3) Resolution handling:* Convolutions naturally handle arbitrary resolutions, while only SigLIP2 [44] 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 (Sec. 3.1), the transformer architecture processing these representations (Sec. 3.2), adversarial-free training objectives (Sec. 3.3), and a progressive curriculum for multimodal learning (Sec. 3.4), followed by implementation details (Sec. 3.5).

3.1. Unified Latent Representation

Unified Modalities – Image, Video and 3D. Our central insight is that all visual modalities can be represented within

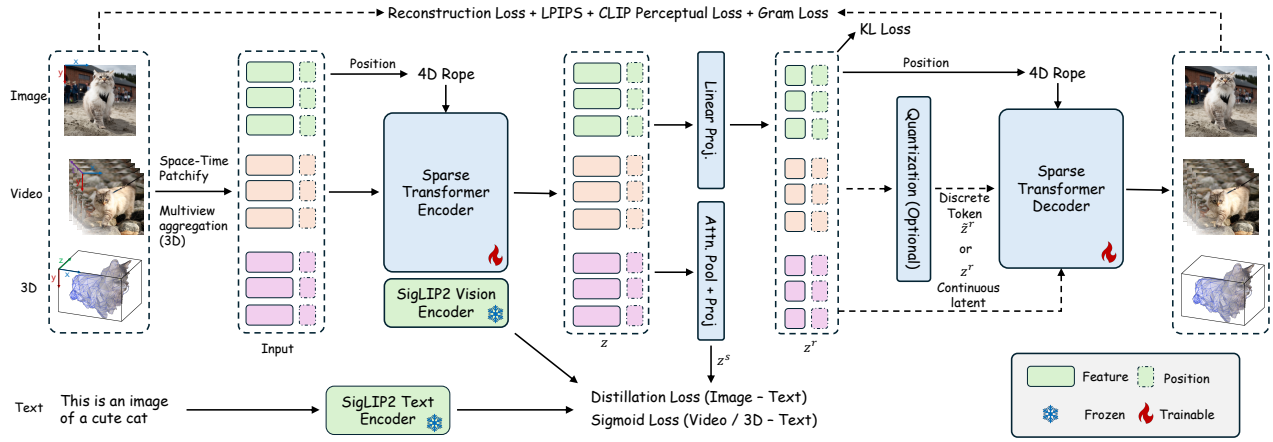

Figure 2. **Overview of our method.** All modalities undergo unified space-time patchification and encoding into sparse 4D latents, which support reconstruction through modality-specific decoders and understanding through attention pooling and text alignment. The architecture optimizes reconstruction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

a shared 4D space. As illustrated in Fig. 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 [56] by rendering multi-view images from spherically sampled cameras, applying our unified patchification, then aggregating features into voxel space (detailed in Sec. 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 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 [37], optionally applying FSQ [30] for discrete codes $\tilde{\boldsymbol{z}}^r = \mathrm{FSQ}(\boldsymbol{z}^r)$. The decoder $\mathcal{D}_\theta$ reconstructs the input from these latents. For understanding, we aggregate latents via attention pooling [36, 44] 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.

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 [56] to our unified pipeline. As shown 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 [56], 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 Fig. 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 [44] 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 image features. Second, we augment learnable 2D position embeddings with 4D RoPE

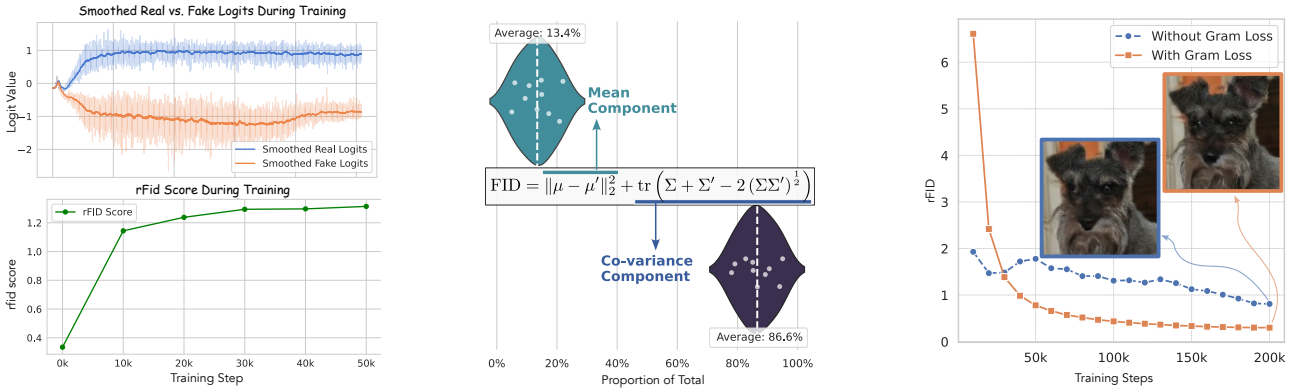

Figure 3. **Adversarial-free training with Gram loss achieves stable, high-fidelity reconstruction.** (a) GAN training fails as the discriminator overpowers the generator, degrading rFID. (b) rFID decomposition shows 86.6% of error from covariance (texture/style) vs. 13.4% from mean. (c) Gram loss directly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

[24] 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:

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

treating images as single-frame videos ($T = 1$) and discarding temporal padding following [34]. For 3D assets, we first decode to pixel-space features, then apply an additional layer to generate Gaussian splatting parameters for efficient rendering:

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

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 [56], 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.

3.3. Training Objectives

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

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

where $\mathcal{L}_{\mathrm{KL}}$ 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 [52] and incompatible with sparse 3D representations.

Reconstruction Loss. While GANs [16] are standard for visual tokenizers, we found them unsuitable for our transformer architecture. Fig. 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 (Fig. 3(b)). The covariance component – capturing second-order statistics like texture and style – dominates at 86.6%, while the mean contributes only 13.4%. This motivated adopting Gram matrix loss [15], 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 Fig. 3(c), this achieves superior performance throughout training. For images, we combine four complementary loss components:

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

where $\mathcal{L}_1 = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_1$ provides pixel supervision, $\mathcal{L}_{\mathrm{LPIPS}}$ [68] measures perceptual similarity, $\mathcal{L}_{\mathrm{GRAM}}$ captures texture, and $\mathcal{L}_{\mathrm{CLIP}}$ enforces semantic consistency. For video and 3D assets, we use $\mathcal{L}_{\mathrm{rec}}^{\mathrm{V/3D}} = \mathcal{L}_1$ for efficiency, relying on cross-modal transfer from images for details:

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

where $\mathcal{L}_{\mathrm{KL}}$ 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 [52] 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

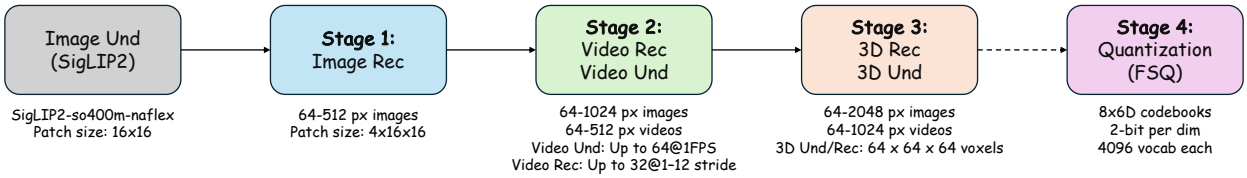

Figure 4. **Progressive training curriculum of AToken.** Our model starts from SigLIP2 image understanding and progressively adds: (1) image reconstruction, (2) video capabilities with temporal modeling, (3) 3D understanding with expanded resolutions, and (4) discrete tokenization via FSQ. Each box shows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

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 optimize alignment using the sigmoid loss from SigLIP [67], which proves more stable for smaller batch sizes in these domains. This dual strategy preserves pretrained image semantics while enabling efficient learning for new modalities.

3.4. Training Strategy

Our training employs a four-stage curriculum (Fig. 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. Training uses variable resolution from 64 to 512 pixels.

Stage 2: Video Dynamics. We extend to temporal sequences, expanding latent dimensions to 48 for motion complexity [38]. Resolution increases to 1024 for images and 512 for videos. We employ temporal tiling with adaptive sampling and KV-caching as illustrate in Fig. 5 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 [30], partitioning 48-dimensional latents into 8 discrete tokens from 4096-entry codebooks, enabling compatibility with discrete generative models across domains.

See Appendix for complete training configurations.

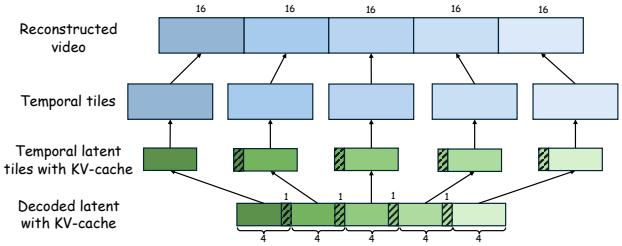

Figure 5. **Overview of the video decoding process.** During decoding, we use KV-caching across temporal tiles to eliminate redundant computation while maintaining temporal coherence.

3.5. Implementation Details

Our encoder and decoder each contain 27 transformer blocks with hidden dimension $d = 1152$ and 16 attention heads. The encoder is initialized from SigLIP-SO400M-patch16-naflex [44], while the decoder is trained from scratch. We optimize using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight decay 0.1. The learning rate follows linear warmup for 2,000 steps to $\eta_{\max} = 3\times10^{-4}$, then cosine annealing to $\eta_{\min} = 3 \times 10^{-5}$. Given the pretrained encoder, we apply a learning rate $\eta_{\text{encoder}} = 0.1 \times \eta_{\text{base}}$ and use exponential moving average with decay rate $\gamma = 0.9999$.

Training utilizes 256 H100 GPUs with adaptive global batch sizes optimized for each task's memory requirements. Image understanding maintains 8,192 samples throughout all stages, while reconstruction tasks scale with complexity: image reconstruction uses 1,024-4,096, video reconstruction uses 512, and 3D reconstruction uses 256-512. Stage 1 uses 64 GPUs, expanding to the full 256 GPUs for Stages 2-3 as modalities are added. The four-stage curriculum trains for 200k, 200k, 50k, and 100k iterations, respectively, with each stage initialized from the previous checkpoint.

Throughout training, we maintain fixed loss coefficients: $\lambda_{\text{rec}} = 0.2$, $\lambda_{\text{sem}} = 1.0$, and $\lambda_{\text{KL}} = 10^{-8}$. Within reconstruction (Eq. 5), we set $\lambda_1 = 1.0$, $\lambda_{\text{LPIPS}} = 10.0$, $\lambda_{\text{GRAM}} = 10^3$, $\lambda_{\text{CLIP}} = 1.0$, and $\tau = 2.0$. We normalize reconstruction losses over patches rather than summing

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. ATOKEN is the only tokenizer supporting all three modalities. 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 [37]	(1, 8, 8)	4	VAE	26.26	0.61	-	-	-	-	-	-	-
VA-VAE [61]	(1, 16, 16)	32	VAE	27.70	0.28	-	-	-	-	-	-	-
GigaTok-XL-XXL [58]	(1, 16, 16)	8	VQ	22.42	0.80	-	-	-	-	-	-	-
Cosmos-0.1-CV8×8 [3]	(4, 8, 8)	16	AE	30.11	7.55	-	34.33	8.34	-	-	-	-
Wan2.2 [45]	(4, 16, 16)	48	VAE	31.25	0.75	-	**36.39**	3.19	-	-	-	-
OmniTokenizer [46]	(4, 8, 8)	8	VQ	24.69	1.41	-	19.89	202.46	-	-	-	-
Cosmos-0.1-DV8×8 [3]	(4, 8, 8)	6	FSQ	26.34	7.86	-	31.42	25.94	-	-	-	-
Trellis-SLAT [56]	-	8	VAE	-	-	-	-	-	-	26.97	**0.054**	-
Understanding Only												
SigLIP2-So/16 [44]	(1, 16, 16)	-	-	-	-	83.4	-	-	41.9	-	-	-
PE_{coreL} [5]	(1, 14, 14)	-	-	-	-	**83.5**	-	-	**50.3**	-	-	-
Reconstruction & Understanding												
VILA-U [55]	(1, 16, 16)	16	RQ	22.24	4.23	78.0	-	-	-	-	-	-
UniTok [29]	(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												
Stage 1	(4, 16, 16)	32	VAE	28.77	0.26	82.3	-	-	-	-	-	-
Stage 2	(4, 16, 16)	48	VAE	29.55	0.25	82.2	35.63	3.63	40.1	-	-	-
Stage 3	(4, 16, 16)	48	VAE	29.72	**0.21**	82.2	36.07	**3.01**	40.2	**28.28**	0.062	**90.9**

[11], providing stable gradients across resolutions.

Training data follows our progressive curriculum: DFN [13], Open Images [21], and internal datasets for images; WebVid [4] and TextVR [54] for video understanding with Panda70M [6] for reconstruction; Objaverse [8] with Cap3D [27] for 3D. Datasets are sampled proportionally to their size, with task ratios detailed in Appendix.

4. Results

We evaluate ATOKEN as the first visual tokenizer to achieve reconstruction and understanding across images, videos, and 3D assets. Our experiments show that unified tokenization achieves competitive performance across all modalities (Sec. 4.1), reveals insights about model scaling and cross-modal benefits (Sec. 4.2), integrates seamlessly into existing pipelines (Sec. 4.3), and enables high-quality generation (Sec. 4.4). The Appendix provides detailed per-modality evaluations and downstream applications.

4.1. Unified Tokenizer Evaluation

Tab. 1 compares visual tokenizers across modalities using ImageNet [10], TokenBench [3], MSR-VTT [59] for video, and Toys4k [41] for 3D. Existing approaches fall into three limited categories: reconstruction-only tokenizers [37, 45, 56] excel at generation but lack semantics; understanding-only encoders [5, 44] provide semantics but cannot reconstruct; recent unified attempts [28, 55] combine both but remain image-only.

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 [45] and Trellis-SLAT. Our discrete variant (ATOKEN-So/D) maintains competitive performance, pioneering discrete tokenization across all modalities.

The progressive training stages reveal unexpected cross-modal synergies. Starting from Stage 1 with 28.77 PSNR and 0.26 rFID, ATOKEN-So/C improves to 29.55 PSNR and 0.25 rFID when video is added in Stage 2, and further to 29.72 PSNR and 0.21 rFID with 3D integration in Stage 3. This 19% reduction in rFID (0.26→0.21) challenges the conventional wisdom that unified models must sacrifice quality for generality. Most strikingly, video reconstruction improves from 35.63 to 36.07 PSNR when 3D is added in Stage 3, with rFVD dropping 17%, suggesting that geometric understanding from 3D data provides valuable inductive biases for temporal modeling. The progressive architecture expansion from 32 to 48 latent channels between Stages 1 and 2 accommodates these complementary signals.

Fig. 6 shows the qualitative examples of image and video reconstruction. Due to space constraints, detailed evaluations with comprehensive baselines on additional benchmarks for all three modalities are provided in Appendix.

4.2. Scaling and Cross-Modal Benefits

To understand why multimodal training enhances rather than degrades performance, we investigate the role of model capacity by comparing So400m model (800M) with a

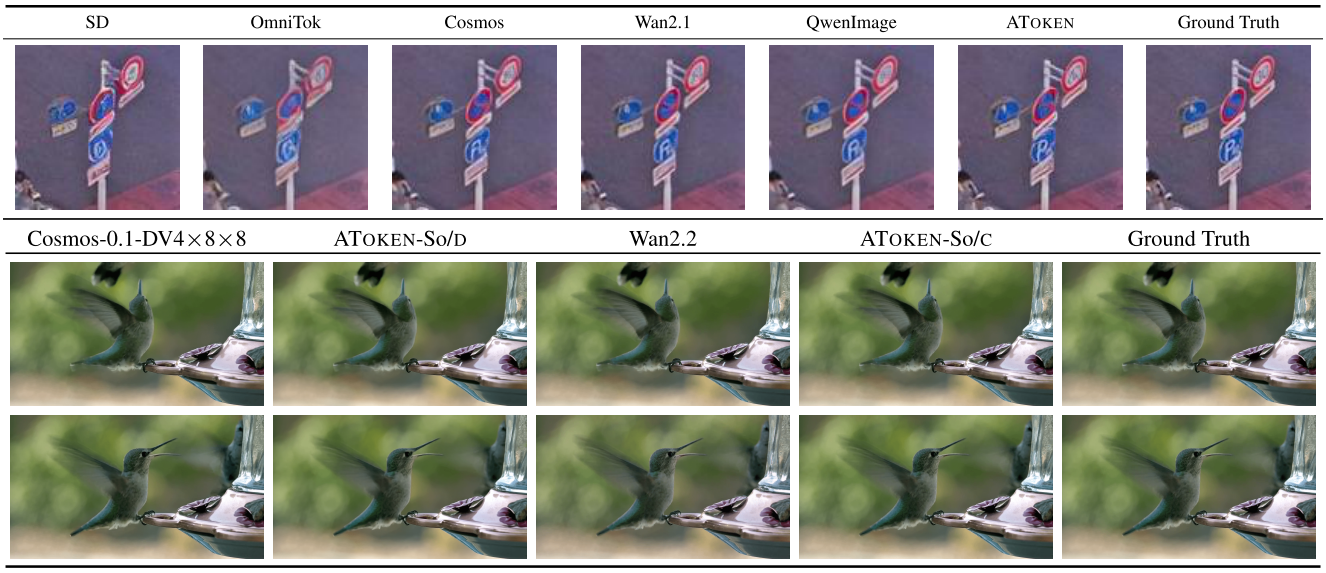

Figure 6. **Qualitative comparison of image and video reconstruction performance.** Despite operating at a higher compression ratio, ATOKEN outperforms SoTA methods in handling high-frequency textures and complex text.

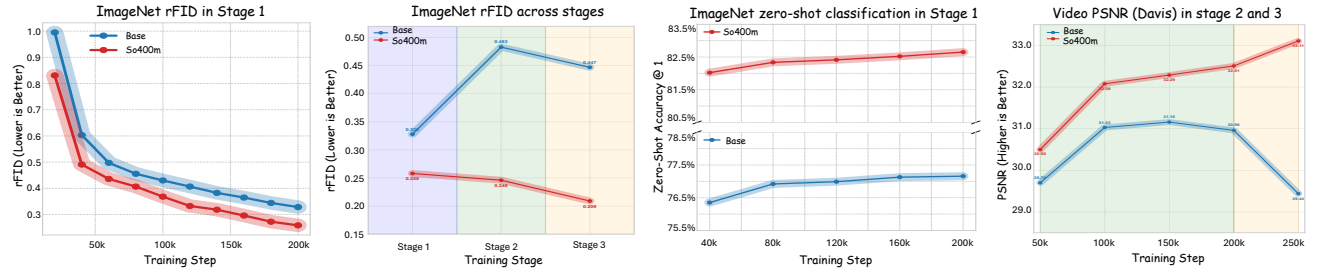

Figure 7. **Architectural scaling comparison: Base vs. So400m models.** (a) ImageNet rFID during Stage 1 training. (b) ImageNet rFID across training stages. (c) ImageNet zero-shot classification accuracy in Stage 1. (d) Video PSNR on DAVIS in Stages 2 and 3. The So400m model maintains or improves performance 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.

smaller Base variant (192M) following identical training procedures. The Base model uses the same architecture initialized from SigLIP-Base-patch16-naflex [44].

Fig. 7 reveals a critical capacity threshold for multimodal tokenization. While both So400m (800M) and Base (192M) models achieve reasonable single-modal performance in Stage 1, their trajectories diverge dramatically with multimodal expansion. The Base model suffers catastrophic interference – ImageNet rFID deteriorates 49% (0.323→0.483) while video PSNR declines monotonically. This demonstrates that multimodal tokenization requires sufficient capacity: below ∼200M, modalities compete destructively for representation space; at 800M, they leverage the complementary signals described above.

4.3. Multimodal LLMs

To validate ATOKEN's effectiveness for multimodal LLMs, we integrate it into SlowFast-LLaVA-1.5 [60], replacing the Oryx-ViT [23] vision encoder with ATOKEN-So/C.

Tab. 2 shows consistent improvements across 7 image understanding benchmarks. The 7B SlowFast-LLaVA with ATOKEN outperforms the Oryx-ViT baseline on all tasks, with notable gains on question-answering (1.3% on RW-QA, 1.3% on TextVQA) and reasoning benchmarks (1.0% on SQA). These improvements demonstrate strong generalization despite ATOKEN being frozen during training.

For video understanding (Tab. 3), ATOKEN achieves competitive performance across both general video QA and long-form tasks. We observe gains on VideoMME (64.5% vs 63.9%) and PercepTest (70.3% vs 69.6%), surpassing specialized models like Qwen2-VL-7B. While Oryx-ViT shows advantages on certain long-form benchmarks (MLVU, LongVideoBench), likely due to video-specific optimizations, ATOKEN provides strong unified performance across modalities without specialized engineering. Results for additional model scales are provided in the Appendix.

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 MLLMs. Additional comparisons are provided in Appendix.

			General & Knowledge					TextRich	
Multimodal LLM	**Vision Encoder**	**# Input Pixels**	**RW-QA (test)**	**AI2D [19] (test)**	**SQA [25] (test)**	**MMMU [66] (val)**	**MathV [26] (testmini)**	**OCRBench [22] (test)**	**TextVQA [40] (val)**
InternVL2.5-8B [50]	InternViT	9.63M	70.1	84.5	-	56.0	64.4	-	79.1
Qwen2-VL-7B [47]	DFN	-	70.1	83.0	-	54.1	58.2	-	84.3
SlowFast-LLaVA-1.5-7B [60]	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. Additional comparisons are provided in Appendix.

			General VideoQA			Long-Form Video Understanding		
Multimodal LLM	**Vision Encoder**	**# Input Tokens**	**VideoMME [14] (w/o sub)**	**PercepTest [35] (val)**	**NExT-QA [57] (test)**	**LongVideoBench [53] (val)**	**MLVU [70] (m-avg)**	**LVBench [48] (avg)**
InternVL2.5-8B [50]	InternViT	16K	64.2	-	85.0	60.0	69.0	43.2
Qwen2-VL-7B [47]	DFN	16K	63.3	62.3	81.2	55.6	69.8	44.7
SlowFast-LLaVA-1.5-7B [60]	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

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

Tokenizer	**CFG**	**gFID↓**	**IS↑**	**Pre.↑**	**Rec.↑**
DiT [33]	1.5	2.27	278.2	**0.83**	0.57
REPA [65]	1.35	1.42	**305.7**	0.80	**0.65**
VAVAE [61]	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.↑**
TikTok-L [64]	-	6.18	182.1	0.80
VQGAN [62]	1.75	2.34	253.9	0.81
UniTok [29]	1	2.51	216.7	**0.82**
TokenBridge [51]	3.1	**1.76**	294.8	0.80
ATOKEN-So/D	3.1	2.23	274.5	0.79

4.4. Multimodal Generation

Image Generation with Continuous Tokens. We evaluate continuous token generation using Lightning-DiT [61], comparing against diffusion methods (DiT [33]) and reconstruction-specialized approaches (REPA [65], VAVAE [61]). Using identical training code to VAVAE for fair comparison. As shown in Tab. 4, ATOKEN-So/C Stage 3 achieves 1.56 gFID, competitive with VAVAE (1.35) and REPA (1.42) despite optimizing for multiple modalities.

Image Generation with Discrete Tokens. We integrate ATOKEN-So/D into TokenBridge [51], 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. Tab. 5 shows ATOKEN-So/D achieves 2.23 gFID, outperforming UniTok (2.51), the only other unified visual tokenizer, while approaching TokenBridge (1.76) despite our larger vocabulary size.

Video and 3D Generation. Beyond images, ATOKEN enables diverse generative applications. For text-to-video, we integrate ATOKEN-So/C into an MMDiT-based model [12], achieving 78.46% on VBench – matching specialized video tokenizers like Wan2.1 (78.60%) and Hunyuan (78.02%) despite being designed for multiple modalities. For image-to-3D synthesis using Trellis-SLAT [56], we successfully generate 3D assets from single images, though our 48-dimensional latents require further optimization compared to task-specific 8-channel approaches. Full comparisons and detailed analysis are provided in the Appendix. These results demonstrate that unified tokenization provides a strong foundation for multimodal generation without modality-specific engineering.

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 generalization. We hope this work sheds light on the next-generation multimodal AI systems built upon unified visual tokenization.

References

[1] 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. 1

[2] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018. 1

[3] 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. 2, 6

[4] 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. 6

[5] 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. 1, 6

[6] 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*, pages 13320–13331, 2024. 6

[7] 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. 1

[8] 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. 6

[9] 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. 2

[10] 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*, pages 248–255. Ieee, 2009. 6

[11] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020. 1, 2, 6

[12] 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. 8

[13] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *ArXiv*, abs/2309.17425, 2023. 6

[14] 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. 8

[15] 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)*, pages 2414–2423, 2016. 4

[16] 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. 4

[17] 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. 1

[18] 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. 1, 2

[19] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016. 8

[20] 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. 2

[21] 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. 6

[22] 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*, 2024. 8

[23] 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*, 2024. 7

[24] 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*, pages 26439–26455, 2024. 4

[25] 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*, 2022. 8

[26] 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*, 2024. 8

[27] Tiange Luo, Justin Johnson, and Honglak Lee. View selection for 3d captioning via diffusion ranking. In *European Conference on Computer Vision*, pages 180–197. Springer, 2024. 6

[28] 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. 2, 6

[29] 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. 6, 8

[30] Fabian Mentzer, David C. Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *ArXiv*, abs/2309.15505, 2023. 3, 5

[31] 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*, pages 4460–4470, 2019. 1

[32] 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. 1

[33] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 8

[34] 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. 1, 2, 4

[35] 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. 8

[36] 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*, pages 8748–8763. PmLR, 2021. 1, 2, 3

[37] 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*, pages 10684–10695, 2022. 1, 2, 3, 6

[38] 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. 5

[39] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv:1508.07909*, 2015. 1

[40] 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. 8

[41] 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*, pages 1798–1808, 2021. 6

[42] 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. 1

[43] 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. 1

[44] 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. 2, 3, 5, 6, 7

[45] 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. 6

[46] 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, 2024. 1, 6

[47] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv:2409.12191*, 2024. 8

[48] 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*, 2024. 8

[49] 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, 2022. 2

[50] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyun Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kaiming Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2.5: Empowering video mllms with long and rich context modeling. *ArXiv*, abs/2501.12386, 2025. 8

[51] 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. 8

[52] 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, 2025. 4

[53] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *NeurIPS*, 2025. 8

[54] 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, 2025. 6

[55] 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*, 2024. 2, 6

[56] 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. 2, 3, 4, 6, 8

[57] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021. 8

[58] 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. 1, 2, 6

[59] 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)*, pages 5288–5296, 2016. 6

[60] 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. 7, 8

[61] Jingfeng Yao and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. *ArXiv*, abs/2501.01423, 2025. 6, 8

[62] 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. 1, 8

[63] 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)*, pages 10459–10469, 2022. 1

[64] 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, 2024. 8

[65] 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*, 2024. 8

[66] 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. 8

[67] 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)*, pages 11941–11952, 2023. 1, 5

[68] 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. 4

[69] 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. 2

[70] 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*, 2024. 8