Method	Encoder Arch.	Decoder Arch.	Reconstruction			Understanding			Discrete Token	Cont. Token	GAN Free	Temporal Comp.	Native Res.
			Image	Video	3D	Image	Video	3D					
Reconstruction Only													
SD-VAE	Conv	Conv	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	-
VQGAN	Conv	Conv	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	-
GigaTok	Hybrid	Hybrid	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
OmniTokenizer	Trans	Trans	✓	✓	✗	✗	✗	✗	✓	✓	✗	✓	✗
MAGVIT-v2	Conv	Conv	✓	✓	✗	✗	✗	✗	✓	✗	✗	✓	-
Cosmos	Conv	Conv	✓	✓	✗	✗	✗	✗	✓	✓	✗	✓	-
ViTok	Trans	Trans	✓	✓	✗	✗	✗	✗	✗	✓	✗	✓	✗
TAE	Conv	Conv	✓	✓	✗	✗	✗	✗	✗	✓	✗	✓	-
Hunyuan	Conv	Conv	✓	✓	✗	✗	✗	✗	✗	✓	✗	✓	-
Wan	Conv	Conv	✓	✓	✗	✗	✗	✗	✗	✓	✗	✓	-
Trellis-SLAT	Trans	Trans	✗	✗	✓	✗	✗	✗	✗	✓	✓	✗	-
Understanding Only													
SigLIP2	Trans	-	✗	✗	✗	✓	✓	✗	-	-	-	✗	✓
PE	Trans	-	✗	✗	✗	✓	✓	✗	-	-	-	✗	✗
VideoPrism	Trans	-	✗	✗	✗	✓	✓	✗	-	-	-	✗	✗
InternVideo	Trans	-	✗	✗	✗	✓	✓	✗	-	-	-	✓	✗
Reconstruction & Understanding													
VILA-U	Trans	Conv	✓	✗	✗	✓	✓	✗	✓	✗	✗	✗	✗
UniTok	Trans	Hybrid	✓	✗	✗	✓	✓	✗	✓	✗	✗	✗	✗
AToken	Trans	Trans	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 6. **Comparison between existing visual tokenizers and AToken.** We categorize methods by task capabilities (reconstruction, understanding, or both) and evaluate their modality coverage, architectural choices, token representations, and key features. AToken is the only method providing support across all dimensions.

The appendix includes the following sections:

- Sec A - Related Work
- Sec B - Model Details & Training Strategies
- Sec C - Additional Tokenizer Results
- Sec D - Additional Downstream Results

A. Related Work

Reconstruction Tokenizers. High-resolution images have been compressed using deep auto-encoders [40, 106], which learn lower-dimensional latent representations for reconstruction. VAEs [49] extended this framework with probabilistic modeling, while VQ-VAE [104] introduced vector quantization to discretize the latent space. Building on these foundations, subsequent works enhanced reconstruction quality through adversarial training [22, 88], developed alternative quantization strategies [54, 69, 73, 157], incorporated semantic guidance [11, 12, 48, 57, 58, 140, 149], and scaled model capacity [130].

Video tokenization extended these image-based methods to temporal domains, employing 3D convolutions [31, 137, 144], decoupled spatial-temporal processing [81], and causal modeling [51, 107, 139]. Beyond convolutional architectures, recent work has explored Vision Transformers [21] as an alternative backbone for both image [36, 141, 145] and video [105, 108, 109, 138] tokenization.

3D generation methods initially applied diffusion models directly to various 3D representations [38, 43, 67, 91, 112], then shifted toward compact latent spaces for improved efficiency [35, 46, 53, 77, 129]. Notably, Trellis [125] introduces structured latents (SLAT) that jointly encode geometry and appearance on sparse 3D grids, enabling flexible decoding to multiple output formats.

Visual Encoders. Image encoders initially leveraged contrastive learning through vision-language alignment [44, 86, 150] and image-only self-supervision [13, 78]. Generative pretraining explored text generation objectives [117], discrete token reconstruction [6], and masked image modeling [8, 37]. Methods like NaViT [16] introduced resolution flexibility with preserved aspect ratios. Recent unified approaches merge contrastive, generative, and self-supervised objectives [103, 142] or leverage intermediate-layer features with task-specific alignment [7].

Video encoders primarily employ self-supervised learning on video-only data [26, 84, 85, 87, 101] or video-language modeling with noisy text supervision [10, 28, 41, 56, 148]. Recent methods treat video as image sequences, focusing on context window expansion [99, 135] or token compression [25, 59, 93, 118, 133].

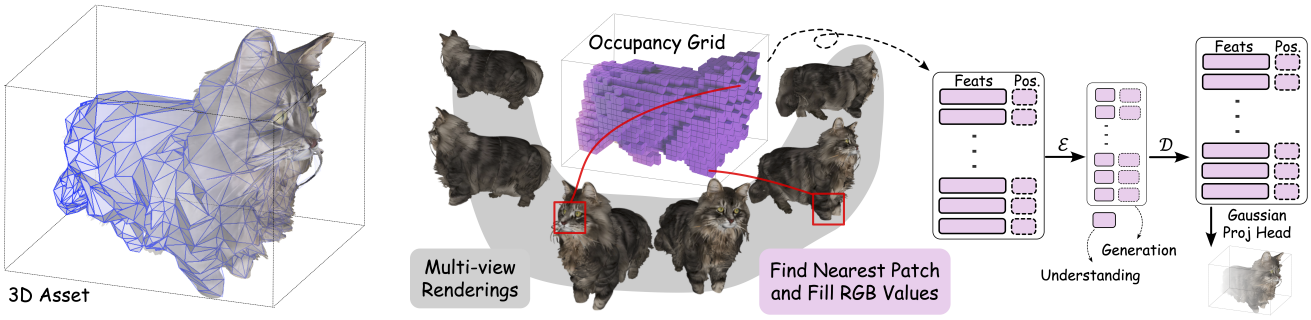

Figure 8. **3D tokenization pipeline.** We extend Trellis-SLAT [125] for multimodal unification through two modifications: directly tokenizing raw RGB patches from multiview renderings (as opposed to using DINOv2 features), and aggregating each voxel's features from its nearest viewpoint (as opposed to averaging across all views). Combined with Gaussian decoding, this approach integrates 3D assets into our unified token space alongside images and videos.

Unified Tokenizers & Multimodal Models. Unified Multimodal Models aim to combine visual understanding and generation within a single framework [63, 76, 110]. Many approaches use decoupled tokenizers while employing various generation paradigms – autoregressive [62, 97, 119], diffusion [158], flow-matching [72], and masked prediction [100, 127]. Recent efforts on unified tokenizers that handle both tasks include VILA-U [124], which combines pixel reconstruction with contrastive learning in a single vision tower; SeTok [122], which groups visual features into semantic units; UniTok [70], which uses multi-codebook quantization for enhanced expressiveness; and UniToken [45], which produces hybrid discrete-continuous representations through dual encoders. Show-o2 [128] extends these approaches by leveraging a 3D causal VAE space with dual-path spatial-temporal fusion, enabling scalability across both image and video modalities while combining autoregressive modeling with flow matching.

B. Model Details & Training Strategies

B.1. 3D Asset Processing Details

Our 3D tokenization pipeline (Fig. 8) extends Trellis-SLAT [125] with two key modifications for multimodal unification. First, we directly tokenize raw RGB patches from multiview renderings rather than using DINOv2 features, enabling end-to-end optimization. Second, we aggregate each voxel's features from its nearest viewpoint rather than averaging across all views, preserving local geometric details.

The pipeline renders 12 views from spherically sampled cameras, applies standard $4 \times 16 \times 16$ patchification to each view, and back-projects features into a 64^3 voxel grid. Each voxel selects features from the closest viewpoint to resolve projection conflicts.

B.2. Video Processing with KV-Caching

Our video encoding employs temporal tiling with KV-caching to eliminate redundant computation. Long sequences are partitioned into overlapping 16-32 frame tiles, encoded to 4-8 latent frames. The caching mechanism stores key-value pairs across tiles, avoiding recomputation while maintaining temporal coherence through smooth boundary transitions.

B.3. Training Strategies

Our training employs a four-stage progressive curriculum (Fig. 4 in the main paper) that builds from image foundations to video dynamics to 3D geometry, with optional discrete quantization. Starting from the pretrained SigLIP2 encoder [103], we gradually introduce more complex objectives and modalities while maintaining semantic understanding across all stages.

We implement this curriculum through round-robin sampling of modalities and tasks, using gradient accumulation to balance image-text distillation with other objectives (reconstruction, video-text alignment, 3D-text alignment) across all stages. This ensures semantic alignment is preserved even as reconstruction capabilities expand. Our sparse transformer architecture facilitates this multi-modal training by separating features and positions, allowing each modality to be processed at its natural resolution without padding or packing.

Stage 1: Image Foundation. Starting from pretrained SigLIP2, we establish core visual representations by adding image reconstruction capabilities. We process images using $4 \times 16 \times 16$ space-time patches with temporal padding for consistency, employing 32 latent dimensions following [140]. Training uses variable resolution sampling from 64 to 512 pixels, with L1 loss computed at native resolution

Table 7. **Training curriculum configuration.** Resolution limits for each modality and task sampling ratios across the four training stages. Superscripts denote reconstruction (r) and understanding (u) tasks.

Training Stage	Image Res.	Video Res.	3D Size	Task Sampling Ratios					#Steps
				I^r	V^u	V^r	$3D^u$	$3D^r$	
Stage 1: Image Foundation	$[64 \rightarrow 512]$	-	-	100%	-	-	-	-	200k
Stage 2: Video Dynamics	$[64 \rightarrow 1024]$	$[64 \rightarrow 512]$	-	22.2%	11.1%	66.6%	-	-	200k
Stage 3: 3D Geometry	$[64 \rightarrow 2048]$	$[64 \rightarrow 1024]$	$[64, 64, 64]$	22.2%	11.1%	44.4%	11.1%	11.1%	50k
Stage 4: Discrete Tokenization	$[64 \rightarrow 2048]$	$[64 \rightarrow 1024]$	$[64, 64, 64]$	22.2%	11.1%	44.4%	11.1%	11.1%	100k

while perceptual losses ($\mathcal{L}_{\text{LPIPS}}$, $\mathcal{L}_{\text{CLIP}}$, $\mathcal{L}_{\text{Gram}}$) use 224×224 interpolation to match their pretrained features.

Stage 2: Video Dynamics. We extend to temporal sequences, expanding latent dimensions from 32 to 48 to accommodate motion complexity [89]. Resolution capabilities increase to 1024 for images and 512 for videos. We employ temporal tiling (16-32 frames $\rightarrow$ 4-8 latent frames) with adaptive sampling: stride 1-3 for temporal consistency or 4-12 for diversity in reconstruction, 1 FPS up to 64 frames for understanding. Our KV-caching mechanism eliminates redundant computation across tiles while maintaining temporal coherence.

Stage 3: 3D Geometry. We incorporate 3D assets as active voxels in 64^3 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 all three modalities prevents catastrophic forgetting while leveraging cross-modal learning. The geometric semantics from 3D and the temporal dynamics from video enhance image reconstruction quality.

Stage 4: Discrete Tokenization. Optionally, we add FSQ quantization [73] for discrete generation tasks. The 48-dimensional latents are partitioned into 8 groups of 6 dimensions, each quantized to 4 levels, yielding 8 discrete tokens from 4096-entry codebooks. We finetune the entire encoder and decoder to adapt all modalities to discrete tokens, enabling compatibility with discrete generative models across all visual domains.

C. Additional Tokenizer Results

C.1. Image Tokenization

We evaluate ATOKEN's image capabilities against specialized tokenizers through reconstruction quality (Tab. 8) and semantic understanding (Tab. 9) benchmarks.

Reconstruction Performance. Tab. 8 presents our comprehensive evaluation, where we re-evaluated all baseline methods using a unified protocol with official implementations to ensure fair comparison. Under this standardized evaluation protocol, we observe that multimodal training enhances rather than compromises image reconstruction. ATOKEN-So/C achieves 0.209 rFID at 16×16 compression, with progressive improvement across training stages: 0.258 (Stage 1)$\rightarrow$0.246 (Stage 2)$\rightarrow$0.209 (Stage 3), a 19% gain through multimodal expansion.

This improvement is particularly notable given three fundamental challenges in the field. First, the compression-dimension trade-off severely constrains 16×16 models: VAVAE [140] requires 32-dimensional latents to achieve 0.279 rFID, while Cosmos-CI16×16 with 16 dimensions degrades to 0.959 rFID. Second, transformer architectures consistently underperform convolutional architectures (OmniTokenizer [109] 26.74 PSNR vs. Hunyuan [51] 33.32 PSNR), explaining why most reconstruction tokenizers avoid transformers. Third, discrete tokenizers struggle with generalization – UniTok [70] degrades from 0.362 rFID on ImageNet to 3.918 on COCO, while GigaTok [130] exhibits even larger gaps.

Our approach addresses all three challenges: achieving strong performance with 48-dimensional latents at 16×16 compression, demonstrating transformer viability through adversarial-free training, and maintaining consistent quality across datasets (0.209 rFID on ImageNet, 2.026 rFID on COCO). These results suggest temporal dynamics from video and geometric understanding from 3D provide complementary signals for image reconstruction.

Semantic Understanding. Tab. 9 evaluates zero-shot classification and retrieval against leading vision encoders. While understanding-only models like CLIP [86] and its variants [24, 96, 131] optimize purely for semantic alignment, ATOKEN need to balance understanding with reconstruction across three modalities.

Despite these constraints, ATOKEN achieves 82.2% ImageNet accuracy – within 1.2% of understanding-only SigLIP2 [103] (83.4%). This narrows the gap compared to previous unified attempts like UniTok (78.6%) and VILA-U (78.0%), while uniquely extending unified capabilities to video and 3D. Across our progressive training stages, accu-

Table 8. **Image reconstruction comparison on ImageNet and COCO.** We evaluate all methods using a unified protocol with official implementations to ensure fair comparison. All images are resized and center-cropped to 256×256, with metrics computed using identical scripts. Note that our reproduced results may differ from original papers due to standardized evaluation settings, but provide consistent cross-model comparison.

Method	Comp. Ratio	Latent Size	Token Type	ImageNet				COCO			
				PSNR↑	SSIM↑	LPIPS↓	rFID↓	PSNR↑	SSIM↑	LPIPS↓	rFID↓
Continuous Latent											
SD-VAE	(1, 8, 8)	4	VAE	26.26	0.745	0.133	0.606	25.99	0.759	0.130	4.142
FLUX.1 [dev]	(1, 8, 8)	16	VAE	32.86	0.917	**0.044**	**0.176**	**32.73**	0.923	**0.041**	**1.343**
VAVAE	(1, 16, 16)	32	VAE	27.70	0.798	0.096	0.279	27.50	0.811	0.093	2.709
OmniTokenizer	(4, 8, 8)	8	VAE	26.74	0.824	0.101	1.023	26.44	0.833	0.099	4.687
Hunyuan	(4, 8, 8)	16	VAE	**33.32**	**0.916**	0.053	0.670	33.25	**0.924**	0.050	2.597
Wan2.2	(4, 16, 16)	48	VAE	31.25	0.878	0.057	0.749	31.10	0.888	0.054	3.279
ATOKEN-So/C											
Stage 1	(1, 16, 16)	32	VAE	28.77	0.814	0.099	0.258	28.66	0.829	0.096	2.336
Stage 2	(4, 16, 16)	48	VAE	29.55	0.845	0.087	0.246	29.49	0.858	0.083	2.180
Stage 3	(4, 16, 16)	48	VAE	29.72	0.848	0.085	0.209	29.67	0.861	0.081	2.026
Discrete Latent											
Cosmos-0.1-DI8×8	(1, 8, 8)	6	FSQ	25.87	0.750	0.155	0.867	25.54	0.760	0.153	5.016
GigaTok-XL-XXL	(1, 16, 16)	8	VQ	22.42	0.613	0.189	0.795	22.03	0.620	0.191	5.757
Vila-U	(1, 16, 16)	16	RQ	22.24	0.612	0.228	4.231	21.89	0.620	0.227	10.997
UniTok	(1, 16, 16)	64	MCQ	25.34	0.742	0.132	**0.362**	24.95	0.750	0.131	3.918
OmniTokenizer	(4, 8, 8)	8	VQ	24.69	0.771	0.138	1.411	24.31	0.779	0.137	6.292
ATOKEN-So/D	(4, 16, 16)	48	FSQ	**27.14**	**0.801**	**0.119**	0.379	**27.00**	**0.815**	**0.115**	**3.270**

Table 9. **Image understanding comparison with semantic encoders.** We evaluate zero-shot classification on ImageNet, ImageNet-v2, and cross-modal retrieval on COCO and Flickr30k. ATOKEN maintains competitive performance across all stages despite joint training on multiple modalities and tasks.

Res.	Seq.	Model	ImageNet-1k		COCO		Flickr	
			val	v2	T→I	I→T	T→I	I→T
224	196	CLIP	68.3	61.9	33.1	52.4	62.1	81.9
		MetaCLIP	72.4	65.1	48.9	–	77.1	–
		EVA-CLIP	74.7	67.0	42.2	58.7	71.2	85.7
		DFN	76.2	68.2	51.9	–	77.3	–
256	256	SigLIP	80.8	74.1	49.4	68.6	80.0	92.1
		SigLIP 2	**83.4**	**77.8**	**55.4**	**71.5**	**84.4**	**94.2**
		ATOKEN-So/C						
		Stage 1	82.7	76.7	54.1	70.4	81.3	93.1
		Stage 2	82.3	76.4	53.8	70.6	80.7	93.0
		Stage 3	82.2	76.1	53.7	70.5	80.5	93.2
		ATOKEN-So/D	82.2	76.2	53.8	70.1	80.9	93.5
384	576	SigLIP 2	**84.1**	**78.4**	**56.0**	**71.2**	**85.3**	**95.9**
		ATOKEN-So/C						
		Stage 1	83.4	77.6	54.8	70.4	81.7	93.8
		Stage 2	82.9	77.1	54.7	71.1	81.9	93.9
		Stage 3	82.9	76.8	54.6	71.3	81.9	93.5
		ATOKEN-So/D	82.8	76.6	54.4	70.9	81.9	93.5
512	1024	SigLIP 2	**84.3**	**79.1**	**56.0**	**71.3**	**85.5**	**95.4**
		ATOKEN-So/C						
		Stage 1	83.5	77.8	54.7	71.1	82.1	94.1
		Stage 2	83.1	77.3	54.7	71.3	82.2	93.6
		Stage 3	82.9	77.2	54.7	71.1	82.3	93.6
		ATOKEN-So/D	82.9	77.0	54.7	71.2	82.3	93.5

racy remains stable (82.7% → 82.3% → 82.2%), with only 0.5% degradation as modalities are added. Discrete quantization also preserves full semantic performance, achieving 82.2% accuracy.

C.2. Video Tokenization

We evaluate ATOKEN's video capabilities through reconstruction quality and semantic understanding benchmarks, demonstrating competitive performance while uniquely supporting both continuous and discrete representations across multiple modalities.

Reconstruction Performance. We evaluate video reconstruction on DAVIS [82] (1080p, 50 videos) and TokenBench [4] (720p, 471 videos), reporting PSNR and SSIM for pixel quality, LPIPS for perceptual similarity, and rFVD for temporal consistency. All baselines were re-evaluated using official implementations with consistent protocols and spatial tiling for memory management. ATOKEN employs temporal tiling with KV-caching, leveraging its native 2048×2048 resolution support.

As shown in Tab. 10, ATOKEN-So/C achieves 33.11 PSNR on DAVIS and 36.07 PSNR on TokenBench, approaching specialized video-only models (Wan2.1 [107]: 33.50 and 36.11, Hunyuan [51]: 32.33 and 36.37). Notably, we demonstrate that transformers can match CNN performance when properly designed – our method dramatically outperforms OmniTokenizer's transformer baseline (21.06 vs 33.11 PSNR on DAVIS) while adding native res-

Table 10. **Video reconstruction comparison on high-resolution benchmarks.** We evaluate quality on DAVIS at 1080p and TokenBench at 720p. All methods are re-evaluated using official implementations with consistent protocols for fair comparison. ATOKEN achieves competitive performance with specialized video-only tokenizers while uniquely supporting both continuous and discrete representations across modalities.

Tokenizer	Comp. Ratio	Latent Size	Token Type	DAVIS				TokenBench			
				PSNR↑	SSIM↑	LPIPS↓	rFVD↓	PSNR↑	SSIM↑	LPIPS↓	rFVD↓
Continuous Latent											
Cosmos-0.1-CV4×8×8	(4, 8, 8)	16	AE	32.25	0.894	0.219	19.15	34.33	0.924	0.155	8.34
OmniTokenizer	(4, 8, 8)	8	VAE	21.06	0.800	0.315	206.34	19.39	0.782	0.275	173.48
Hunyuan	(4, 8, 8)	16	VAE	32.33	**0.907**	0.194	22.94	36.37	**0.944**	0.129	3.78
Wan2.2	(4, 16, 16)	48	VAE	33.06	**0.907**	0.184	12.65	**36.39**	0.942	**0.126**	3.19
ATOKEN-So/C											
Stage 2	(4, 16, 16)	48	VAE	32.29	0.902	0.196	13.50	35.63	0.937	0.139	3.63
Stage 3	(4, 16, 16)	48	VAE	33.11	**0.907**	0.189	**10.76**	36.07	0.940	0.135	**3.01**
Discrete Latent											
OmniTokenizer	(4, 8, 8)	8	VQ	20.62	0.770	0.346	240.20	19.89	0.787	0.293	202.46
Cosmos-0.1-DV4×8×8	(4, 8, 8)	6	FSQ	27.26	0.798	0.310	110.33	31.20	0.892	**0.190**	25.94
ATOKEN-So/D	(4, 16, 16)	48	FSQ	**29.75**	**0.846**	**0.288**	**41.42**	**33.12**	**0.913**	0.193	**22.16**

Table 11. **Zero-shot video-text retrieval on MSRVTT and MSVD.** We compare ATOKEN against understanding-focused encoders on standard video retrieval benchmarks. Despite optimizing for both reconstruction and understanding across three modalities, ATOKEN maintains reasonable retrieval performance.

Methods	Res.	MSRVTT (1K-A)						MSVD					
		Text → Video			Video → Text			Text → Video			Video → Text		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP-ViT-B/32	224	31.2	53.7	63.3	26.4	49.9	61.7	36.4	63.3	73.1	57.8	84.1	90.7
SigLIP2-So400m	256	41.9	66.3	75.7	32.4	55.4	65.9	**55.5**	**81.2**	87.8	72.7	91.7	96.1
VideoPrism-g	288	**52.7**	**77.2**	-	**51.7**	**75.2**	-	-	-	-	-	-	-
PE-Core-B16	224	45.8	70.1	78.1	45.5	70.9	80.0	48.7	75.5	84.1	79.1	96.7	98.8
PE-Core-L14	336	49.1	73.3	**81.6**	50.9	74.4	**82.7**	54.4	**81.2**	**88.4**	**82.5**	**98.2**	**99.4**
ATOKEN-So/C–224													
Stage 1	224	40.8	65.3	75.2	31.0	55.0	63.7	53.9	79.9	87.3	72.4	93.0	95.4
Stage 2	224	40.1	64.9	75.2	30.9	53.7	64.0	53.4	79.6	87.1	71.6	91.9	95.5
Stage 3	224	40.2	64.9	75.2	30.5	53.1	63.2	53.5	79.5	87.1	72.4	91.6	95.4
ATOKEN-So/D	224	40.3	65.0	74.6	30.3	51.8	61.7	53.8	79.7	87.2	71.5	91.8	95.2

olution support. Furthermore, our progressive training reveals cross-modal benefits: incorporating 3D in Stage 3 improves video reconstruction from 35.63 to 36.07 PSNR on TokenBench, indicating that geometric understanding may enhance temporal modeling.

For discrete tokenization, ATOKEN-So/D pioneers multimodal video support, achieving 29.75 PSNR on DAVIS—surpassing Cosmos-0.1-DV (27.26) and dramatically outperforming OmniTokenizer (20.62), while maintaining reasonable perceptual quality (0.288 LPIPS) for downstream tasks.

Semantic Understanding. Tab. 11 evaluates zero-shot video-text retrieval on MSRVTT [132] and MSVD [9]. Following standard protocols [66, 114], we use frame embedding averaging with zero-padding. ATOKEN achieves 40.2% R@1 on MSRVTT and 53.5% on MSVD, maintaining reasonable semantic alignment despite optimizing primarily for reconstruction across three modalities. We note that alternative pooling strategies without frame averaging yielded lower performance, likely due to the limited video-text pairs in our training data compared to dedicated video understanding models. While understanding-only models trained on large-scale video-text data achieve higher scores, our results validate that unified tokenization successfully balances reconstruction quality with semantic understanding.

C.3. 3D Tokenization.

We evaluate ATOKEN's 3D capabilities on Toys4k [95] for reconstruction and semantic understanding. For reconstruction, ATOKEN-So/C achieves 28.28 PSNR and 0.062 LPIPS (Tab. 12), surpassing the specialized Trellis-SLAT [125] baseline (26.97 PSNR, 0.054 LPIPS) despite jointly training across three modalities. This demonstrates that our unified

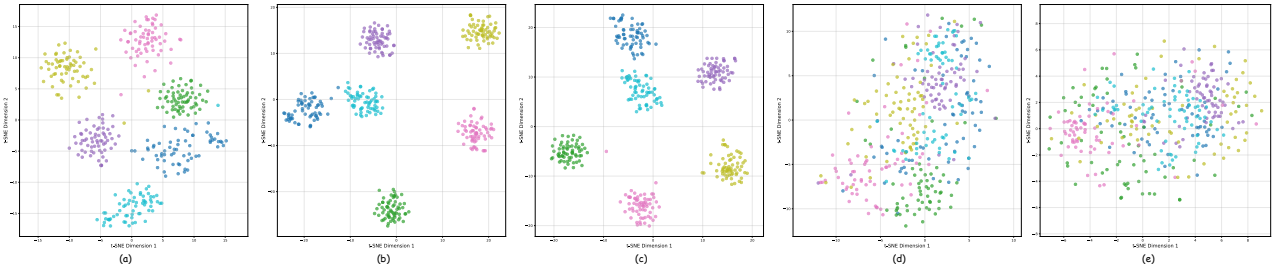

Figure 9. **Learned representations across training stages.** T-SNE visualizations of ImageNet class embeddings (colors indicate different classes). (a) Stage 1: image-only training. (b) Stage 2: with video. (c) Stage 3: dense features before projection. (d) Stage 3: projected 48-dim latents. (e) Stage 4: before FSQ quantization. Dense features (a-c) show clear semantic clustering, while dimensional reduction (d-e) leads to more mixed class distributions, suggesting a trade-off between compression and semantic separability.

Table 12. **3D reconstruction comparison on Toys4k.** We average metrics across rendered multi-view images. ATOKEN achieves comparable performance to specialized Trellis-SLAT despite jointly optimizing for three modalities, demonstrating unified training maintains strong 3D capabilities.

Method	PSNR↑	SSIM↑	LPIPS↓
Specialized 3D Tokenizer			
Trellis-SLAT	26.97	0.943	**0.054**
*Our Unified Tokenizer (*ATOKEN*)*			
ATOKEN -So/C	**28.28**	**0.951**	0.062
ATOKEN -So/D	28.17	**0.951**	0.063

4D representation effectively captures geometric structure without requiring dedicated 3D architectures.

For semantic understanding, ATOKEN-So/C achieves 90.9% zero-shot classification accuracy on Toys4k, validating that our approach maintains strong semantic representations for 3D objects alongside reconstruction capabilities. Combined with our image and video results, this confirms that all three modalities can coexist within a single tokenizer without significant trade-offs.

C.4. Ablation Study

Representation Structure Analysis. Fig. 9 visualizes learned representations through T-SNE projections across training stages. Dense features (a-c) show clear semantic clustering with distinct ImageNet class separation. However, projection to 48-dimensional latents (d-e) results in more intermixed distributions, likely due to KL regularization without post-projection alignment loss.

Despite this apparent mixing in T-SNE visualizations, the model maintains strong reconstruction and understanding performance, suggesting that semantic information may be encoded in ways not captured by 2D projections. This raises an interesting question: whether explicit semantic clustering in low-dimensional spaces – as emphasized by methods like VAVAE [140] – is necessary for strong performance, or whether larger models can effectively leverage seemingly intermixed representations. Our results suggest the latter, though we leave detailed investigation of semantic preservation through aggressive dimensionality reduction for future work.

Reconstruction Visualization. Figures 10-12 provide qualitative comparisons of reconstruction quality across all three modalities. For images (Fig. 10), ATOKEN operates at a higher compression ratio (16×) than most baselines yet achieves superior visual fidelity, particularly in preserving high-frequency details such as text clarity, fine textures, and complex patterns. The comparison reveals that methods optimized for lower compression ratios (e.g., SD-VAE and OmniTok at 8×) struggle with text legibility and texture preservation, while ATOKEN maintains sharp details. For video reconstruction (Fig. 11), ATOKEN demonstrates temporal consistency comparable to specialized video tokenizers like Wan2.2, with both continuous and discrete variants preserving motion smoothness across 720p sequences. The 3D reconstruction results (Fig. 12) highlight ATOKEN's advantage in color consistency. While Trellis-SLAT exhibits color shifts and artifacts, our unified training across modalities transfers color understanding from images and videos to improve 3D reconstruction.

D. Additional Downstream Results

D.1. Multimodal LLMs

To validate ATOKEN's effectiveness for vision-language understanding, we integrate it into SlowFast-LLaVA-1.5 [134], replacing the Oryx-ViT [61] vision encoder with ATOKEN-So/C while keeping all other settings identical. To assess generalization, the ATOKEN parameters are frozen during training, with only the SlowFast projector and LLM updated. We evaluate using the `lmms-eval` [152] toolkit and report official metrics without output filtering.

Image Understanding. Tab. 13 shows the image understanding results on 7 standard benchmarks, including

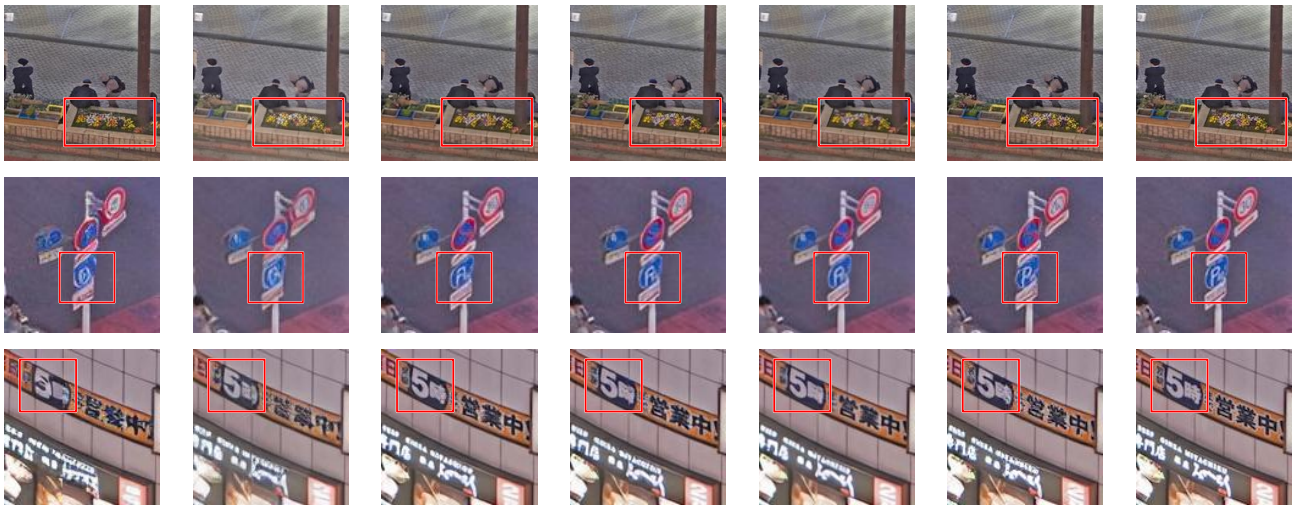

Figure 10. **Qualitative comparison of image reconstruction performance across different tokenization methods.** The latent shape for a 256 × 256 image patch is shown under each method name. Despite operating at higher compression ratios, ATOKEN demonstrates superior reconstruction quality, particularly excelling in preserving high-frequency textures, fine details, and complex text elements.

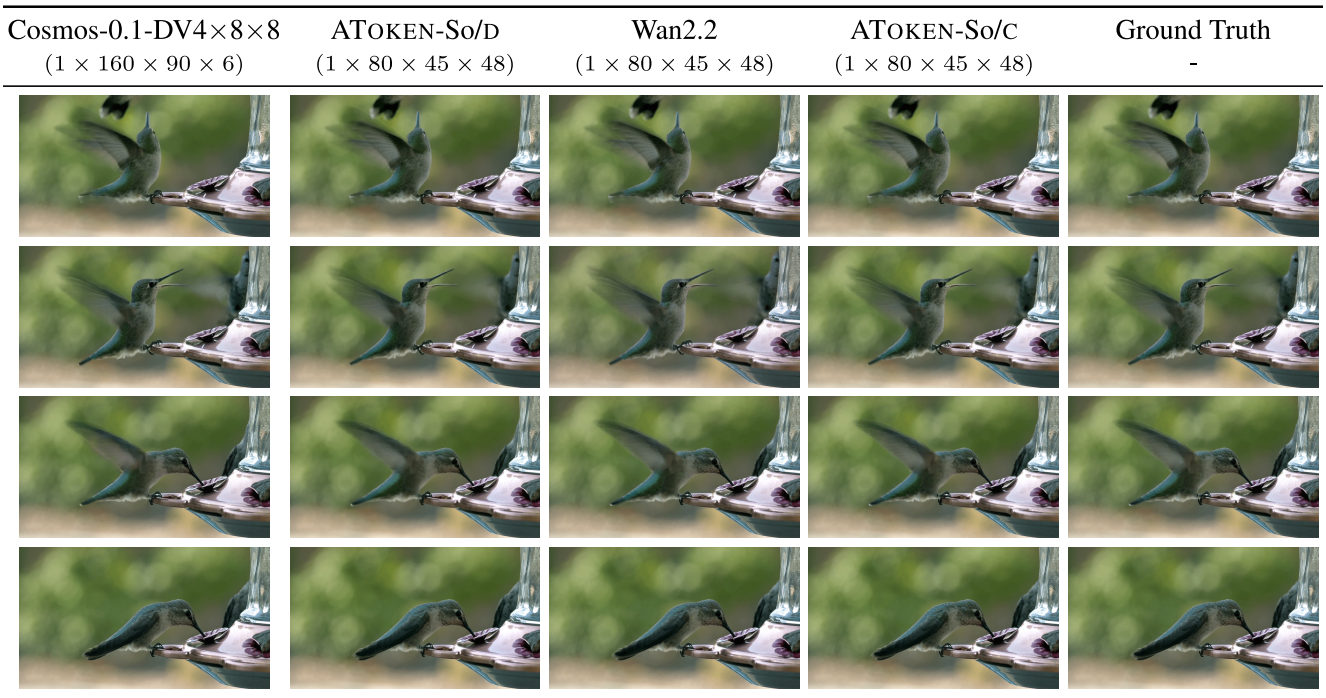

Figure 11. **Qualitative comparison of video reconstruction performance on 720p video sequences.** The latent shape for each video tokenization method is indicated under the method name. ATOKEN achieves comparable quality to specialized video-only methods while uniquely supporting both continuous and discrete representations in a unified framework.

Table 13. **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. Results shown for 7 benchmarks (general QA and text-rich understanding) across 1B, 3B, and 7B model scales.

			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									
LLaVA-OV-0.5B	SigLIP	5.31M	55.6	57.1	67.2	31.4	34.8	-	-
MM1.5-1B	CLIP	4.52M	53.3	59.3	82.1	35.8	37.2	60.5	72.5
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									
BLIP3-4B	SigLIP	-	60.5	-	88.3	41.1	39.6	-	71.0
MM1.5-3B	CLIP	4.52M	56.9	65.7	85.8	37.1	44.4	65.7	76.5
Phi-3.5-V-4B	CLIP	-	-	78.1	91.3	43.0	43.9	-	72.0
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									
LLaVA-OV-7B	SigLIP	5.31M	66.3	81.4	96.0	48.8	63.2	-	-
MM1.5-7B	CLIP	4.52M	62.5	72.2	89.6	41.8	47.6	63.5	76.5
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 14. **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. Results shown for 6 benchmarks (general and long-form video understanding) across 1B, 3B, and 7B model scales.

			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								
Apollo-1.5B	SigLIP	3K	53.0	61.0	-	54.1	63.3	-
InternVL2.5-2B	InternViT	16K	51.9	-	77.2	52.0	61.4	37.9
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								
InternVL2-4B	InternViT	16K	53.9	53.9	71.1	53.0	59.9	35.1
LinVT-Blip3-4B	SigLIP	-	58.3	-	80.1	56.6	67.9	-
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								
Oryx1.5-7B	Oryx-ViT	14K	58.8	70.0	81.8	56.3	67.5	39.0
LLaVA-Video-7B	SigLIP	11K	63.3	66.9	83.2	58.2	70.8	-
Apollo-7B	SigLIP	3K	61.3	67.3	-	58.5	70.9	-
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

Table 15. **Text-to-image and text-to-video generation benchmarks.** We compare ATOKEN Stages 2-3 with specialized video tokenizers (Cosmos, Hunyuan, Wan) under resource-constrained settings. Higher scores indicate better performance across all metrics. All models trained with identical data and model sizes for fair comparison.

Tokenizer	Comp. Ratio	Latent Size	Patch Size	T2I			T2V: VBench		
				CLIP	Pick	GenEval	Quality	Semantic	Total
Cosmos-0.1-CV4×8×8	(4, 8, 8)	16	2	32.16	21.47	62.14%	77.27%	65.13%	74.84%
Hunyuan	(4, 8, 8)	16	2	32.49	21.66	66.11%	79.52%	72.03%	78.02%
Wan2.1	(4, 8, 8)	16	2	32.45	21.62	65.57%	79.74%	74.01%	78.60%
ATOKEN-So/C									
Stage 2	(4,16,16)	48	1	32.44	21.59	63.08%	79.30%	72.42%	77.92%
Stage 3	(4,16,16)	48	1	32.50	21.74	64.61%	79.82%	73.04%	78.46%

RW-QA[1], AI2D [47], SQA [64], and MMMU [147], and MathVISTA [65] for general image QA, as well as OCR-Bench [60] and TextVQA [92] for text and document understanding. To position our models relative to state-of-the-art methods, we compare it against LLaVA-OV [55], MM1.5 [151], Molmo [18], BLIP3 [136], Phi-3.5-V [1], InternVL2.5 [153], and Qwen2-VL [111].

Here we highlight some key observations. *First,* compared to Oryx-ViT, a specific vision encoder for multimodal understanding, SlowFast-LLaVA-1.5 with ATOKEN as vision encoder shows overall better performance on image understanding across different model scales. Specifically, Tab. 13 shows that SlowFast-LLaVA-1.5-7B with ATOKEN outperforms Oryx-ViT under the same MLLM by 1.3% on RW-QA, 1.0% on SQA, and 1.3% on TextVQA. *Second*, ATOKEN shows strong generalization ability across different tasks and model scales. For reference, using ATOKEN, SlowFast-LLaVA-1.5-3B achieves superior results on almost all benchmarks. On RW-QA and AI2D, ATOKEN outperforms Oryx-ViT across the 1B, 3B, and 7B scales and achieves very competitive performance.

Video Understanding. The video understanding results are summarized in Tab. 14, covering a range of video tasks. Video-MME [27], PercepTest [83], and NExT-QA [126] assess general video QA, whereas LongVideoBench [121], MLVU [159], and LVBench [113] focus on temporal understanding on long-range context. We compared with both video specialist models, such as Apollo [160], LLaVA-Video [155], and LinVT [29], and unified image-video MLLMs, such as Oryx1.5 [61], InternVL2.5 [153], and Qwen2VL [111].

We outline several key observations. *First*, ATOKEN excels at smaller model scales. For reference, SlowFast-LLaVA-1.5-1.5B with ATOKEN achieves state-of-the-art performance on almost all benchmarks (*e.g.*, outperforming Oryx-ViT by 0.8% on LongVideoBench and 1.4% on LVBench). *Second*, ATOKENprovides more performance gain on general video QA benchmarks. Specifically, it achieves state-of-the-art results on VideoMME (*e.g.*, 64.5% with 7B LLM) and PercepTest (*e.g.*, 70.3% with 7B LLM) across scales. *Third*, we note the strong performance of Oryx-ViT on long-form video understanding, particularly on MLVU. We hypothesize that this advantage arises because (*i*) Oryx-ViT was specifically designed for video understanding in LLMs and (*ii*) it was trained on long-video retrieval tasks. Future work to address this gap includes incorporating more long videos into our training data to strengthen temporal modeling over long-range context.

D.2. Text to Video Generation

To assess the text-to-video (T2V) capabilities of the ATOKEN-So/C tokenizers, we integrate them into a video generation model. Our model is built upon the MMDiT backbone [23] and incorporates design elements from recent video architectures [51, 80, 107]. Due to computational constraints, we conduct experiments with smaller models and limited training data, maintaining consistent settings across all tokenizers for fair comparison. Following a standard two-stage training approach, we first pretrain the model from scratch on text-to-image (T2I) tasks with each tokenizer. We then adapt this image model for video generation, enabling evaluation on both T2I and T2V benchmarks. To provide a fair and efficient basis for comparing tokenizers, all training is conducted at low resolutions, using 256×256 for images and 192×336 for videos.

For T2I evaluation, we report CLIP-Score [39], PickScore [50], and GenEval [32]. For T2V tasks, we evaluate performance using the VBench benchmark [42]. We compare our results against state-of-the-art video tokenizers, namely Cosmos [4], Hunyuan [51], and Wan [107]. To ensure a fair comparison, we normalize the effective token budget for video generation across all tokenizers by adjusting the patch size. For example, we use a patch size of 2×2 for 8×8 spatial compression and 1×1 for 16×16 compression. Additionally, for T2V generation, we adjust the clas-

[1]https://huggingface.co/datasets/xai-org/RealworldQA

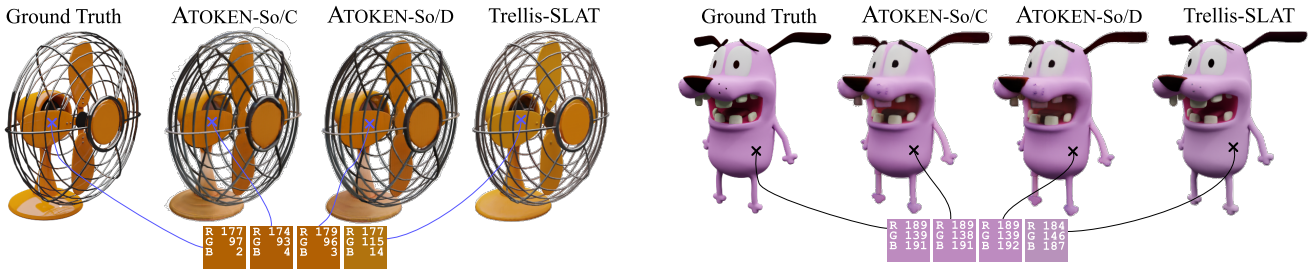

Figure 12. **3D Reconstruction Visualization on Toys4k.** ATOKEN's improved color consistency results in a higher PSNR compared to specialized 3D tokenizer Trellis-SLAT.

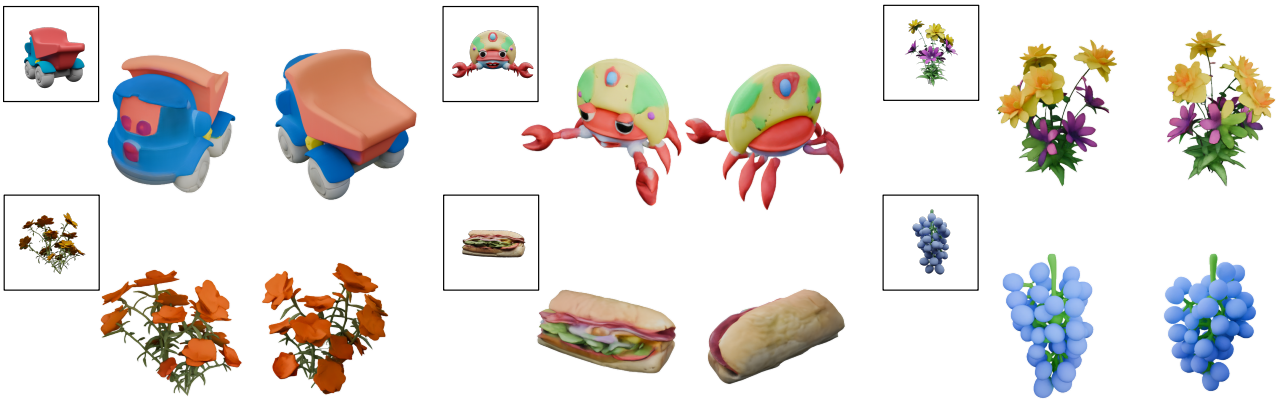

Figure 13. **Image-to-3D Generation Visualization on Toys4k.**

sifier free guidance (CFG) scale to account for differences in channel size, using a scale of 9.0 for a channel size of 48 and 4.5 for a channel size of 16.

As shown in Tab. 15, our ATOKEN-So/C tokenizers achieve results comparable to specialized video-optimized tokenizers across all metrics, outperforming Cosmos and matching the performance of Hunyuan and Wan, even though ours are designed for a broader range of tasks.

D.3. Image to 3D Synthesis

To validate the utility of our learned discrete tokens for downstream generative tasks, we train an image-to-3D synthesis model. Following the methodology of Trellis-SLAT [125], we adopt their diffusion model architecture and training regimen. We replace their original 3D tokens with the tokens generated by our ATOKEN-So/C. For a fair comparison, all inference hyperparameters, such as the number of diffusion steps and classifier-free guidance scale, are kept identical to those reported in the original work.

As shown in Fig. 13, our approach successfully generates 3D assets from single conditioning images, demonstrating that our tokens are suitable for complex generative modeling. However, we observe that the performance does not yet match the fidelity of the original Trellis-SLAT model. Specifically, while our tokenizer demonstrates excellent reconstruction capabilities that preserve color and structure (as in Fig. 12), the generative model sometimes struggles to maintain this consistency. The generated assets do not always adhere strictly to the color and style of the input image.

We hypothesize that this discrepancy arises from the significantly larger latent channel dimension of our tokenizer. ATOKEN-So/C uses 48 latent channels to accommodate rich multimodal information, a substantial increase from the 8 channels used in Trellis-SLAT. A diffusion model operating in this higher-dimensional space likely requires further optimization of training and inference hyperparameters (e.g., conditioning strength, diffusion schedule) to leverage the conditioning signal fully. We leave the exploration of these optimizations as a promising direction for future work.

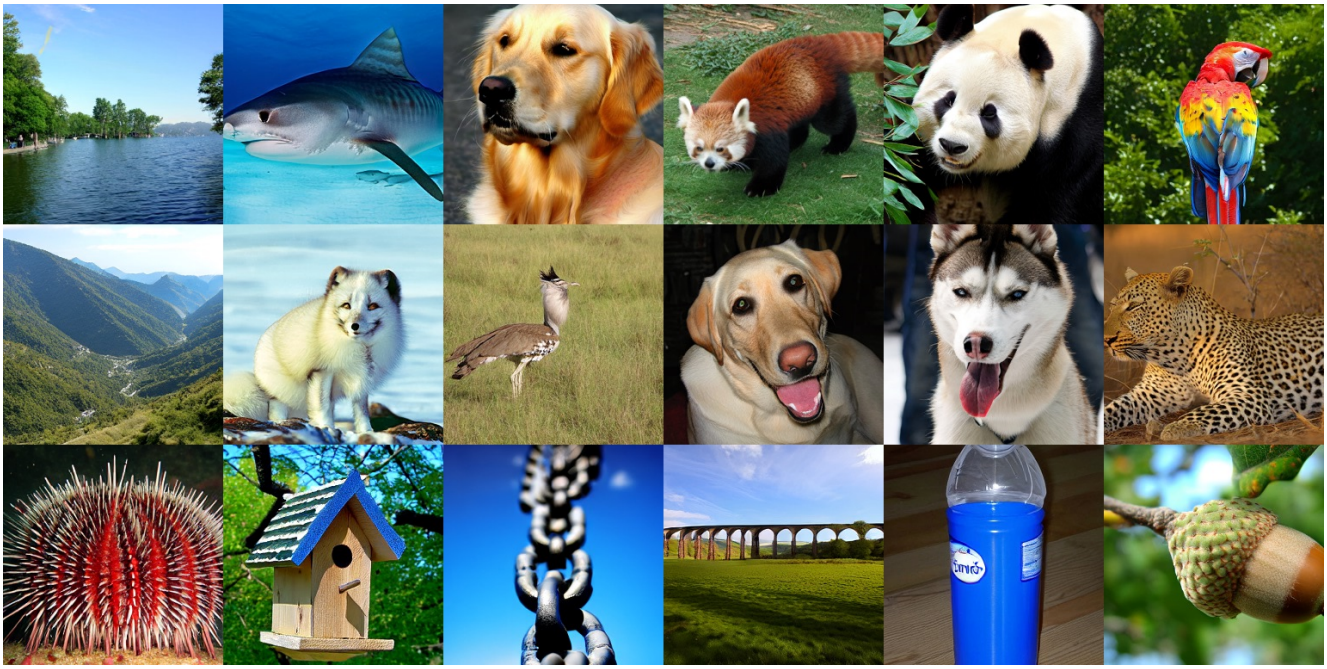

Figure 14. **ImageNet Generation Samples Using Continuous Token.**

Figure 15. **ImageNet Generation Samples Using Discrete Token.**

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv:2404.14219*, 2024. 17

[2] 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

[3] 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

[4] 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, 12, 17

[5] 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

[6] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv:2106.08254*, 2021. 9

[7] 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, 9

[8] 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. 9

[9] 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*, pages 190–200, 2011. 13

[10] 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. 9

[11] 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, 2024. 9

[12] 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. 9

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 9

[14] 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

[15] 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

[16] 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. 9

[17] 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

[18] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv:2409.17146*, 2024. 17

[19] 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

[20] 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

[21] 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. 9

[22] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020. 1, 2, 6, 9

[23] 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, 17

[24] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *ArXiv*, abs/2309.17425, 2023. 6, 11

[25] 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. 9

[26] 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*, pages 3299–3309, 2021. 9

[27] 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, 17

[28] 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. 9

[29] Lishuai Gao, Yujie Zhong, Yingsen Zeng, Haoxian Tan, Dengjie Li, and Zheng Zhao. Linvt: Empower your image-level large language model to understand videos. *arXiv:2412.05185*, 2024. 17

[30] 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

[31] 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*, pages 102–118. Springer, 2022. 9

[32] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *NeurIPS*, 2023. 17

[33] 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

[34] 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

[35] Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv:2303.05371*, 2023. 9

[36] 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, 9

[37] 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*, pages 16000–16009, 2022. 9

[38] 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*, pages 463–479. Springer, 2024. 9

[39] 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. 17

[40] 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. 9

[41] 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)*, pages 14856–14866, 2022. 9

[42] 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. 17

[43] 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*, pages 1–9, 2022. 9

[44] 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*, pages 4904–4916. PMLR, 2021. 9

[45] 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*, pages 3600–3610, 2025. 10

[46] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv:2305.02463*, 2023. 9

[47] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016. 8, 17

[48] 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. 9

[49] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, 2013. 9

[50] 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. 17

[51] 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, 9, 11, 12, 17

[52] 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

[53] 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*, pages 112–130. Springer, 2024. 9

[54] 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)*, pages 11513–11522, 2022. 9

[55] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv:2408.03326*, 2024. 17

[56] 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)*, pages 23119–23129, 2022. 9

[57] 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, 2024. 9

[58] 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, 2024. 9

[59] 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. 9

[60] 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, 17

[61] 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, 14, 17

[62] 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, 2022. 10

[63] 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, 10

[64] 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, 17

[65] 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, 17

[66] 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. 13

[67] 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*, pages 2837–2845, 2021. 9

[68] 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

[69] 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, 2024. 9

[70] 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, 10, 11

[71] 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

[72] 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*, 2024. 10

[73] Fabian Mentzer, David C. Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *ArXiv*, abs/2309.15505, 2023. 3, 5, 9, 11

[74] 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

[75] 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

[76] 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. 10

[77] 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. 9

[78] 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. 9

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

[80] 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. 17

[81] 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, 9

[82] 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. 12

[83] 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, 17

[84] 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*, pages 6964–6974, 2021. 9

[85] 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. 9

[86] 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, 9, 11

[87] 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)*, pages 1235–1245, 2021. 9

[88] 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, 9

[89] 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, 11

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

[91] 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*, pages 20875–20886, 2023. 9

[92] 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, 17

[93] 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)*, pages 18221–18232, 2023. 9

[94] 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

[95] Stefan Stojanov, Anh Thai, and James M. Rehg. Using shape to categorize: Low-shot learning with an explicit shape bias. 2021. 13

[96] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv:2303.15389*, 2023. 11

[97] Chameleon Team and Jacob Kahn. Chameleon: Mixed-modal early-fusion foundation models. *ArXiv*, abs/2405.09818, 2024. 10

[98] 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

[99] 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. 9

[100] 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. 10

[101] 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. 9

[102] 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

[103] 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, 9, 10, 11

[104] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 9

[105] 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. 9

[106] 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*, pages 1096–1103, 2008. 9

[107] 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, 9, 12, 17

[108] Hanyu Wang, Saksham Suri, Yixuan Ren, Hao Chen, and Abhinav Shrivastava. Larp: Tokenizing videos with a learned autoregressive generative prior. *arXiv:2410.21264*, 2024. 9

[109] 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, 9, 11

[110] 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*, 2022. 10

[111] 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, 17

[112] 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*, pages 4563–4573, 2023. 9

[113] 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, 17

[114] 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, 13

[115] 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

[116] 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

[117] 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. 9

[118] 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. 9

[119] 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, 2024. 10

[120] 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

[121] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *NeurIPS*, 2025. 8, 17

[122] 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*, 2024. 10

[123] 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

[124] 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, 10

[125] 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, 9, 10, 13, 18

[126] 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, 17

[127] 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. 10

[128] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *ArXiv*, abs/2506.15564, 2025. 10

[129] 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. 9

[130] 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, 9, 11

[131] 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. 11

[132] 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, 13

[133] 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. 9

[134] 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, 14

[135] 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. 9

[136] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv:2408.08872*, 2024. 17

[137] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv:2104.10157*, 2021. 9

[138] 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. 9

[139] 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. 9

[140] Jingfeng Yao and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. *ArXiv*, abs/2501.01423, 2025. 6, 8, 9, 10, 11, 14

[141] 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, 9

[142] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv:2205.01917*, 2022. 9

[143] 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

[144] 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*, pages 10459–10469, 2023. 9

[145] 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, 9

[146] 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

[147] 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, 17

[148] 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)*, pages 16354–16366, 2022. 9

[149] 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. 9

[150] 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, 9

[151] Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, et al. MM1. 5: Methods, analysis & insights from multimodal llm fine-tuning. *ICLR*, 2025. 17

[152] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024. 14

[153] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv:2407.03320*, 2024. 17

[154] 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

[155] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv:2410.02713*, 2024. 17

[156] 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

[157] 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. 9

[158] 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 multimodal model. *ArXiv*, abs/2408.11039, 2024. 10

[159] 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, 17

[160] Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. Apollo: An exploration of video understanding in large multimodal models. *arXiv:2412.10360*, 2024. 17