TUNA: Taming Unified Visual Representations for Native Unified Multimodal Models

Supplementary Material

6. VBench Column Names

We list the full column names in Table 5 below:

QS: Quality Score, SS: Semantic Score, SC: Subject Consistency, BC: Background Consistency, TF: Temporal Flickering, MS: Motion Smoothness, DD: Dynamic Degree, AQ: Aesthetic Quality, IQ: Imaging Quality, OC: Object Class, MO: Multiple Objects, HA: Human Action, C: Color, SR: Spatial Relationship, S: Scene, AS: Appearance style, TS: Temporal Style, OC': Overall Consistency.

7. Extended Related Work

7.1. Large Multimodal Models

Large multimodal models (LMMs) aim to generate text responses from multimodal inputs spanning images, videos, and text. Early LMMs such as Flamingo [1] and Idefics [35] introduced cross-attention layers to enable interaction between visual and linguistic features. Modern LMMs generally follow the LLaVA paradigm [52], where visual inputs are encoded by a vision encoder (e.g., CLIP [73]) and then concatenated with text tokens for joint processing by a language model decoder. Recent research advances focus on improving instruction-following through higher-quality training data [3, 9, 39, 42, 53, 74, 103, 127], developing stronger vision encoders capable of handling higher-resolution images [5, 36, 53, 97], extending LMMs to interleaved image [32, 36, 39] and video understanding [42, 44, 48, 63, 126], and incorporating reinforcement learning with thinking modes [14, 19, 28] or pixel-space reasoning [57, 83, 85].

7.2. Diffusion Generative Models

Diffusion generative models have become the de facto backbone of high-fidelity image [6, 17, 45, 105] and video [34, 78, 94] synthesis. Modern large-scale visual generation models typically apply diffusion in a continuous latent space defined by a learned VAE, following the Latent Diffusion Model (LDM) paradigm [75], which offers superior perceptual quality and sampling efficiency compared to autoregressive decoding of long sequences of discrete tokens based on VQ-VAE [16, 93]. Within diffusion itself, latent-space models [17, 71, 75] are generally preferred over pixel-space approaches [15, 77] because they reduce computational cost, ease scaling to higher resolutions, and allow the denoising network to focus on semantically meaningful structure rather than low-level pixel noise. Architecturally, diffusion backbones have evolved from convolutional U-Net designs [26, 76] to diffusion transformers (DiT) [60, 70]; In parallel, the learning objective has been generalized from Gaussian noise prediction and score matching [26, 81] to more expressive formulations such as rectified flows [55] and flow matching objectives [2, 51].

7.3. Representation in Multimodal Models

Recent studies have explored learning better representations to enhance multimodal understanding and generation models. From the perspective of improving understanding models, methods such as Ross [96], GenHancer [61] and ASVR [95] enhance multimodal understanding by introducing generation or reconstruction objectives, encouraging the model to capture fine-grained visual details. Conversely, to improve generative models, approaches such as REPA [122] and VA-VAE [120] align diffusion transformers or VAE representations with semantic vision encoders, thereby achieving stronger generative performance. Similarly, Dispersive Loss [98] introduces an auxiliary contrastive-like objective to further enhance generation quality.

8. Additional Implementation Details

We list the detailed hyperparameter settings for each of our training stages in Table 7.

9. Additional Experimental Results

In this section, we present additional experimental results on image generation, image editing and video understanding benchmarks.

Image generation. We further evaluate TUNA on OneIG-Bench [7]. As shown in Table 8, TUNA achieves the best overall performance among both 1.5B and 7B unified models. Notably, TUNA shows a substantial advantage in text rendering quality, indicating its strong semantic understanding capability when generating images from complex instructions containing visual text-related information. Moreover, TUNA's performance approaches that of state-of-the-art generation-only models such as Qwen-Image [105], highlighting its strong image generation capability despite having a much smaller model size.

Image editing. We use GEdit-Bench [54] to further assess TUNA on image editing tasks. As shown in Table 9, although TUNA performs slightly below the best generation-only model (Qwen-Image [105]), it again achieves the highest overall score among all unified models. TUNA's consistently strong results on both ImgEdit-Bench and GEdit-

Table 7. TUNA's training recipe for each training stage.

Hyperparameters	Pretraining	Continue Pretraining	Supervised Finetuning
Learning rate	1.0×10^{-4}	5.0×10^{-5}	1×10^{-5}
LR scheduler	Constant	Constant	Constant
Weight decay	0.0	0.0	0.0
Gradient norm clip	1.0	1.0	1.0
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1.0 \times 10^{-15}$)		AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$)
Loss weight (CE : MSE)	0.2 : 1	0.2 : 1	0.2 : 1
Warm-up steps	2000	2000	3000
Training steps	200K	100K	30K
EMA ratio	0.9999	0.9999	0.995
Gen resolution (min short side, max long side)	(384, 672)	(384, 672)	(384, 1024)
Und resolution (min short side, max long side)	(384, 672)	(384, 672)	(384, 672)
Diffusion timestep shift	3.0	3.0	4.0
Data ratio (gen: und: text)	4:1:0	3:1:0	4:2:1

Table 8. Image generation results on OneIG-Bench. **Bold**: best results among each section. Underline: second-best.

Models	Size	Alignment	Text	Reasoning	Style	Diversity	Average
Generation-only Models							
FLUX.1 [Dev][6]	12B	0.79	0.52	0.25	0.37	0.24	0.43
Qwen-Image [105]	20B	0.88	0.89	0.31	0.42	0.20	0.54
Composite UMMs							
BLIP3-o [8]	8B	0.71	0.01	0.22	0.36	0.23	0.31
OmniGen2[106]	7B	0.80	0.68	0.27	0.38	0.24	0.48
1.5B-scale Native UMMs							
Show-o [114]	1.3B	0.70	0.00	0.21	0.36	0.24	0.25
Show-o2 [115]	1.5B	0.80	0.13	**0.27**	0.35	0.19	0.35
TUNA	1.5B	**0.82**	**0.77**	0.25	**0.36**	**0.20**	**0.48**
7B-scale Native UMMs							
Janus-Pro [11]	7B	0.55	0.00	0.14	0.28	0.37	0.27
Show-o2 [115]	7B	0.82	0.00	0.23	0.32	0.18	0.31
BAGEL [13]	14B	0.77	0.24	0.17	0.37	**0.25**	0.36
TUNA	7B	**0.84**	**0.82**	**0.27**	**0.40**	0.19	**0.50**

Table 9. Image editing results on GEdit-Bench. "G-SC" and "G-PQ" denote "G-Semantic Consistency" and "G-Perceptual Quality", respectively. **Bold**: best results among each section. Underline: second-best.

Models	Size	G-SC	G-PQ	G-Overall
Generation-only Models				
FLUX.1 Kontext [Pro][6]	-	7.02	7.60	6.56
Qwen-Image [105]	20B	8.00	7.86	7.56
Native or Composite UMMs				
UniWorld-V1 [49]	12B	4.93	7.43	4.85
OmniGen[111]	3.8B	5.96	5.89	5.06
OmniGen2[106]	4B	7.16	6.77	6.41
BAGEL [13]	14B	7.36	6.83	6.52
TUNA	7B	**7.79**	**7.48**	**7.29**

Table 10. Experimental results on video understanding benchmarks. **Bold**: best results. Underline: second-best.

Models	Size	MVBench	Video-MME		LongVidBench	LVBench
		test	w/o sub		val	test
Understanding-only Models (LMMs)						
GPT-4o [67]	-	-	71.9	66.7		48.9
Gemini-1.5-Pro [89]	-	54.2	75.0	64.0		33.1
LongVA [126]	7B	49.2	52.6	51.8		-
VideoLLaMA2 [12]	7B	54.6	47.9	-		-
LLaVA-OV [39]	7B	56.7	58.2	56.5		26.9
1.5B-scale Native UMMs						
Show-o2 [115]	1.5B	49.8	48.0		49.2	-
TUNA	1.5B	**54.4**	**49.1**		**49.7**	27.4

Bench underscore its robust image editing capability and highlight the effectiveness of our unified visual representation when handling visual generation tasks that require precise semantic understanding and accurate prompt following.

Video understanding. We employ four video understanding benchmarks to evaluate TUNA: MVBench [42], Video-MME [21], LongVideoBench [107] and LVBench [100]. As shown in Table 10, TUNA outperforms Show-o2 on MVBench and Video-MME, while achieving competitive results on LongVideoBench and LVBench. Notably, despite being only a 1.5B-parameter model, TUNA performs on par with larger understanding-only models on MVBench and LVBench, demonstrating the efficiency and effectiveness of our unified representation for video understanding tasks.

10. Comparison with Show-o2

As discussed in Section 3.3, both TUNA and Show-o2 [115] employ unified visual representations for understanding and generation, but they construct these representations in fundamentally different ways. In this section, we first describe Show-o2's unified visual representation design in detail and then highlight the key differences compared to TUNA.

As illustrated in Figure 5, Show-o2 constructs unified visual representations using a dual-path feature fusion mechanism. The input image or video is first encoded by a VAE encoder, after which the latent is processed through two parallel branches. The semantic projection branch feeds the VAE latents into a set of semantic layers to extract features for understanding tasks. The VAE projection branch applies 2D patch embedding layers to produce features tailored for generation tasks. Importantly, the semantic layers are *pre-distilled* using a frozen representation encoder: given the same image, their outputs are first aligned with a pretrained SigLIP model before conducting end-to-end training of the

Table 11. Full ablation study results.

Models	ID	Training Data	Understanding				Generation	
			MME-p	MMMU	SEED	GQA	GenEval	DPG
Show-o2 (Wan 2.1 VAE + SigLIP)	1	Understanding Data Only	1351	36.1	62.1	56.8	-	-
Decoupled (SigLIP 2 only)	2		1392	38.2	62.9	58.1	-	-
TUNA (Wan 2.2 VAE + SigLIP 2)	3		1386	37.6	62.9	57.4	-	-
Show-o2 (Wan 2.1 VAE + SigLIP)	4	Generation Data Only	-	-	-	-	76.2	82.56
Decoupled (Wan 2.2 VAE only)	5		-	-	-	-	77.3	82.87
TUNA (Wan 2.2 VAE + SigLIP 2)	6		-	-	-	-	77.8	83.33
Show-o2 (Wan 2.1 VAE + SigLIP)	7	Understanding and Generation Data	1339	35.5	61.7	55.9	75.9	82.32
Decoupled (Wan 2.2 VAE + SigLIP 2)	8		1346	37.2	61.4	56.5	78.3	83.50
TUNA (Wan 2.1 VAE + SigLIP)	9		1358	35.9	64.2	57.2	77.2	83.29
TUNA (Wan 2.2 VAE + SigLIP)	10		1349	36.3	64.6	57.4	76.9	83.10
TUNA (Wan 2.1 VAE + SigLIP 2)	11		1379	37.7	65.9	58.4	79.1	83.98
TUNA (Wan 2.2 VAE + SigLIP 2)	12		1361	38.1	66.5	58.2	79.4	84.20
TUNA (Wan 2.2 VAE + DINOv3)	13		1396	37.3	65.6	58.6	78.9	84.08

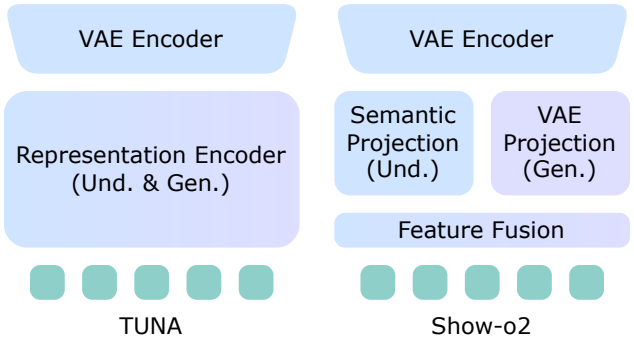


Figure 5. Comparison between TUNA and Show-o2 on how unified visual representations are produced.

Show-o2 model. This pre-distillation stage is proposed to preserve semantic understanding capability. Finally, Show-o2 merges the outputs of both paths using a feature fusion layer to obtain its unified visual representation.

While this dual-path fusion design is intended to combine understanding- and generation-oriented features, our analysis in Section 3.3 shows that Show-o2's fused representation remains biased toward semantic features, resulting in weaker generation quality. In contrast, TUNA learns a more balanced unified representation that performs well on both understanding and generation tasks. We attribute this to TUNA's end-to-end training of the unified representation on both objectives, which enables early fusion of understanding and generation signals at every layer of the representation encoder. This layer-wise interaction captures richer cross-task dependencies and is inherently more robust than the late-fusion strategy adopted in Show-o2.

11. Additional Ablation Studies

Our ablation study in Section 3.3 yields the following conclusions:

1. TUNA's unified visual representation outperforms decoupled representations on both understanding and generation tasks.
2. TUNA's unified representation improves with stronger pretrained representation encoders.
3. TUNA's unified representation enables mutual reinforcement between understanding and generation.
4. Although both models employ unified visual representations, TUNA consistently outperforms Show-o2 on both understanding and generation tasks.

Due to space constraints, we provide additional ablation results in this section to further support these conclusions. To better compare different ablation experiments, we include all results from Table 6 plus the newly added results in Table 11, with the additional entries highlighted in yellow.

Unified vs. decoupled visual representation. Comparing Models 2 and 3 with Model 8, we observe that training a unified model using decoupled visual representations results in substantial degradation on understanding tasks. In contrast, Model 12 surpasses Model 3 on most understanding benchmarks and outperforms Model 6 across all generation benchmarks. These results indicate that our unified representation suffers far less from representation conflicts than decoupled designs, enabling stronger performance in both understanding and generation.

Selection of representation encoders. Comparing Model 9 to Model 11 and Model 10 to Model 12, we observe that regardless of which VAE encoder is used in the model, replacing SigLIP with SigLIP 2 in the representation encoder consistently improves performance across all understanding and generation benchmarks. This further reinforces our conclusion that TUNA benefits from stronger pretrained representation encoders.

Understanding-generation synergy. Although the comparison between Model 2 and Model 3 shows that TUNA's

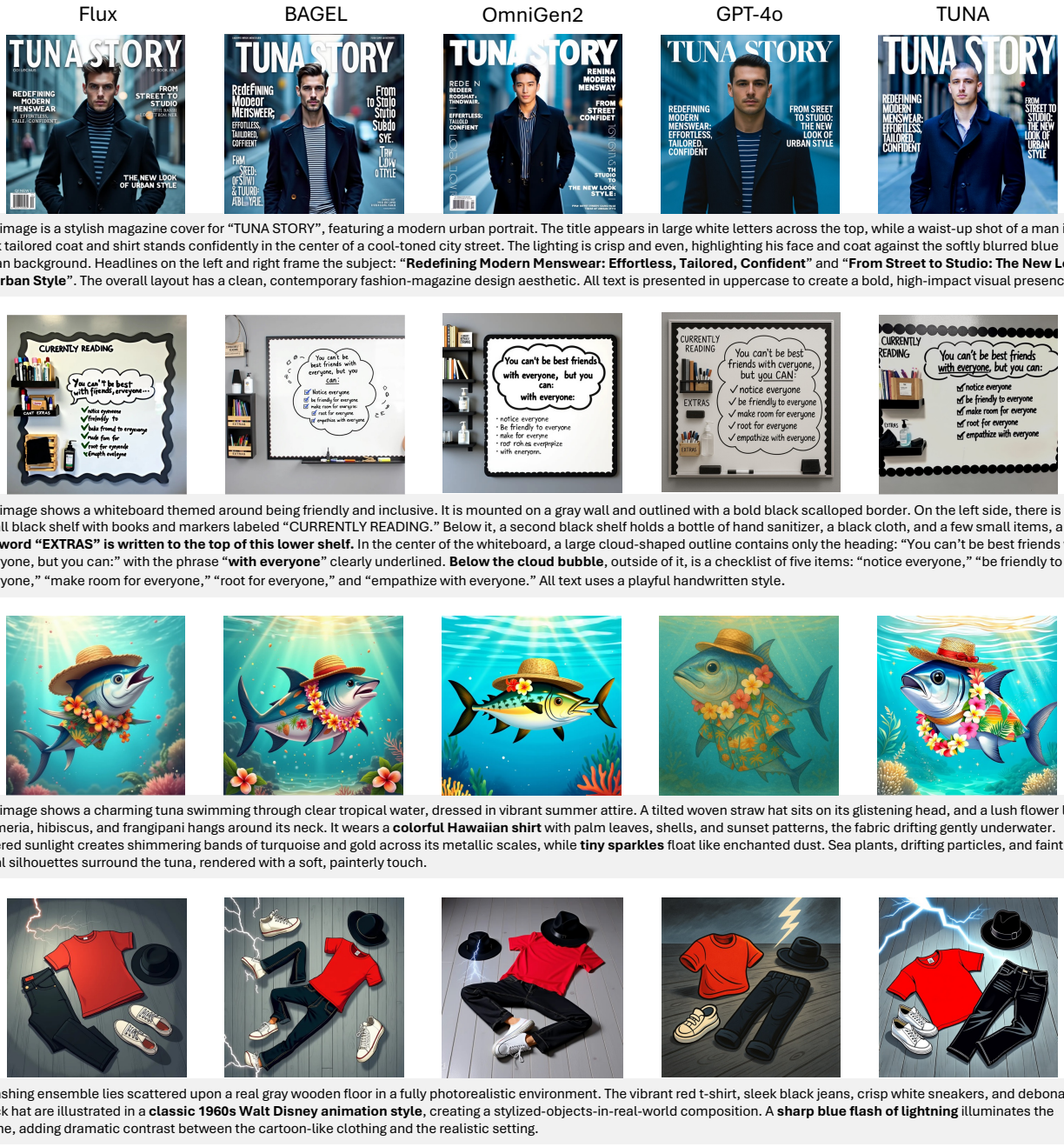


Figure 6. Qualitative comparison between TUNA and baseline models on image generation tasks. The instructions that are correctly reflected in our results but failed in some of the baseline models are **bolded**.

VAE + representation encoder architecture incurs a slight performance drop relative to using only a representation encoder (the standard setup for understanding-only models), our joint understanding + generation training pipeline largely compensates for this degradation. Specifically, Model 12 recovers its understanding performance and becomes comparable to or even better than Model 2 on several understanding benchmarks. Moreover, Model 12 substantially outperforms both Model 5 and Model 6 on all generation benchmarks. These results demonstrate the mutual enhancement between understanding and generation made possible by our unified visual representation design.

Comparison with Show-o2. Comparing Model 7 with Model 9, we see that even when initializing the VAE and representation encoder with relatively weaker pretrained weights (Wan 2.1 VAE + SigLIP), TUNA still consistently outperforms Show-o2 across all understanding and generation benchmarks. This further demonstrates the superiority

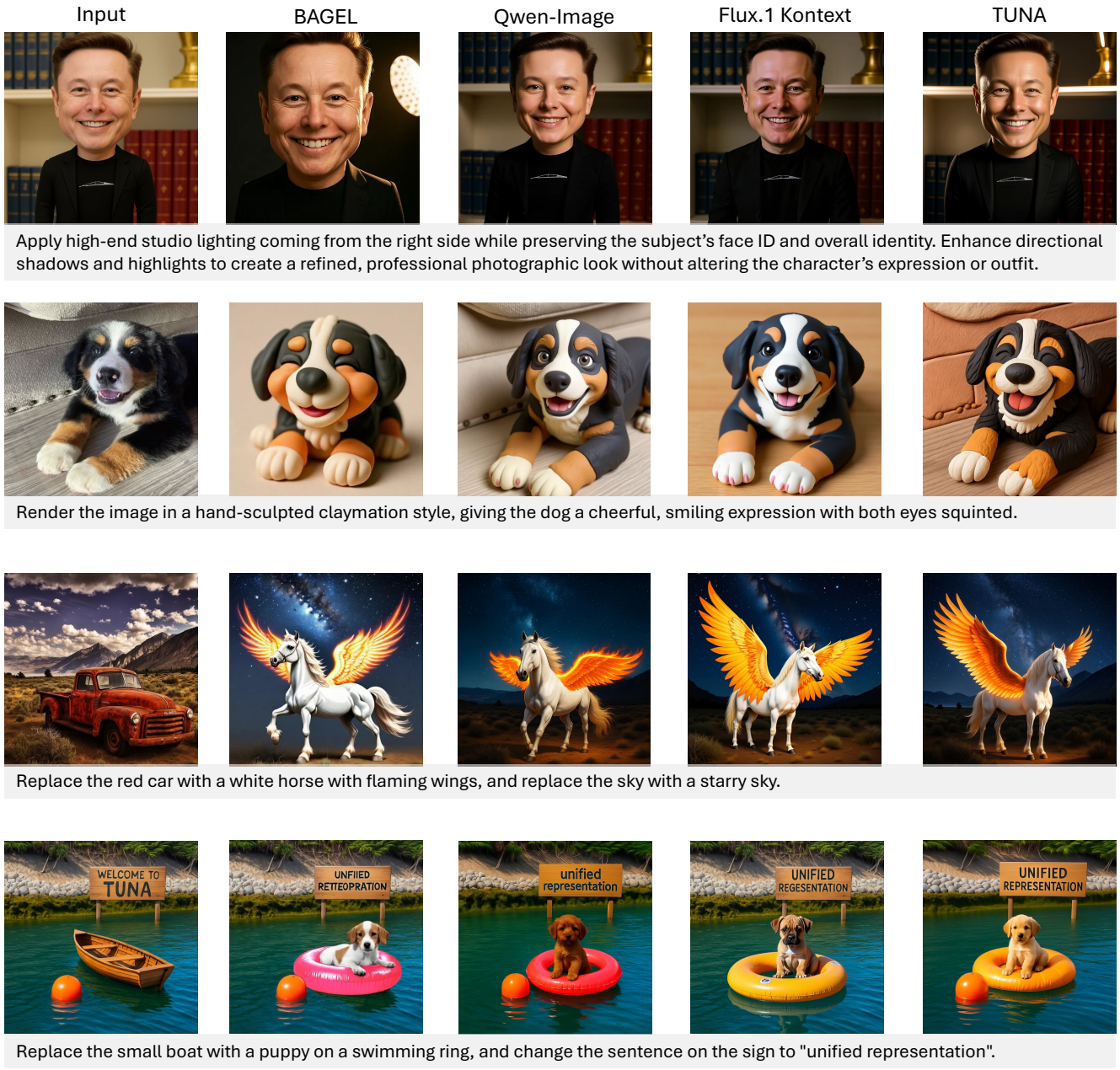


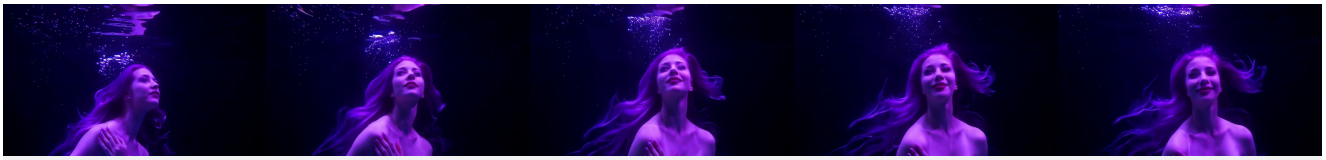
Figure 7. Qualitative comparison between TUNA and baseline models on image editing tasks.

and robustness of our unified visual representation design compared to Show-o2.

12. Qualitative Results

Image generation. We compare TUNA with state-of-the-art generation-only and unified models across diverse image generation instructions in Figure 6. In the first two examples, TUNA exhibits strong text rendering ability, accurately reproducing all visual text in the prompts without errors. In the whiteboard example, TUNA is the only model that correctly places an underline beneath "with everyone", demonstrating precise prompt-following capability. Moreover, TUNA accurately generates two black shelves, one containing books and markers on top and the other containing black cloth and hand sanitizer at the bottom, each in the correct position. Other models either fail to produce the correct number of shelves or place the wrong items on them. These results show that TUNA excels at compositional image generation, enabled by its unified visual representation with strong semantic understanding capabilities. In the "tuna" example, both TUNA and Flux [6] success-

A bird building a nest from twigs and leaves.



A woman with long, flowing purple hair and a strapless top floats underwater, gazing upwards with her hands placed on her chest. The dark background is illuminated by a soft, blue-purple light that casts an ethereal glow on her face and hair. Small, shimmering particles are scattered throughout the water, adding to the dreamy atmosphere. The woman's expression is serene, with a subtle smile playing on her lips. As the video progresses, her hair flows gently around her, and her hands remain still on her chest. The camera captures her from a medium close-up angle, with the darkness of the water serving as a striking backdrop.

A person stands at the edge of an infinity pool, arms outstretched, gazing out at a vibrant sunset over the ocean. The individual is silhouetted against the colorful sky, wearing a long-sleeved shirt and shorts. The pool's edge is barely visible, with ripples on the water's surface reflecting the sky's hues. In the background, a small island with palm trees and rocks meets the horizon, where the sun sets over the calm ocean waters. The sky transitions from blue to orange, yellow, and pink, with scattered clouds. The scene exudes a serene, peaceful atmosphere, capturing a moment of contemplation or joy at the natural beauty of the surroundings.

A majestic male lion with a thick, dark-brown mane and golden-brown fur lies on the ground, facing right. His front paws are stretched out in front of him, and his head is slightly raised, gazing into the distance. The lion's mane is long and shaggy, framing his face and neck. In the background, a tree trunk and some dry grass are visible, along with a large rock formation to the right. The ground is covered in light-brown dirt, and the overall atmosphere suggests a sunny day in a savannah or zoo enclosure. The lion's expression is calm and regal, exuding a sense of power and serenity.

The video shows a close-up view of a hand placing colorful magnetic letters and numbers on a white refrigerator in a kitchen setting. The hand, belonging to a fair-skinned individual, enters the frame from the bottom left corner and places a red '8' next to other magnets already on the fridge, including a green 'A', blue '2', yellow 'A', and orange '0'. The kitchen background features light brown cabinets, a brown tile backsplash, and a white electric kettle on the countertop. The hand then places a red '8' next to the existing magnets. The video captures the hand's movement and the interaction with the fridge magnets, set against the blurred kitchen environment.

Figure 8. Qualitative results for TUNA on the task of text-to-video generation.

fully render the Hawaiian shirt, while other models either fail to depict the shirt or generate an incorrect tuna body. Finally, in the "red t-shirt" example, TUNA accurately reflects the "classic 1960s Walt Disney animation style" and correctly includes all required elements from the prompt, maintaining a coherent and well-structured composition.

Image editing. We compare TUNA with BAGEL [13], Qwen-Image [105], and Flux.1 Kontext [6] on image editing tasks in Figure 7. As shown, TUNA not only correctly performs explicit editing operations, such as style transfer (photorealistic → hand-sculpted claymation in the "dog" example), environment change (daylight → nighttime in the "red car" example), and object replacement (boat → puppy with a swim ring in the "boat" example), but also handles more implicit and nuanced instructions, such as applying lighting from the right side in the "person" example. These results further highlight TUNA's strong semantic understanding and high-fidelity image generation capabilities.

Video generation. We present TUNA 's video generation results in Figure 8. The model produces high-fidelity videos across a wide range of instructions, demonstrating the strength of its unified visual representation space for jointly modeling both images and videos.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736, 2022. 1

[2] Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023. 1

[3] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025. 1

[4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 5

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 4, 5, 1

[6] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025. 3, 5, 6, 1, 2

[7] Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. *arXiv preprint arXiv:2506.07977*, 2025. 1

[8] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025. 5, 6, 8, 2

[9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024. 1

[10] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024. 2, 5

[11] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 5, 6, 8, 2

[12] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 2

[13] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 2, 4, 5, 6, 8

[14] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[15] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[16] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2, 1

[17] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 3, 6, 7, 1

[18] Lijie Fan, Luming Tang, Siyang Qin, Tianhong Li, Xuan Yang, Siyuan Qiao, Andreas Steiner, Chen Sun, Yuanzhen Li, Tao Zhu, et al. Unified autoregressive visual generation and understanding with continuous tokens. *arXiv preprint arXiv:2503.13436*, 2025. 8

[19] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025. 1

[20] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2025. 5

[21] Chaoyou Fu, Yuhan Dai, Yongdong 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 2

[22] Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, et al. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025. 5

[23] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023. 2, 6

[24] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3

[25] Jiaming Han, Hao Chen, Yang Zhao, Hanyu Wang, Qi Zhao, Ziyan Yang, Hao He, Xiangyu Yue, and Lu Jiang. Vision as a dialect: Unifying visual understanding and generation via text-aligned representations. *arXiv preprint arXiv:2506.18898*, 2025. 5, 6

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[27] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 6

[28] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025. 1

[29] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 6

[30] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 5

[31] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024. 7

[32] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024. 1

[33] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer, 2016. 5

[34] 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 preprint arXiv:2412.03603*, 2024. 1

[35] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023. 1

[36] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 1

[37] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2206–2217, 2023. 3

[38] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 5

[39] 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 preprint arXiv:2408.03326*, 2024. 5, 1, 2

[40] Han Li, Xinyu Peng, Yaoming Wang, Zelin Peng, Xin Chen, Rongxiang Weng, Jingang Wang, Xunliang Cai, Wenrui Dai, and Hongkai Xiong. Onecat: Decoder-only auto-regressive model for unified understanding and generation. *arXiv preprint arXiv:2509.03498*, 2025. 8

[41] Hao Li, Changyao Tian, Jie Shao, Xizhou Zhu, Zhaokai Wang, Jinguo Zhu, Wenhan Dou, Xiaogang Wang, Hongsheng Li, Lewei Lu, et al. Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29767–29779, 2025. 5

[42] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 1, 2

[43] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445, 2024. 2

[44] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024. 1

[45] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuandit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024. 1

[46] Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2779–2790, 2025. 6

[47] Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation. *arXiv preprint arXiv:2505.05472*, 2025. 2, 5, 6, 8

[48] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual rep-

resentation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 1

[49] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025. 6, 8, 2

[50] Haokun Lin, Teng Wang, Yixiao Ge, Yuying Ge, Zhichao Lu, Ying Wei, Qingfu Zhang, Zhenan Sun, and Ying Shan. Toklip: Marry visual tokens to clip for multimodal comprehension and generation, 2025. 3, 8

[51] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 1

[52] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 5, 1

[53] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next. https://llava-vl.github.io/blog/2024-01-30-llava-next/, 2024. Accessed: 2025-02-14. 1

[54] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025. 1

[55] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 1

[56] 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*, 67(12):220102, 2024. 5

[57] Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv preprint arXiv:2505.12081*, 2025. 1

[58] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[59] 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 preprint arXiv:2502.20321*, 2025. 3, 8

[60] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pages 23–40. Springer, 2024. 1

[61] Shijie Ma, Yuying Ge, Teng Wang, Yuxin Guo, Yixiao Ge, and Ying Shan. Genhancer: Imperfect generative models are secretly strong vision-centric enhancers. *arXiv preprint arXiv:2503.19480*, 2025. 1

[62] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7739–7751, 2025. 5, 8

[63] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 1

[64] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279, 2022. 5

[65] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025. 3

[66] OpenAI. gpt-image-1. Model documentation. 6

[67] OpenAI. Gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. 2

[68] 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 preprint arXiv:2304.07193*, 2023. 3

[69] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025. 6, 8

[70] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 4, 1

[71] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[72] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2545–2555, 2025. 5, 8

[73] 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. 3, 1

[74] Weiming Ren, Huan Yang, Jie Min, Cong Wei, and Wenhu Chen. Vista: Enhancing long-duration and high-resolution video understanding by video spatiotemporal augmentation. *arXiv preprint arXiv:2412.00927*, 2024. 1

[75] 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

[76] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[77] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1

[78] 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 preprint arXiv:2504.08685*, 2025. 4, 1

[79] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 7

[80] Wei Song, Yuran Wang, Zijia Song, Yadong Li, Haoze Sun, Weipeng Chen, Zenan Zhou, Jianhua Xu, Jiaqi Wang, and Kaicheng Yu. Dualtoken: Towards unifying visual understanding and generation with dual visual vocabularies. *arXiv preprint arXiv:2503.14324*, 2025. 8

[81] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1

[82] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449, 2021. 7

[83] Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025. 1

[84] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063, 2024. 4

[85] Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025. 1

[86] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 3

[87] Hao Tang, Chenwei Xie, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. Unilip: Adapting clip for unified multimodal understanding, generation and editing. *arXiv preprint arXiv:2507.23278*, 2025. 8

[88] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 2

[89] 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 preprint arXiv:2403.05530*, 2024. 2

[90] Wan-Video team. Wan: Open and advanced large-scale video generative models (wan2.2), 2025. GitHub repository, Apache-2.0 License. 4

[91] Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image captioners are scalable vision learners too. *Advances in Neural Information Processing Systems*, 36:46830–46855, 2023. 4

[92] 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 preprint arXiv:2502.14786*, 2025. 4, 7

[93] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 1

[94] 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 preprint arXiv:2503.20314*, 2025. 2, 3, 1

[95] Dianyi Wang, Wei Song, Yikun Wang, Siyuan Wang, Kaicheng Yu, Zhongyu Wei, and Jiaqi Wang. Autoregressive semantic visual reconstruction helps vlms understand better. *arXiv preprint arXiv:2506.09040*, 2025. 1

[96] Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. Reconstructive visual instruction tuning. *arXiv preprint arXiv:2410.09575*, 2024. 2, 1

[97] 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 preprint arXiv:2409.12191*, 2024. 1

[98] Runqian Wang and Kaiming He. Diffuse and disperse: Image generation with representation regularization. *arXiv preprint arXiv:2506.09027*, 2025. 1

[99] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 3

[100] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF In-*

ternational Conference on Computer Vision, pages 22958–22967, 2025. 2

[101] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 5, 6, 7

[102] Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision. In *The Thirteenth International Conference on Learning Representations*, 2024. 5

[103] Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. Finevision: Open data is all you need. *arXiv preprint arXiv:2510.17269*, 2025. 5, 1

[104] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12966–12977, 2025. 8

[105] 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. *arXiv preprint arXiv:2508.02324*, 2025. 6, 1, 2

[106] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 6, 2

[107] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024. 2

[108] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. *arXiv preprint arXiv:2503.21979*, 2025. 2, 5, 6, 8

[109] 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 preprint arXiv:2409.04429*, 2024. 5, 7

[110] xAI. Grok-1.5 vision preview. Company news post. 5

[111] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13294–13304, 2025. 6, 2

[112] Yicheng Xiao, Lin Song, Rui Yang, Cheng Cheng, Zunnan Xu, Zhaoyang Zhang, Yixiao Ge, Xiu Li, and Ying Shan. Haploomni: Unified single transformer for multimodal video understanding and generation. *arXiv preprint arXiv:2506.02975*, 2025. 7

[113] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024. 8

[114] 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 preprint arXiv:2408.12528*, 2024. 4, 5, 6, 8, 2

[115] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025. 2, 4, 5, 6, 7, 8

[116] Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding. *arXiv preprint arXiv:2411.17762*, 2024. 5, 8

[117] Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24135–24146, 2025. 6

[118] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3

[119] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 7

[120] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15703–15712, 2025. 1

[121] Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025. 6

[122] 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. 2, 3, 1

[123] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 5

[124] Zhengrong Yue, Haiyu Zhang, Xiangyu Zeng, Boyu Chen, Chenting Wang, Shaobin Zhuang, Lu Dong, KunPeng Du, Yi Wang, Limin Wang, et al. Uniflow: A unified pixel flow tokenizer for visual understanding and generation. *arXiv preprint arXiv:2510.10575*, 2025. 8

[125] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 2

[126] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 1, 2

[127] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 5, 1

[128] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025. 3

[129] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024. 2, 6, 8

[130] Muzhi Zhu, Yang Liu, Zekai Luo, Chenchen Jing, Hao Chen, Guangkai Xu, Xinlong Wang, and Chunhua Shen. Unleashing the potential of the diffusion model in few-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 37:42672–42695, 2024. 3

[131] Ran Zuo, Haoxiang Hu, Xiaoming Deng, Cangjun Gao, Zhengming Zhang, Yukun Lai, Cuixia Ma, Yong-Jin Liu, and Hongan Wang. Scenediff: Generative scene-level image retrieval with text and sketch using diffusion models. 2024. 3