CAPYBARA: A Unified Visual Creation Model

Celia Research Team

Abstract

Visual content creation contains two tightly coupled capabilities: generation, which synthesizes images or videos, and editing, which transforms existing visual inputs while preserving identity, structure, and temporal coherence. However, many existing works focus on a single modality or a subset of creation functionalities, resulting in separate solutions with incompatible interfaces that limit unified creation workflows. In this report, we introduce **Capybara**, a unified visual creation foundation model that performs both generation and editing under one framework. We define unified as operating on a single model that accepts multi-modal in-context inputs, including text, images, and videos, and expresses diverse tasks by varying the provided context and instructions. Under this formulation, Capybara supports four major families of creation tasks: (1) text-to-image/video generation; (2) in-context generation conditioned on visual context such as sketches or reference frames; (3) instruction-based editing that applies textual edit instructions to an input image or video; and (4) in-context editing driven by visual references or multi-modal context, enabling consistent transformations across modalities. Capybara is designed to unify these task families with a shared conditioning interface and a single generation backbone, enabling flexible composition of textual intent and visual context for both static and dynamic content creation.

Correspondence: Authors are listed in Project Contributors.

1 Introduction

Recent advances in Multimodal Large Language Models (MLLMs) have rapidly expanded the landscape of visual content creation, which encompasses two tightly coupled capabilities: generation that synthesizes images or videos, and editing that transforms existing visual inputs while preserving identity, structure, and temporal coherence. Image-centric works such as NanoBanana-Pro [1], together with emerging video generators like Kling-Omni [2] and Seeddance2.0 [3], have been widely adopted by academia and industry. However, it still remains largely fragmented: many works focus on a single modality or only a subset of creation functionalities, leading to separate solutions with incompatible interfaces; meanwhile, in-context conditioning for generation (e.g., sketches, subject images, or reference frames) and in-context editing driven by visual references are often introduced as task-specific add-ons, which makes it difficult to build a single system that supports unified creation workflows with diverse multimodal inputs. Naturally, this raises a question:

Can we unify these separate tasks into a single model with a shared multi-modality interface?

Specifically, we propose **Capybara**, a unified visual creation model: a single model accepts multi-modal in-context inputs including text, images, and videos. It also realizes diverse creation behaviors by varying the provided context and instructions, rather than switching architectures or training separate specialists.

We unify visual creation into a single conditioning interface. Each training or inference instance is specified by a common condition package including: (1) a text input (prompt or edit instruction), (2) a primary visual context (image, video, a starting frame, or sparse key-frames), and (3) optional auxiliary conditions (additional references, style/identity examples, or structured controls such as sketches, depth).

(1) Text-to-image/video generation (T2I/T2V). Only text input is provided; the model generates an image or a video from scratch.

(2) In-context generation (e.g., S2I/S2V, C2I/C2V, I2V). In-context generation produces images or videos conditioned on a multi-modal context beyond text. S2I/S2V uses a subject reference image to anchor identity and appearance, while the model synthesizes novel content consistent with the subject. C2I/C2V conditions generation on additional visual prompts, ranging from structured controls (e.g., sketches, layouts, pose, depth/edge maps) to more general visual exemplars. I2V further instantiates this paradigm for temporal synthesis, where the generation is conditioned on a starting frame to ensure temporal consistency.

(3) Instruction-based editing (TI2I/TV2V). Given a source image or video as the primary context, the model applies a textual edit instruction while preserving non-edited regions and maintaining overall fidelity, including identity, structure, and temporal coherence. We also treat dense prediction (e.g., depth, normal, segmentation) as a special case of instruction-based editing, where the instruction requests structured outputs aligned with the input content.

(4) In-context editing (II2I/IV2V/VV2V, propagation). In-context editing is driven by multi-modal context beyond text instructions, including additional reference images/videos, style or identity exemplars, and structured or region-specific guidance. Keyframe propagation is a natural instantiation of in-context editing: given sparse edited keyframes together with unedited context frames, the model propagates the intended changes across time while preserving identity, structure, and temporal coherence.

We reformulate visual creation as the composition of textual conditioning and multi-modal exemplars under a unified backbone. It naturally extends to long-video editing, and with higher throughput could further enable streaming video editing with online updates. The same interface also supports compositional multi-modality workflows, e.g., mixing images and videos as references in one request (identity, motion, structure) for flexible multi-task creation.

2 Data

To support unified visual creation, we curate a joint image–video corpus that provides training signals for text-to-image/video generation, in-context generation, instruction-based editing, and in-context editing. Accordingly, our data includes both standard text-to-image/video pairs for from-scratch synthesis, as well as context-rich tuples that contain text with visual inputs: subject references for S2I/S2V, visual prompts or structured controls (e.g., sketches, layouts, pose, depth/edge maps) for C2I/C2V, starting-frame-conditioned clips for I2V, paired source–instruction–target examples for instruction-based editing, and reference-driven edit tuples (source plus one or more visual exemplars) for in-context editing. For propagation task, we random sample data from the TV2V dataset as our training data.

We employ a systematic multi-stage processing workflow to transform heterogeneous raw collections into high-quality training data. The pipeline consists of: (1) **Quality filtering** using automated classifiers to remove defective content (blur, artifacts, harmful material) and extraneous overlays (watermarks, subtitles); (2) **Semantic deduplication** through embedding-based clustering to retain diverse, non-redundant samples; (3) **Distribution rebalancing** to ensure adequate representation across subject categories, scene types, and visual attributes; (4) **Dense recaptioning** using a bilingual (Chinese/English) vision-language model trained on high-quality annotations, generating detailed descriptions of both dynamic elements (motions, camera movements) and static features (appearances, aesthetics, styles). For editing tasks specifically, we develop large-scale synthesis pipelines generating paired data (source, edited result, instruction).

3 Model Design & Training

3.1 Unified Architecture: Decoupling Understanding from Generation

To build a unified visual creation model, the core challenge is to accept various in-context inputs: text, images, and videos, and fuse them into a single conditioning space that can drive both generation and editing.

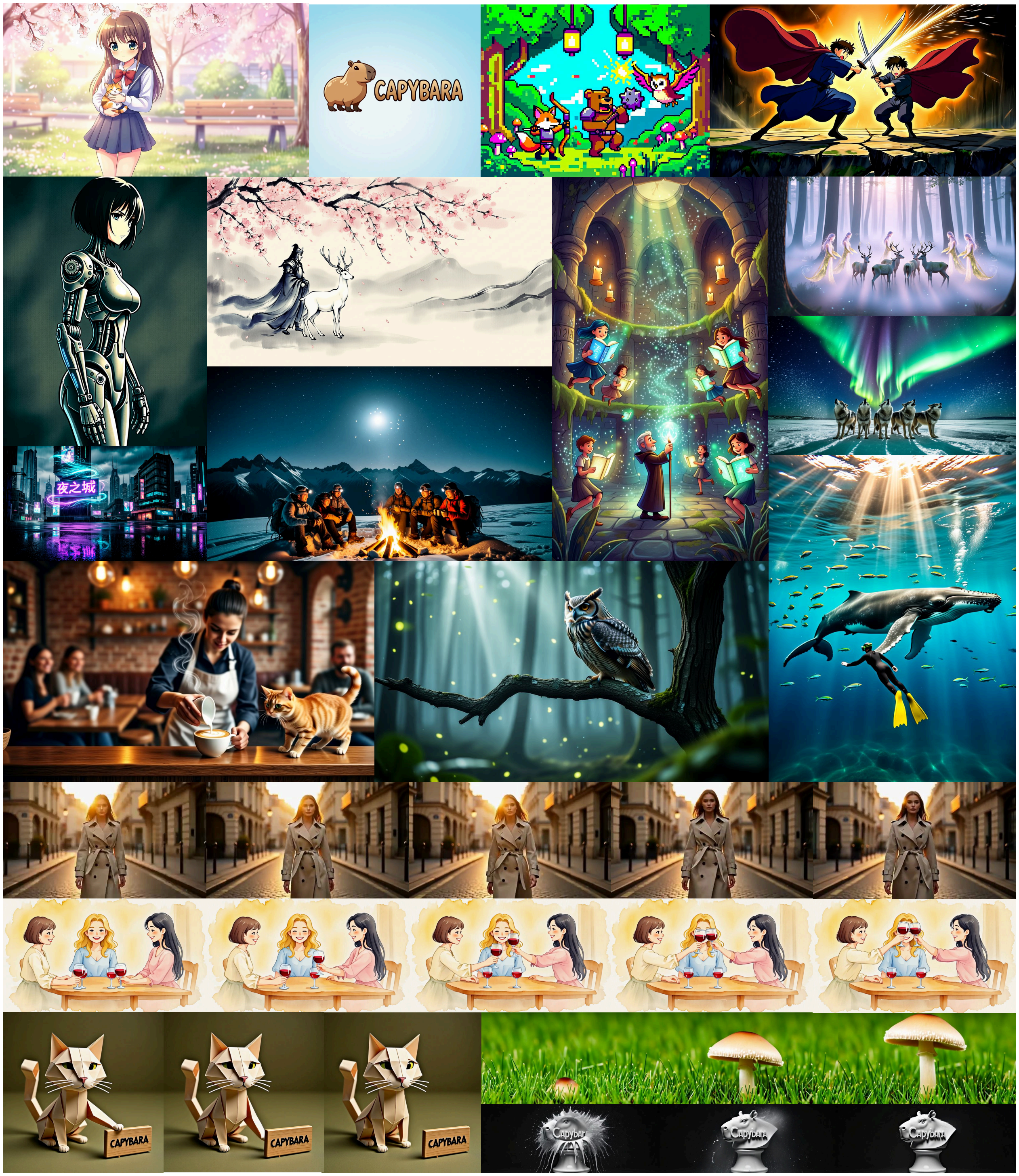

Figure 1 Qualitative results of generation tasks We show two generation tasks under our unified model. The top section presents text-to-image results, illustrating high-fidelity synthesis across diverse styles. The bottom rows show text-to-video results, demonstrating temporally coherent generation with natural motion for both realistic and stylized content.

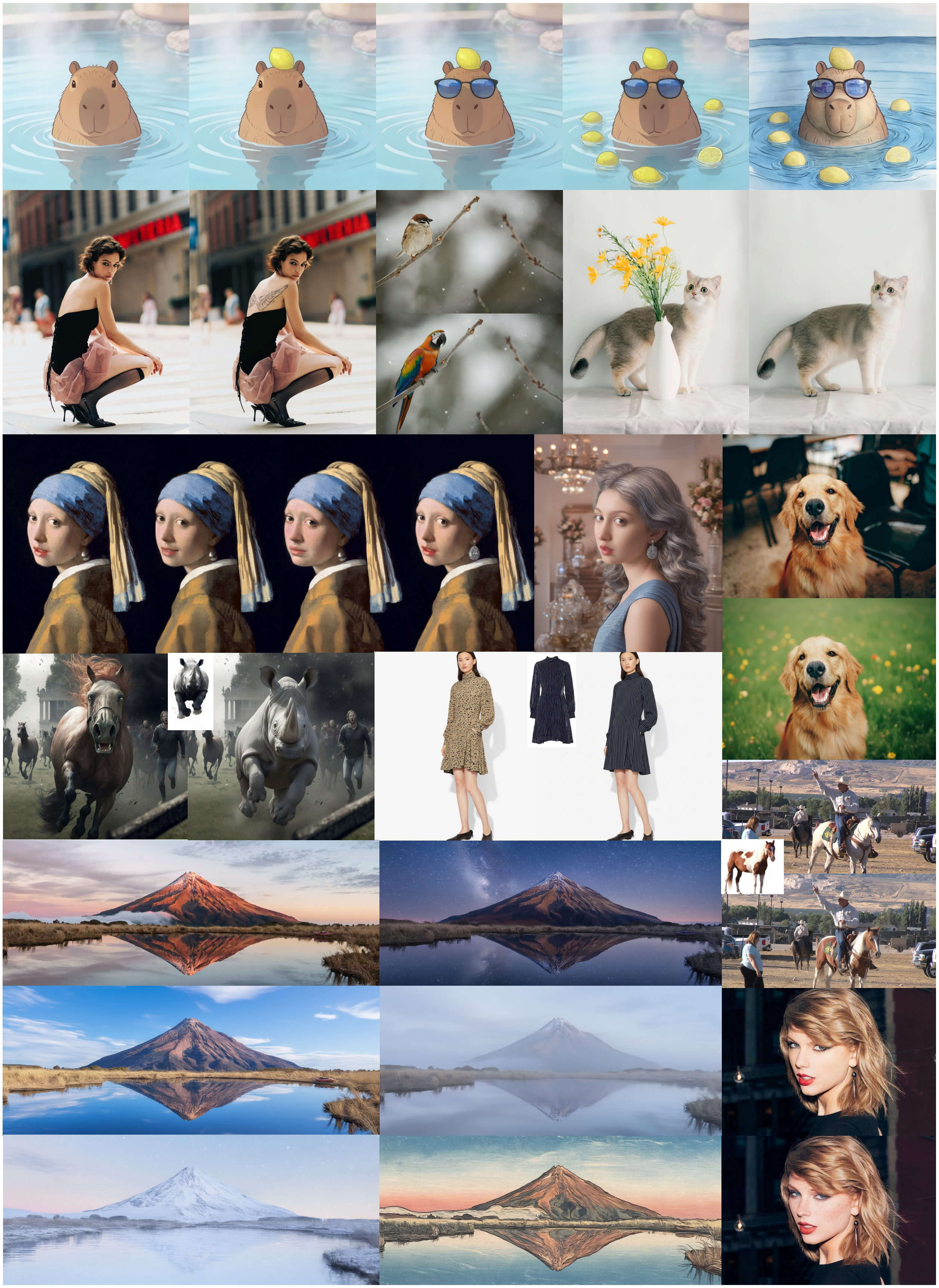

Figure 2 Qualitative results of image editing tasks. We show the results of both instruction-based image editing and in-context image editing. The examples cover local and global edits (e.g., time-of-day and style changes), background replacement, and expression control. We further demonstrate multi-turn editing, where edits are applied sequentially. We also show in-context editing guided by a refenece image.

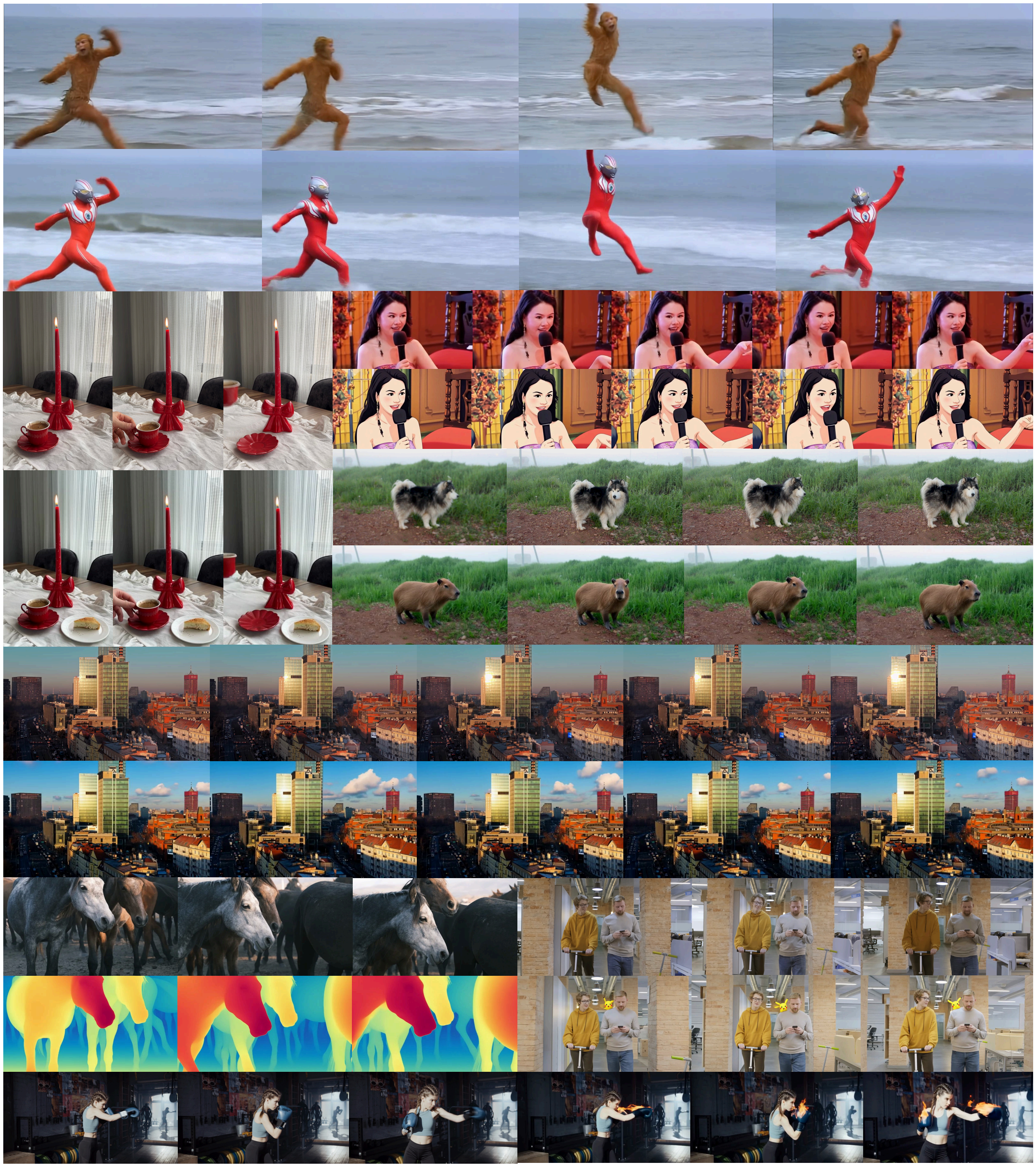

Figure 3 Qualitative results of instruction-based video editing task. We showcase instruction-based editing (TV2V) under our unified creation interface, covering local edits, global edits, dense prediction, and dynamic edits. Each example presents input frames and the edited outputs, highlighting temporally coherent transformations that preserve identity and overall structure.

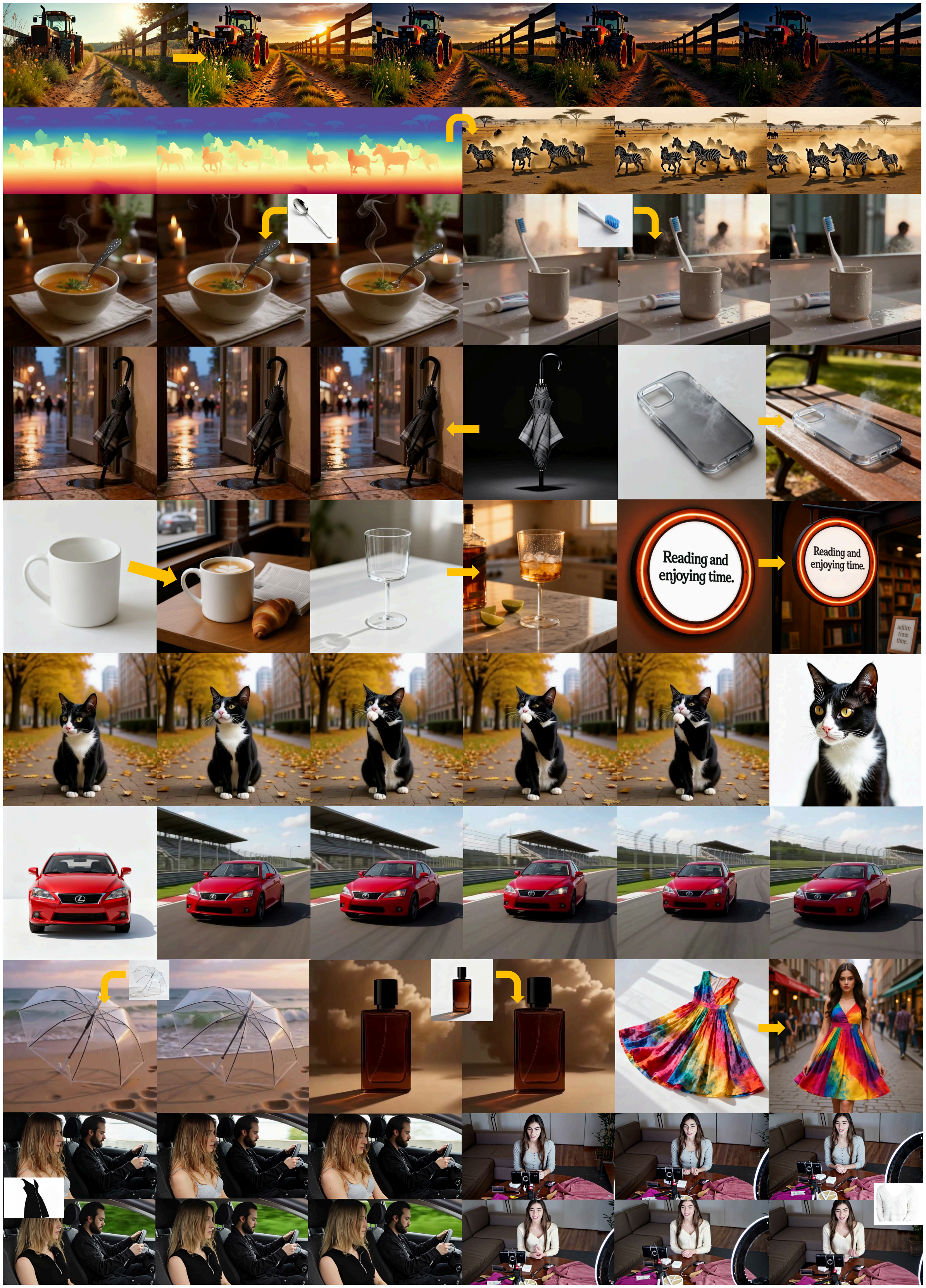

Figure 4 Qualitative results of in-context visual creation. We show in-context generation and in-context editing results , including subject-conditioned generation (S2V/S2I), conditional generation(C2V), image-to-video(I2V), reference-driven editing (II2I/IV2V).

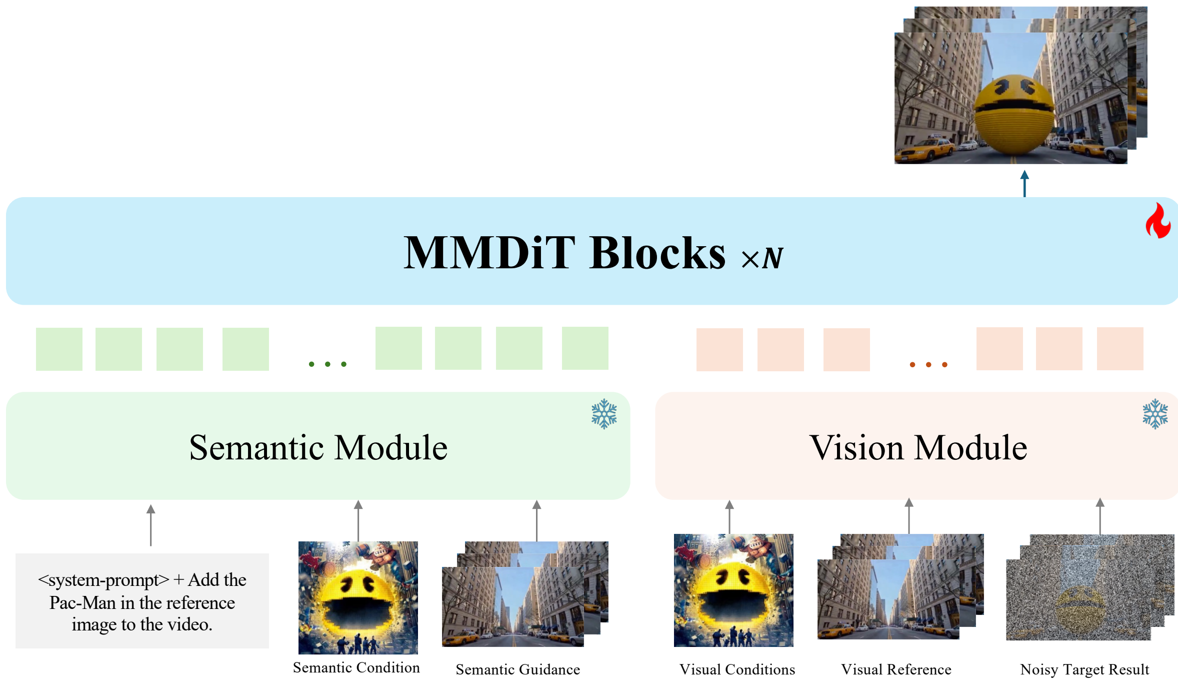

Figure 5 Pipeline overview. Given a system prompt and an instruction (e.g., "Add the Pac-Man in the reference image to the video."), a frozen Semantic Module encodes the textual input into vision-semantic guidance, while a frozen Vision Module extracts visual reference features. These conditions are fused by stacked MMDiT blocks ($\times N$) to denoise the latent representation and synthesize the final output, enabling unified instruction-driven image/video generation and editing.

We therefore choose a dual-stream decoupled architecture that separates multi-modal understanding from diffusion-based synthesis: a Semantic Perception Module focuses on handling user input and reasoning over multi-modal context, while a Visual Integration Module incorporates aligned semantic and visual features into the denoising backbone for high-fidelity synthesis. By structurally decoupling comprehension from generation, we avoid forcing one set of blocks to simultaneously perform high-level interpretation and low-level denoising, enabling a single model to support diverse creation tasks by simply varying the provided context and instructions.

Semantic Module We propose the Semantic Module to consolidate various conditions (e.g., text, image, and video) into a unified latent representation. This module performs contextual reasoning to extract intent-specific features while remaining structurally isolated from the denoising network. This design provides a robust semantic prior, guiding the generative process to strictly adhere to the user's creative intent.

Vision Module The Vision Module is responsible for diffusion denoising process and the precise integration of detailed pixel-level conditions. Complementing the high-level guidance from the Semantic Module, the Vision Module incorporates granular visual conditions. This architecture allocates generative capacity toward faithful reconstruction and spatiotemporal consistency, ensuring strict adherence to multi-modal constraints within a unified framework.

Diffusion Transformer Backbone Our model is initialized from the pre-trained Hunyuan-Video 1.5 [4], inheriting its VAE, DiT architecture, and spatial-temporal modeling capabilities. Building upon this foundation, we introduce a dual-stream decoupled modeling design: a semantic module processes all conditional inputs into unified representations, while a visual module focuses on processing lowlevel feature. This architectural modification enables flexible multi-condition modeling while preserving strong generation priors from pretraining.

3.2 Training Strategy

To establish a unified visual generation framework, we employ a progressive three-stage training curriculum. This strategy is designed to systematically address the distinct challenges associated with unifying various tasks and conditioning signals. The training trajectory evolves the model from robust reconstruction to broad multi-task generalization, culminating in high-fidelity instruction alignment.

Stage I: Reconstruction & In-context generation training. We start from a strong generative prior (initialized from HunyuanVideo-1.5 [4]). The goal is to ensure that conditioning signals produced by the Semantic Module can be reliably consumed by the Vision Module without causing degradation, which is especially critical for editing where non-edited regions must remain consistent. Furthermore, We also trian a mix of standard and in-context generation tasks (S2I/S2V, C2I/C2V, I2V) to introduce pixel-level conditioning capabilities.

Stage II: Editing Tasks Training. After Stage I establishes a stable multi-modality conditioning interface for generation tasks, we expand training to cover editing under the same unified formulation. Specifically, we introduce instruction-based editing (TI2I/TV2V), including dense prediction as a special case where the instruction requests structured outputs aligned with the input content. We further scale to in-context editing (II2I/IV2V/VV2V) driven by additional visual references, style/identity exemplars, and structured or region-specific guidance, and include propagation sequences where sparse edited keyframes supervise temporally consistent change transfer across longer videos.

Stage III: Quality Tuning (QT). Finally, we perform quality tuning to improve instruction adherence, visual fidelity, and temporal stability across both generation and editing. This stage emphasizes difficult cases, such as fine-grained edit locality, identity/appearance preservation, complex multi-modality constraints, and long-range temporal consistency. We collect higher-quality and harder examples and apply targeted tuning to reduce artifacts and strengthen alignment between inputs and outputs.

3.3 Agentic Visual Creation

For iterative video editing, we adopt an **agent-in-the-loop** closed-loop pipeline: **plan** → **edit** → **evaluate/-diagnose** → **refine**. The agent translates a high-level intent into an edit plan that defines what to change (content/style/motion) and what to preserve, with constraints on identity, locality, and temporal scope. It then calls a video editor (e.g., T2V/V2V, optionally with masks/boxes, references, or segment-wise schedules) to generate candidate clips.

A critic scores the results with a small set of metrics—goal alignment, subject consistency, temporal stability, and constraint satisfaction—and outputs structured feedback indicating incorrect changes and where artifacts occur. The agent converts this feedback into tighter instructions and updated controls (prompt edits, strength schedules, temporal windows, region constraints, anchors), and iterates for a few rounds until metrics stabilize or meet a threshold. This is iterative steering via explicit diagnostics, rather than one-shot prompting [5].

4 Related Work

4.1 Diffusion Models for Unified Video Frameworks

With the rapid evolution of Diffusion Transformers (DiTs), video diffusion has moved from specialized text-to-video models toward more general frameworks that unify generation with diverse editing/control interfaces [6–9].

Recent DiT-based generators achieve strong fidelity and scalability for long, high-resolution synthesis from natural language prompts, e.g., OpenSora [10, 11], OpenSora-Plan [12], SanaVideo [13], HunyuanVideo [14], WAN [15], and CogVideoX [16], forming common priors and training recipes for downstream controllable generation.

Video editing methods adapt text-driven image editing paradigms (e.g., Prompt-to-Prompt [17] and InstructPix2Pix [18]) to preserve temporal consistency via correspondence tracking and stable updates, including VideoGrain [19], Pix2Video [20], VideoP2P [21], TokenFlow [22], FateZero [23], InstructVid2Vid [24], CoDeF [25], VEGGIE [26], Ditto [27], InSVIE [28], Senorita [29], LucyEdit [30], OpenVE [31], and MagicEdit [32]. Beyond instruction-based editing, reference- and motion-conditioned control further constrains appearance and dynamics, e.g., IV2V with a reference image [33] and trajectory-conditioned editing such as ReVideo [34].

To reduce fragmentation, unified systems integrate generation and multiple editing/control modes in a single architecture. Some focus on standardized condition interfaces and in-context composition, e.g., VACE [35], UNIC [36], and EditVerse [37]; others leverage MLLMs/VLMs for instruction understanding and project semantics into diffusion backbones, such as UniVideo [38], UniVid [39], Kling-Omni [2], and VINO [40]. However, compressing visual conditions into abstract embeddings can lose fine-grained details, often requiring re-injection of low-level signals (e.g., VAE latents) to recover fidelity [38, 40].

4.2 Visual Encoding in Unified Image and Video Generation Frameworks

A core challenge in unified generation frameworks is how to encode visual inputs such that both semantic understanding and fine-grained controllability are preserved [41–44]. Early diffusion-based systems predominantly adopt CLIP-family encoders [45] to project images into global semantic embeddings. This paradigm is widely used in both image and video models, including Wanx [46], longcat video [47], stable diffusion [48]. While CLIP provides strong text–image alignment, its global pooling design inevitably compresses spatial details, often requiring auxiliary low-level signal injection to recover generation fidelity.

To mitigate this bottleneck, a line of multimodal models explicitly aims to unify visual understanding and generation within a shared encoding space [49–51]. chameleon [52], transfusion [53], show-o [54], EMU3 [55], bagel [56], Emma [57] and lumia [58] integrate visual tokens into large language backbones, enabling joint reasoning and conditional generation. Despite the elegance of unified frameworks, there exists an inherent trade-off between generative fidelity and discriminative power, often resulting in a performance gap when compared to specialized architectures.

More recent works [44, 47, 59–62], such as Qwen-Image [61] and hunyuan image [63] variants, have explored an alternative paradigm by leveraging frozen pre-trained Vision-Language Models as foundational backbones. By keeping the VLM parameters fixed, these methods preserve the model's inherently powerful semantic reasoning while effectively channeling these rich representations to guide the generative process, achieving impressive results in complex instruction following. Further pushing this boundary, Janus-Pro [64] advocates for a decoupled visual processing strategy within such unified architectures. Instead of forcing a single representation to multitask, it employs SigLIP features for high-level semantic understanding and discrete visual tokens for image synthesis. This demonstration suggests that utilizing stronger, specialized encoders within a unified framework can significantly reduce the need for auxiliary injection pathways, effectively bridging the gap between unified flexibility and specialist-level fidelity.

5 Conclusion

We have introduced Capybara, a unified visual creation foundation model that effectively bridges the gap between static and dynamic content generation. By unifying multiple paradigms—ranging from Text-to-Image to complex Video Editing—Capybara excels in precise instruction following, structural stability, and photorealistic visual quality. We presented our core technical innovations in the native unified architecture, intrinsic 3D-aware perception mechanisms, and comprehensive multi-task training strategies, which are effectively integrated to achieve a robust and versatile system. It demonstrates exceptional capabilities in handling complex multi-condition scenarios, maintaining physics-grounded temporal coherence, and enabling a seamless, professional-grade workflow for omni-visual creation.

Acknowledgements

We thank ByteDance Seed for inspiring the paper template. We thank the support from the HKUST Central High Performance Computing Cluster, and we appreciate the generous support from Haoran Yang, Siyue Xie, Cheuk-Him Chau, Yanbin Wei, Chufeng Xiao, Ming Zhang, Xintong Guo, and Binxiao Huang.

Project Contributors

- **Core Contributors:**
  - **Algorithm & Training:** Zhefan Rao, Haoxuan Che, Ziwen Hu, Bin Zou, Yaofang Liu, Xuanhua He
  - **Data Pipeline:** Chong-Hou Choi, Yuyang He, Haoyu Chen, Jingran Su
  - **Benchmarking:** Yanheng Li
  - **Agent:** Meng Chu
- **Contributors:** Chenyang Lei, Guanhua Zhao, Zhaoqing Li, Xichen Zhang, Anping Li, Lin Liu
- **Project Sponsors:** Rui Liu, Dandan Tu
- **Project Leader:** Haoxuan Che

References

[1] Google. Nano banana pro. https://blog.google/innovation-and-ai/products/nano-banana-pro/, 2025. Accessed: 2026-02-13.

[2] Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, Xiao Hu, Xiaohua Hu, Boyuan Jiang, Fangyuan Kong, Hang Li, Jie Li, Qingyu Li, Shen Li, Xiaohan Li, Yan Li, Jiajun Liang, Borui Liao, Yiqiao Liao, Weihong Lin, Quande Liu, Xiaokun Liu, Yilun Liu, Yuliang Liu, Shun Lu, Hangyu Mao, Yunyao Mao, Haodong Ouyang, Wenyu Qin, Wanqi Shi, Xiaoyu Shi, Lianghao Su, Haozhi Sun, Peiqin Sun, Pengfei Wan, Chao Wang, Chenyu Wang, Meng Wang, Qiulin Wang, Runqi Wang, Xintao Wang, Xuebo Wang, Zekun Wang, Min Wei, Tiancheng Wen, Guohao Wu, Xiaoshi Wu, Zhenhua Wu, Da Xie, Yingtong Xiong, Yulong Xu, Sile Yang, Zikang Yang, Weicai Ye, Ziyang Yuan, Shenglong Zhang, Shuaiyu Zhang, Yuanxing Zhang, Yufan Zhang, Wenzheng Zhao, Ruiliang Zhou, Yan Zhou, Guosheng Zhu, and Yongjie Zhu. Kling-omni technical report, 2025. URL https://arxiv.org/abs/2512.16776.

[3] ByteDance. Seedance 2.0. https://seed.bytedance.com/en/seedance2_0, 2026. Accessed: 2026-02-13.

[4] Tencent Hunyuan Foundation Model Team. Hunyuanvideo 1.5 technical report, 2025. URL https://arxiv.org/abs/2511.18870.

[5] Meng Chu, Senqiao Yang, Haoxuan Che, Suiyun Zhang, Xichen Zhang, Shaozuo Yu, Haokun Gui, Zhefan Rao, Dandan Tu, Rui Liu, and Jiaya Jia. Visiondirector: Vision-language guided closed-loop refinement for generative image synthesis, 2026. URL https://arxiv.org/abs/2512.19243.

[6] Xuanhua He, Quande Liu, Zixuan Ye, Weicai Ye, Qiulin Wang, Xintao Wang, Qifeng Chen, Pengfei Wan, Di Zhang, and Kun Gai. Fulldit2: Efficient in-context conditioning for video diffusion transformers. arXiv preprint arXiv:2506.04213, 2025.

[7] Zixuan Ye, Quande Liu, Cong Wei, Yuanxing Zhang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhan Luo. Visual-aware cot: Achieving high-fidelity visual consistency in unified models. arXiv preprint arXiv:2512.19686, 2025.

[8] Xiangpeng Yang, Ji Xie, Yiyuan Yang, Yan Huang, Min Xu, and Qiang Wu. Unified video editing with temporal reasoner. arXiv preprint arXiv:2512.07469, 2025.

[9] Zhoujie Fu, Xianfang Zeng, Jinghong Lan, Xinyao Liao, Cheng Chen, Junyi Chen, Jiacheng Wei, Wei Cheng, Shiyu Liu, Yunuo Chen, et al. imontage: Unified, versatile, highly dynamic many-to-many image generation. arXiv preprint arXiv:2511.20635, 2025.

[10] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404, 2024.

[11] Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, Yuhui Wang, Anbang Ye, Gang Ren, Qianran Ma, Wanying Liang, Xiang Lian, Xiwen Wu, Yuting Zhong, Zhuangyan Li, Chaoyu Gong, Guojun Lei, Leijun Cheng, Limin Zhang, Minghao Li, Ruijie Zhang, Silan Hu, Shijie Huang, Xiaokang Wang, Yuanheng Zhao, Yuqi Wang, Ziang Wei, and Yang You. Open-sora 2.0: Training a commercial-level video generation model in $200k. arXivpreprintarXiv: 2503.09642$, 2025.

[12] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. arXiv preprint arXiv:2412.00131, 2024.

[13] Junsong Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, Junyu Chen, Shuai Yang, Xianbang Wang, Yicheng Pan, Daquan Zhou, Huan Ling, et al. Sana-video: Efficient video generation with block linear diffusion transformer. arXiv preprint arXiv:2509.24695, 2025.

[14] 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.

[15] 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.

[16] 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.

[17] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022. URL https://arxiv.org/abs/2208.01626.

[18] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. URL https://arxiv.org/abs/2211.09800.

[19] Xiangpeng Yang, Linchao Zhu, Hehe Fan, and Yi Yang. VideoGrain: Modulating space-time attention for multi-grained video editing. In ICLR, 2025.

[20] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2Video: Video editing using image diffusion. In ICCV, 2023.

[21] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-P2P: Video editing with cross-attention control. In CVPR, 2024.

[22] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. TokenFlow: Consistent diffusion features for consistent video editing. In ICLR, 2024.

[23] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. FateZero: Fusing attentions for zero-shot text-based video editing. In ICCV, 2023.

[24] Bosheng Qin, Juncheng Li, Siliang Tang, Tat-Seng Chua, and Yueting Zhuang. Instructvid2vid: Controllable video editing with natural language instructions. In ICME, 2024.

[25] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. CoDeF: Content deformation fields for temporally consistent video processing. In CVPR, 2024.

[26] Shoubin Yu, Difan Liu, Ziqiao Ma, Yicong Hong, Yang Zhou, Hao Tan, Joyce Chai, and Mohit Bansal. VEGGIE: Instructional editing and reasoning video concepts with grounded generation. arXiv preprint arXiv:2503.14350, 2025.

[27] Qingyan Bai, Qiuyu Wang, Hao Ouyang, Yue Yu, Hanlin Wang, Wen Wang, Ka Leong Cheng, Shuailei Ma, Yanhong Zeng, Zichen Liu, Yinghao Xu, Yujun Shen, and Qifeng Chen. Scaling instruction-based video editing with a high-quality synthetic dataset. arXiv preprint arXiv:2510.15742, 2025.

[28] Yuhui Wu, Liyi Chen, Ruibin Li, Shihao Wang, Chenxi Xie, and Lei Zhang. InsViE-1M: Effective instruction-based video editing with elaborate dataset construction. In ICCV, 2025.

[29] Bojia Zi, Penghui Ruan, Marco Chen, Xianbiao Qi, Shaozhe Hao, Shihao Zhao, Youze Huang, Bin Liang, Rong Xiao, and Kam-Fai Wong. Señorita-2M: A high-quality instruction-based dataset for general video editing by video specialists. arXiv preprint arXiv:2502.06734, 2025.

[30] DecartAI Team. Lucy edit: Open-weight text-guided video editing. 2025. URL https://d2drjpuinn46lb.cloudfront.net/Lucy_Edit__High_Fidelity_Text_Guided_Video_Editing.pdf.

[31] Haoyang He, Jie Wang, Jiangning Zhang, Zhucun Xue, Xingyuan Bu, Qiangpeng Yang, Shilei Wen, and Lei Xie. Openve-3m: A large-scale high-quality dataset for instruction-guided video editing. arXiv preprint arXiv:2512.07826, 2025.

[32] Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. Magicedit: High-fidelity and temporally coherent video editing. arXiv preprint arXiv:2308.14749, 2023.

[33] Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. AnyV2V: A tuning-free framework for any video-to-video editing tasks. arXiv preprint arXiv:2403.14468, 2024.

[34] Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. Revideo: Remake a video with motion and content control. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, Advances in Neural Information Processing Systems, volume 37, pages 18481–18505. Curran Associates, Inc., 2024. doi: 10.52202/079017-0586. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/20e6b4dd2b1f82bc599c593882f67f75-Paper-Conference.pdf.

[35] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 17191–17202, 2025.

[36] Zixuan Ye, Xuanhua He, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Qifeng Chen, and Wenhan Luo. Unic: Unified in-context video editing, 2025. URL https://arxiv.org/abs/2506.04216.

[37] Xuan Ju, Tianyu Wang, Yuqian Zhou, He Zhang, Qing Liu, Nanxuan Zhao, Zhifei Zhang, Yijun Li, Yuanhao Cai, Shaoteng Liu, Daniil Pakhomov, Zhe Lin, Soo Ye Kim, and Qiang Xu. Editverse: Unifying image and video editing and generation with in-context learning. arXiv preprint arXiv:2509.20360, 2025. URL https://arxiv.org/abs/2509.20360.

[38] Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. Univideo: Unified understanding, generation, and editing for videos. arXiv preprint arXiv:2510.08377, 2025.

[39] Jiabin Luo, Junhui Lin, Zeyu Zhang, Biao Wu, Meng Fang, Ling Chen, and Hao Tang. Univid: The open-source unified video model. arXiv preprint arXiv:2509.24200, 2025.

[40] Junyi Chen, Tong He, Zhoujie Fu, Pengfei Wan, Kun Gai, and Weicai Ye. Vino: A unified visual generator with interleaved omnimodal context. arXiv preprint arXiv:2601.02358, 2026.

[41] Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Feize Wu, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, et al. Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. arXiv preprint arXiv:2510.16888, 2025.

[42] Shanshan Zhao, Xinjie Zhang, Jintao Guo, Jiakui Hu, Lunhao Duan, Minghao Fu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, et al. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. arXiv preprint arXiv:2505.02567, 2025.

[43] Run Luo, Xiaobo Xia, Lu Wang, Longze Chen, Renke Shan, Jing Luo, Min Yang, and Tat-Seng Chua. Next-omni: Towards any-to-any omnimodal foundation models with discrete flow matching. arXiv preprint arXiv:2510.13721, 2025.

[44] Zhiyu Tan, Hao Yang, Luozheng Qin, Jia Gong, Mengping Yang, and Hao Li. Omni-video: Democratizing unified video understanding and generation. arXiv preprint arXiv:2507.06119, 2025.

[45] 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.

[46] 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.

[47] Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, et al. Longcat-video technical report. arXiv preprint arXiv:2510.22200, 2025.

[48] 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.

[49] 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.

[50] Shengbang Tong, David Fan, Jiachen Li, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 17001–17012, 2025.

[51] 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.

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

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] Xin He, Longhui Wei, Jianbo Ouyang, Minghui Liao, Lingxi Xie, and Qi Tian. Emma: Efficient multimodal understanding, generation, and editing with a unified architecture, 2025. URL https://arxiv.org/abs/2512.04810.

[58] Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. arXiv preprint arXiv:2510.06308, 2025.

[59] Xiang Wang, Zhifei Zhang, He Zhang, Zhe Lin, Yuqian Zhou, Qing Liu, Shiwei Zhang, Yijun Li, Shaoteng Liu, Haitian Zheng, et al. Hbridge: H-shape bridging of heterogeneous experts for unified multimodal understanding and generation. arXiv preprint arXiv:2511.20520, 2025.

[60] Bin Xia, Bohao Peng, Yuechen Zhang, Junjia Huang, Jiyang Liu, Jingyao Li, Haoru Tan, Sitong Wu, Chengyao Wang, Yitong Wang, et al. Dreamomni2: Multimodal instruction-based editing and generation. arXiv preprint arXiv:2510.06679, 2025.

[61] 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.

[62] Siqi Kou, Jiachun Jin, Zetong Zhou, Ye Ma, Yugang Wang, Quan Chen, Peng Jiang, Xiao Yang, Jun Zhu, Kai Yu, et al. Think-then-generate: Reasoning-aware text-to-image diffusion with llm encoders. arXiv preprint arXiv:2601.10332, 2026.

[63] Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. arXiv preprint arXiv:2509.23951, 2025.

[64] 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.