PISCO: Precise Video Instance Insertion with Sparse Control

Anonymous CVPR submission

Paper ID

Abstract

*AI video generation is moving beyond general generation, which often relies on exhaustive prompt engineering and cherry-picking, toward fine-grained, controllable generation and high-fidelity post-processing. In **professional AI-assisted filmmaking**, a key capability is **video instance insertion**: placing a specific object into an existing video at a designated spatial-temporal location while preserving scene dynamics and physically consistent interactions such as shadows and reflections. We present* `PISCO`*, a video diffusion framework for precise video instance insertion under arbitrary sparse keyframe control.* `PISCO` *supports single-frame, start-and-end-frame, or sparse keyframes at arbitrary timestamps, and automatically propagates appearance, motion, and interaction. To stabilize sparse conditioning, we introduce Variable-Information Guidance (VIG) and Distribution-Preserving Temporal Masking (DPTM), together with geometry-aware conditioning. We further construct* `PISCO-Bench`*, a benchmark with verified instance annotations and paired clean background videos. Experiments show that* `PISCO` *consistently outperforms representative baselines and improves monotonically as additional control signals are provided.*

1. Introduction

Recent advances in large-scale video generative models [46] have enabled high-fidelity video generation with increasingly realistic motion. As these models continue to evolve, the focus of video generation is shifting beyond producing visually plausible content toward highly controllable video generation and editing [41], with the long-term goal of enabling **Hollywood-grade AI-assisted filmmaking** [37, 39]. In this new paradigm, the priority is no longer to **cherry-pick** a barely acceptable output from numerous generations, but to reliably **achieve precise user intent** with minimal iteration.

A particularly demanding yet under-explored capability in this context is **precise video instance insertion**: inserting a specific object into an existing video at a precise user-specified spatial location and temporal position, while preserving the identity and dynamics of the original video. In professional visual effects and post-production workflows, instance insertion is not merely about adding a visually plausible object, but about achieving precise control with minimal iteration. In this setting, such precision entails several requirements: ❶ **Instance-level controllability**, enabling users to explicitly specify when and where an object appears; ❷ **Physically plausible temporal propagation**, where the inserted object automatically evolves over time with coherent pose and motion, following physically reasonable dynamics; ❸ **Consistent scene adaptation**, adjusting the background for insertion-induced effects such as shadows and reflections; ❹ **Background dynamics preservation**, maintaining original motions, identities, and temporal patterns without disruption; ❺ **Low-effort user interaction**, where achieving the above goals does not require dense per-frame annotations or manual editing across the entire video.

Despite recent advances [57], these requirements remain largely unmet by existing paradigms. **Video inpainting** [42, 67] demands exhaustive per-frame mask annotations (❺) and its copy-and-fill nature struggles with holistic scene adaptation such as realistic shadowing (❸). **Video-to-video editing** [22, 54] can propagate appearance but lacks fine-grained spatiotemporal controllability for precise object placement (❶). **Agentic pipelines** (image editing + I2V) [18, 24] discard temporal context, often causing severe background drift and disrupted scene dynamics (❹). Finally, **geometry-heavy 4D approaches** [23, 36] improve spatial reasoning but incur substantial computational overhead, rely on fragile geometric priors, and struggle with complex dynamic object motions (❷).

To overcome these limitations, we propose `PISCO` (Precise Instance insertion with Sparse COntrol), a video diffusion framework designed for professional-grade insertion with minimal user effort. `PISCO` allows users to specify instance conditions at arbitrary sparse keyframes (e.g., first frame only, or first & last frames). Built upon the Wan backbone [46], it incorporates a multi-channel context adapter to

ingest instance RGB, masks, depth, and an explicit availability signal. However, sparse control introduces a critical challenge: naïvely masking missing frames causes severe distribution shifts in pretrained temporal VAEs, leading to flickering and rendering artifacts. PISCO solves this via **Variable-Information Guidance (VIG)** to train robustly across varying condition densities, and **Distribution-Preserving Temporal Masking (DPTM)** to stabilize VAE inputs. Coupled with geometry-aware conditioning, these mechanisms enable coherent propagation and scene-consistent interaction.

To rigorously evaluate this task, we construct PISCO-Bench, a curated benchmark derived from BURST [4], featuring verified masks and paired clean background videos generated via side-effect-aware removal [38]. Extensive reference-based and reference-free experiments show that PISCO consistently outperforms strong inpainting and video editing baselines, while exhibiting monotonic improvements as additional sparse control frames are provided. Our contributions are summarized as follows:

- We introduce PISCO, a video diffusion framework for precise video instance insertion under arbitrary sparse keyframe control.
- We propose VIG and DPTM, two dedicated mechanisms that address sparse-control temporal ambiguity and temporal-VAE distribution shift.
- We construct PISCO-Bench and demonstrate that PISCO significantly outperforms representative baselines while benefiting consistently from additional sparse control signals.

2. Related Work

Video Inpainting for Video Instance Insertion. Video inpainting [42] has been widely studied for object removal, insertion, and restoration. Earlier methods relied on 3D-CNNs [9, 19, 48], optical flow [13, 25, 28, 30, 56, 68], and later Transformers [8, 33, 34, 66] to improve temporal coherence. More recent approaches adopt diffusion models and video foundation models [6, 14–16, 26, 27, 29, 46, 47, 57, 63, 67]. However, standard inpainting typically requires dense per-frame masks at inference, which is impractical for precise insertion of non-existent dynamic objects. In contrast, our method operates from sparse keyframe conditions.

Reference-Guided Video-to-Video Editing. Reference-guided video editing uses image or text conditions to guide synthesis. Recent frameworks such as UniVideo [54], VACE [22], VideoDirector [52], and ContextFlow [11] expand controllable video generation, while object-centric editing has also been explored for insertion, replacement, and storytelling [12, 17, 45]. Despite these advances, most methods still behave as global video-to-video translation systems and lack the fine-grained spatial-temporal control required for precise instance insertion.

Image-to-Video Generation and Propagation. Another line of work follows a generate-then-animate paradigm: edit one or several keyframes using image editing models [7, 35, 55], then animate them with I2V or transition models [18, 24, 31, 40, 50, 51, 59]. Motion-specialized architectures further improve motion control [43, 60, 64]. However, because these methods regenerate the full video rather than conditioning on the original one, they often suffer from background hallucination and temporal drift. Our approach instead preserves the original background video as a persistent condition.

4D Reconstruction and Manipulation. To improve geometric consistency, some methods lift videos into 3D or 4D representations. Place Anything into Any Video [36], Anything in Any Scene [5], and InsertAnywhere [23] exploit depth, flow, or scene geometry for physical placement and occlusion reasoning. While effective for geometry-aware editing, these methods are typically computationally heavy, sensitive to geometric priors, and less suitable for diverse dynamic insertions. Our method instead uses lightweight geometry-aware conditioning to improve compositing and occlusion handling without requiring full 4D reconstruction.

3. Method

3.1. Problem Formulation

Given paired videos $\{\hat{V}, V\}$ of length T, where $\hat{V} = \{\hat{V}_t\}_{t=1}^{T}$ contains a foreground instance and $V = \{V_t\}_{t=1}^{T}$ depicts the same scene without that instance, our goal is to synthesize an edited video $\tilde{V}$ in which the instance is naturally inserted into V with accurate spatial placement, coherent motion, and plausible occlusion relationships.

We represent the instance-side conditions as an RGB instance clip $I = \{I_t\}_{t=1}^{T}$, its mask $M = \{M_t\}_{t=1}^{T}$, and instance depth $D_I = \{D_{I,t}\}_{t=1}^{T}$. We additionally compute background depth D_V from V to provide geometric cues.

Availability mask. To model sparse user guidance, we define an availability mask $A = \{A_t\}_{t=1}^{T}$, where $A_t \in \{0, 1\}$ indicates whether instance-side conditions are available at time t. The masked conditions are

$$I^A = A \odot I, \qquad M^A = A \odot M, \qquad D_I^A = A \odot D_I, \quad (1)$$

while D_V remains fully available.

Conditional diffusion objective. Let z_0 denote the target-video latent from a pretrained temporal VAE, and let z_t be the noisy latent at diffusion step t. We train a denoising

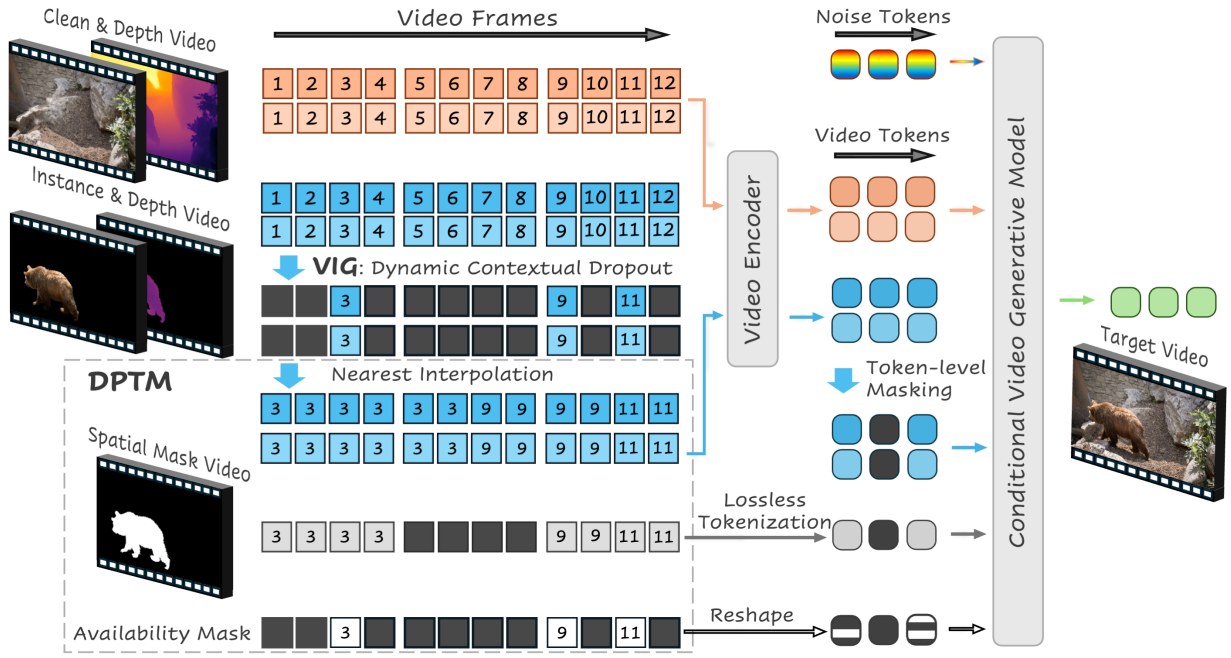

Figure 1. **Overview of PISCO.** PISCO performs precise video instance insertion under sparse keyframe control. VIG exposes the model to diverse conditioning densities, while DPTM stabilizes sparse conditioning for pretrained temporal VAEs. RGB, mask, depth, and availability cues are injected through a multi-channel context adapter.

model ϵ_θ with

$$\mathcal{L} = \mathbb{E}_{z_0,\epsilon,t,A}\left[\left\|\epsilon - \epsilon_\theta\left(z_t,\, t,\, V,\, D_V,\, I^A,\, D_I^A,\, M^A,\, A\right)\right\|_2\right]. \tag{2}$$

In practice, ϵ_θ is implemented with the Wan backbone [46] and a modified VACE context adapter [22].

3.2. Core Design

Variable-Information Guidance. To ensure robustness across different user-control densities, we introduce **Variable-Information Guidance (VIG)**, which samples availability masks during training. The model is exposed to a mixture of extremely sparse, sparse, dense, and fully supervised regimes, enabling it to propagate motion under limited guidance while retaining appearance fidelity when stronger supervision is available.

Distribution-Preserving Temporal Masking. Modern video generators often rely on pretrained temporal VAEs [1, 10, 32, 46, 49, 58, 65]. Directly zero-masking unavailable frames causes severe distribution shift. We therefore introduce **Distribution-Preserving Temporal Masking (DPTM)**, which first fills missing frames in pixel space to maintain natural encoder statistics, then applies masking in token space and provides an aligned availability channel to explicitly indicate which conditions are observed.

Geometry- and appearance-aware training. We further improve insertion realism with depth-aware conditioning, pseudo-amodal instance augmentation [2, 3], and relighting augmentation using IC-Light [61]. These components improve depth ordering, occlusion handling, and illumination compatibility without requiring heavy 4D reconstruction.

4. Experiments

Benchmark and settings. We introduce PISCO-Bench, a curated benchmark of 100 diverse real-world videos derived from BURST [4]. We manually refine the instance masks and use ROSE [38] to generate paired clean background videos. We evaluate both PISCO-1.3B and PISCO-14B under **First-frame** and **First & Last-frame** control, and additionally report a **Five-Frame** setting for PISCO. Since no open-source method supports identical sparse keyframe conditioning, we adapt representative baselines from three categories: **Image Editing + I2V Generation**, **Video Inpainting**, and **Reference-guided V2V Editing**. All methods are evaluated on 49-frame videos at 832×480 resolution.

Reference-based evaluation. Using ground-truth pairs $(V, \hat{V})$, we evaluate generated videos with FVD [44], LPIPS [62], PSNR [21], and SSIM [53], measured on both the whole video and the foreground region. As shown in Table 1, PISCO consistently outperforms all baselines. Image Editing + I2V suffers from background hallucination, inpainting methods preserve backgrounds but lag in temporal coherence and object fidelity, and VACE is the strongest baseline overall. Still, PISCO-14B with First & Last control

Table 1. **Reference-based quantitative results.** Comparisons on both whole-video and foreground regions. Gray rows indicate the Five-Frame setting, reported to demonstrate scalability with additional sparse controls.

Method	Whole Video				Foreground			
	FVD ↓	LPIPS ↓	PSNR ↑	SSIM ↑	FVD ↓	LPIPS ↓	PSNR ↑	SSIM ↑
Image Editing + I2V Generation								
First Only	826	0.451	15.47	0.55	297	0.030	30.24	0.97
First + Last	624	0.392	16.44	0.56	250	0.030	30.38	0.98
Video Inpainting								
CoCoCo	590	0.191	23.62	0.80	398	0.031	30.26	0.97
VideoPainter	524	0.154	23.11	0.78	384	0.035	29.27	0.97
Video-to-Video Editing								
$VACE_{14B}$	371	0.103	25.55	0.88	273	0.028	30.55	0.98
UniVideo	485	0.211	19.22	0.61	310	0.031	29.21	0.97
PISCO $_{1.3B}$ (Ours)								
First Only	398	0.121	25.35	0.87	243	0.029	30.51	0.98
First + Last	269	0.103	27.01	0.88	171	0.024	32.99	0.98
Five Frames	172	0.089	28.53	0.89	104	0.018	35.07	0.98
PISCO $_{14B}$ (Ours)								
First Only	337	0.116	24.81	0.88	222	0.029	30.61	0.97
First + Last	204	0.097	26.58	0.89	138	0.022	33.58	0.98
Five Frames	136	0.084	28.01	0.90	75	0.015	35.94	0.98

Table 2. **Reference-free VBench comparison.** PISCO achieves the strongest overall perceptual quality among standard settings and continues to improve as additional sparse keyframes are provided.

Method	Background Consistency	Subject Consistency	Aesthetic Quality	Imaging Quality	Motion Smoothness	Overall Consistency	Temporal Flickering	Temporal Style	Average
Image Editing + I2V Generation									
First Only	91.43	83.86	48.83	59.01	97.89	15.11	95.89	15.11	63.39
First + Last	92.28	85.09	49.47	59.71	98.18	15.58	96.26	15.58	64.02
Video Inpainting									
CoCoCo	94.63	89.55	47.81	55.76	98.67	14.61	97.69	14.61	64.17
VideoPainter	94.51	89.14	47.82	57.68	98.97	13.85	97.77	13.85	64.20
Video-to-Video Editing									
$VACE_{14B}$	94.21	90.29	48.69	60.95	98.90	14.87	97.56	14.87	65.04
UniVideo	94.04	89.88	49.41	60.84	98.80	15.51	97.01	15.92	65.18
PISCO $_{1.3B}$ (Ours)									
First Only	93.72	87.16	47.70	60.67	98.85	14.92	97.54	14.92	64.43
First + Last	94.07	91.33	49.48	61.18	98.86	15.58	97.51	15.58	65.45
Five Frames	94.23	91.45	50.01	62.09	98.86	15.61	97.67	15.61	65.89
PISCO $_{14B}$ (Ours)									
First Only	93.84	87.26	48.24	60.80	98.79	15.14	97.26	15.14	64.56
First + Last	94.20	91.57	50.08	62.00	98.79	15.64	97.21	15.64	65.64
Five Frames	94.42	91.98	51.45	62.87	98.89	15.82	97.34	15.57	66.04

reduces whole-video FVD from 371 to 204 and foreground FVD from 273 to 138. Additional sparse controls further improve performance in the Five-Frame setting.

Reference-free evaluation. For reference-free evaluation, we employ VBench [20] and adapt it with instance masks to separately assess subject and background consistency. As shown in Table 2, PISCO achieves the strongest overall perceptual quality among standard settings. In particular, it yields better subject consistency and visual quality than prior baselines, while remaining competitive on background preservation. Performance further improves when more sparse controls are provided.

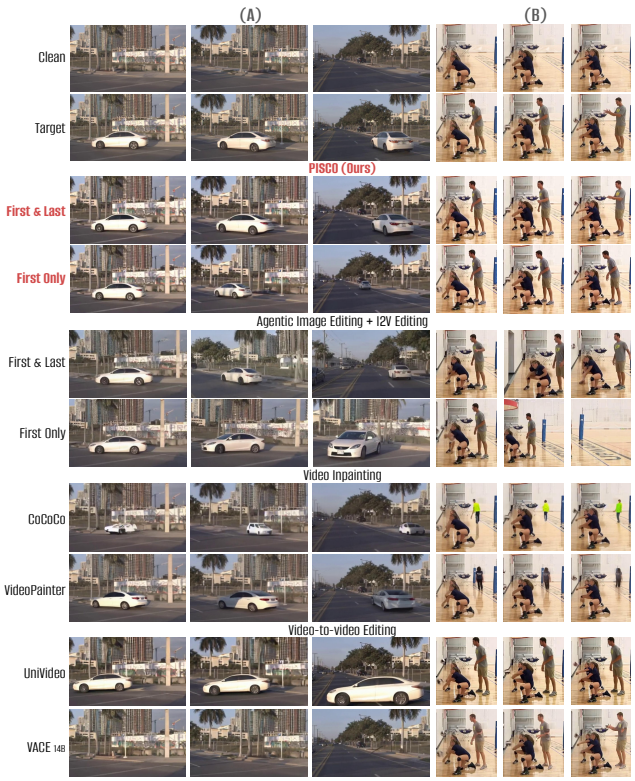

Figure 2. **Qualitative comparison.** Existing methods struggle with background preservation, instance fidelity, or spatial-temporal alignment. In contrast, PISCO produces stronger object fidelity, more consistent scale and motion, and better preservation of original scene dynamics, especially under sparse First & Last control.

Qualitative results. Figure 2 shows that prior methods often fail in different ways: agentic pipelines tend to regenerate the scene and drift in the background, inpainting methods weaken or blur the inserted object, and V2V baselines may lose precise long-term control. In contrast, PISCO achieves stronger spatiotemporal alignment and scene consistency, especially when sparse endpoint constraints are provided.

5. Conclusion

We presented PISCO, a video diffusion framework for precise video instance insertion under sparse user control. By combining Variable-Information Guidance, Distribution-Preserving Temporal Masking, and geometry-aware conditioning, PISCO addresses key challenges in sparse-control video generation and significantly outperforms representative inpainting, video editing, and agentic baselines on PISCO-Bench. Moreover, it improves monotonically as more sparse control frames are provided, supporting a controllable and scalable paradigm for instance-centric video editing.

References

[1] 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 preprint arXiv:2501.03575*, 2025. 3

[2] Jiayang Ao, Qiuhong Ke, and Krista A Ehinger. Image amodal completion: A survey. *Computer Vision and Image Understanding*, 229:103661, 2023. 3

[3] Jiayang Ao, Yanbei Jiang, Qiuhong Ke, and Krista A Ehinger. Open-world amodal appearance completion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6490–6499, 2025. 3

[4] Ali Athar, Jonathon Luiten, Paul Voigtlaender, Tarasha Khurana, Achal Dave, Bastian Leibe, and Deva Ramanan. Burst: A benchmark for unifying object recognition, segmentation and tracking in video. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1674–1683, 2023. 2, 3

[5] Chen Bai, Zeman Shao, Guoxiang Zhang, Di Liang, Jie Yang, Zhuorui Zhang, Yujian Guo, Chengzhang Zhong, Yiqiao Qiu, Zhendong Wang, et al. Anything in any scene: Photorealistic video object insertion. *arXiv preprint arXiv:2401.17509*, 2024. 2

[6] Yuxuan Bian, Zhaoyang Zhang, Xuan Ju, Mingdeng Cao, Liangbin Xie, Ying Shan, and Qiang Xu. Videopainter: Any-length video inpainting and editing with plug-and-play context control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2025. 2

[7] Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Shijie Huang, Zhaohui Hou, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025. 2

[8] Jiayin Cai, Changlin Li, Xin Tao, Chun Yuan, and Yu-Wing Tai. Devit: Deformed vision transformers in video inpainting. In *Proceedings of the 30th ACM international conference on multimedia*, pages 779–789, 2022. 2

[9] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9066–9075, 2019. 2

[10] Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxiang Hao, Hui Wu, Zhichao Lai, Yifei Hu, et al. Goku: Flow based video generative foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23516–23527, 2025. 3

[11] Yiyang Chen, Xuanhua He, Xiujun Ma, and Yue Ma. Contextflow: Training-free video object editing via adaptive context enrichment. *arXiv preprint arXiv:2509.17818*, 2025. 2

[12] Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025. 2

[13] Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In *European Conference on Computer Vision*, pages 713–729. Springer, 2020. 2

[14] Dylan Green, William Harvey, Saeid Naderiparizi, Matthew Niedoba, Yunpeng Liu, Xiaoxuan Liang, Jonathan Lavington, Ke Zhang, Vasileios Lioutas, Setareh Dabiri, et al. Semantically consistent video inpainting with conditional diffusion models. *arXiv preprint arXiv:2405.00251*, 2024. 2

[15] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[16] Yuwei Guo, Ceyuan Yang, Anyi Rao, Chenlin Meng, Omer Bar-Tal, Shuangrui Ding, Maneesh Agrawala, Dahua Lin, and Bo Dai. Keyframe-guided creative video inpainting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13009–13020, 2025. 2

[17] Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*, 2023. 2

[18] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: controllable image-to-video generation with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18219–18228, 2022. 1, 2

[19] Yuan-Ting Hu, Heng Wang, Nicolas Ballas, Kristen Grauman, and Alexander G Schwing. Proposal-based video completion. In *European Conference on Computer Vision*, pages 38–54. Springer, 2020. 2

[20] 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. 4

[21] Bernd Jähne. *Digital image processing*. Springer, 2005. 3

[22] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. 1, 2, 3

[23] Hoiyeong Jin, Hyojin Jang, Jeongho Kim, Junha Hyung, Kinam Kim, Dongjin Kim, Huijin Choi, Hyeonji Kim, and Jaegul Choo. Insertanywhere: Bridging 4d scene geometry and diffusion models for realistic video object insertion. *arXiv preprint arXiv:2512.17504*, 2025. 1, 2

[24] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024. 1, 2

[25] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5792–5801, 2019. 2

[26] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 2

[27] Minhyeok Lee, Suhwan Cho, Chajin Shin, Jungho Lee, Sunghun Yang, and Sangyoun Lee. Video diffusion models are strong video inpainter. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4526–4533, 2025. 2

[28] Ang Li, Shanshan Zhao, Xingjun Ma, Mingming Gong, Jianzhong Qi, Rui Zhang, Dacheng Tao, and Ramamohanarao Kotagiri. Short-term and long-term context aggregation network for video inpainting. In *European Conference on Computer Vision*, pages 728–743. Springer, 2020. 2

[29] Xiaowen Li, Haolan Xue, Peiran Ren, and Liefeng Bo. Diffueraser: A diffusion model for video inpainting. *arXiv preprint arXiv:2501.10018*, 2025. 2

[30] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17562–17571, 2022. 2

[31] Ziye Li, Hao Luo, Xincheng Shuai, and Henghui Ding. Anyi2v: Animating any conditional image with motion control. *arXiv preprint arXiv:2507.02857*, 2025. 2

[32] Jinlai Liu, Jian Han, Bin Yan, Hui Wu, Fengda Zhu, Xing Wang, Yi Jiang, Bingyue Peng, and Zehuan Yuan. Infinitystar: Unified spacetime autoregressive modeling for visual generation. *arXiv preprint arXiv:2511.04675*, 2025. 3

[33] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Decoupled spatial-temporal transformer for video inpainting. *arXiv preprint arXiv:2104.06637*, 2021. 2

[34] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14040–14049, 2021. 2

[35] 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. 2

[36] Ziling Liu, Jinyu Yang, Mingqi Gao, and Feng Zheng. Place anything into any video. *arXiv preprint arXiv:2402.14316*, 2024. 1, 2

[37] Lev Manovich. The rise of generative media. Manovich.net, 2023. From the collection "Artificial Aesthetics". 1

[38] Chenxuan Miao, Yutong Feng, Jianshu Zeng, Zixiang Gao, Hantang Liu, Yunfeng Yan, Donglian Qi, Xi Chen, Bin Wang, and Hengshuang Zhao. Rose: Remove objects with side effects in videos. *arXiv preprint arXiv:2508.18633*, 2025. 2, 3

[39] Justine Moore. Why 2023 was ai video's breakout year, and what to expect in 2024, 2024. Accessed: 2024-05-20. 1

[40] Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David B Lindell. Sg-i2v: Self-guided trajectory control in image-to-video generation. *arXiv preprint arXiv:2411.04989*, 2024. 2

[41] 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 preprint arXiv:2410.13720*, 2024. 1

[42] Weize Quan, Jiaxi Chen, Yanli Liu, Dong-Ming Yan, and Peter Wonka. Deep learning-based image and video inpainting: A survey. *International Journal of Computer Vision*, 132(7): 2367–2400, 2024. 1, 2

[43] Shuwei Shi, Biao Gong, Xi Chen, Dandan Zheng, Shuai Tan, Zizheng Yang, Yuyuan Li, Jingwen He, Kecheng Zheng, Jingdong Chen, et al. Motionstone: Decoupled motion intensity modulation with diffusion transformer for image-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22864–22874, 2025. 2

[44] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3626–3636, 2022. 3

[45] Yuanpeng Tu, Hao Luo, Xi Chen, Sihui Ji, Xiang Bai, and Hengshuang Zhao. Videoanydoor: High-fidelity video object insertion with precise motion control, 2025. 2

[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. 1, 2, 3

[47] Zhen Wan, Chenyang Qi, Zhiheng Liu, Tao Gui, and Yue Ma. Unipaint: Unified space-time video inpainting via mixture-of-experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1861–1871, 2025. 2

[48] Chuan Wang, Haibin Huang, Xiaoguang Han, and Jue Wang. Video inpainting by jointly learning temporal structure and spatial details. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5232–5239, 2019. 2

[49] 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. 3

[50] Wenhao Wang and Yi Yang. Tip-i2v: A million-scale real text and image prompt dataset for image-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14898–14908, 2025. 2

[51] Xiaojuan Wang, Boyang Zhou, Brian Curless, Ira Kemelmacher-Shlizerman, Aleksander Holynski, and Steven M Seitz. Generative inbetweening: Adapting image-to-video models for keyframe interpolation. *arXiv preprint arXiv:2408.15239*, 2024. 2

[52] Yukun Wang, Longguang Wang, Zhiyuan Ma, Qibin Hu, Kai Xu, and Yulan Guo. Videodirector: Precise video editing via text-to-video models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2589–2598, 2025. 2

[53] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 3

[54] 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. 1, 2

[55] 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. 2

[56] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3723–3732, 2019. 2

[57] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM computing surveys*, 56(4): 1–39, 2023. 1, 2

[58] 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. 3

[59] Zuhao Yang, Jiahui Zhang, Yingchen Yu, Shijian Lu, and Song Bai. Versatile transition generation with image-to-video diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16981–16990, 2025. 2

[60] Guy Yariv, Yuval Kirstain, Amit Zohar, Shelly Sheynin, Yaniv Taigman, Yossi Adi, Sagie Benaim, and Adam Polyak. Through-the-mask: Mask-based motion trajectories for image-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18198–18208, 2025. 2

[61] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[62] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3

[63] Zhixing Zhang, Bichen Wu, Xiaoyan Wang, Yaqiao Luo, Luxin Zhang, Yinan Zhao, Peter Vajda, Dimitris Metaxas, and Licheng Yu. Avid: Any-length video inpainting with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7162–7172, 2024. 2

[64] Zhongwei Zhang, Fuchen Long, Zhaofan Qiu, Yingwei Pan, Wu Liu, Ting Yao, and Tao Mei. Motionpro: A precise motion controller for image-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27957–27967, 2025. 2

[65] 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. 3

[66] Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10477–10486, 2023. 2

[67] Bojia Zi, Shihao Zhao, Xianbiao Qi, Jianan Wang, Yukai Shi, Qianyu Chen, Bin Liang, Rong Xiao, Kam-Fai Wong, and Lei Zhang. Cococo: Improving text-guided video inpainting for better consistency, controllability and compatibility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11067–11076, 2025. 1, 2

[68] Xueyan Zou, Linjie Yang, Ding Liu, and Yong Jae Lee. Progressive temporal feature alignment network for video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16448–16457, 2021. 2