Appendix

This appendix provides comprehensive technical details and additional results for MagicMirror, encompassing dataset preparation, architectural specifications, implementation, and extensive experimental validations. We include additional qualitative results and in-depth analyses to support our main findings. **We strongly encourage readers to examine the project page `https://julianjuaner.github.io/projects/Magic-Mirror/`** for dynamic video demonstrations. The following contents are organized for efficient navigation.

Appendix Contents

A. Experiment Details

A.1. Training Data Preparation

Our training dataset is constructed through a rigorous preprocessing pipeline, as illustrated in Fig. 8. For the image pretrain data, we start downloading 5 million images from LAION-face [53], then undergo strict quality filtering based on face detection confidence scores and resolution requirements. The filtered subset of 107K images is then processed through an image captioner [33], where we exclude images containing texts. This results in a curated set of 50K high-quality face image-text pairs. To enhance identity diversity, we incorporate the synthetic SFHQ dataset [3]. To fit the model output, we standardize these images by adding black borders and pairing them with a consistent prompt template: *"A squared ID photo of ..., with pure black on two sides."* This preprocessing ensures uniformity while maintaining the dataset's diverse identity characteristics.

For FFHQ [27], we leverage a state-of-the-art identity-preserving prior PhotoMakerV2 [34] to generate synthetic images of the same identity, but with different face poses. We filter redundant identities using pairwise facial similarity metrics, with prompts sampled from our 50K video keyframe captions. We use the Pexels-400K and Mixkit datasets from [35] for construction of image-video pairs. The videos undergo a systematic preprocessing pipeline, including face detection and motion-based filtering to ensure high-quality dynamic content. We generate video descriptions using CogVLM video captioner [57]. Following our FFHQ processing strategy, we employ PhotoMakerV2 to synthesize identity-consistent images from the detected faces, followed by quality-based filtering.

A.2. Test Data Preparation

Face Images Preparation We construct a comprehensive evaluation set for identity preservation assessment across video generation models. Our dataset comprises 50 distinct identities across seven demographic categories: man, woman, elderly man, elderly woman, boy, girl, and baby. The majority of faces are sourced from PubFig dataset [30], supplemented with public domain images for younger categories. Each identity is represented by 1-4 reference images to capture variations in pose and expression.

Prompt Preparation Our test prompts are derived from VBench [25], focusing on human-centric actions. For detailed descriptions, we sample from the initial 200 GPT-4-enhanced prompts and select 77 single-person scenarios. Each prompt is standardized with consistent subject descriptors and augmented with the `img` trigger word for model compatibility. We assign four category-appropriate prompts to each identity, ensuring demographic alignment. For the "baby" category, which lacks representation in VBench, we craft four custom prompts to maintain evaluation consistency across all categories.

A.3. Comparisons

ID-Animator [20] We utilize enhanced long prompts for evaluation, although some undergo partial truncation due to CLIP's 77-token input constraint.

In our main comparisons Tab. 1, we evaluated ID-Animator at a resolution of 480×720. This choice was made to ensure that SD-based ID-Animator comparisons used matching resolutions, thereby ensuring equal content capacity—a decision justified by the inherent resolution independence of the UNet architecture. To provide a fair and comprehensive evaluation, we additionally present some results at the default 512×512 resolution here in Tab. 4. These results confirm that our comparisons remain robust and consistent across different resolution settings.

Method	Base	Resolution	txt-align↑	FM_{inter}↑	ID↑	Smooth↓
ID-Animator	SD1.5	(480, 720)	0.211	0.181	0.923	0.515
ID-Animator	SD1.5	(512, 512)	0.217	0.179	0.921	0.501
MagicMirror	CogVideoX	(480, 720)	0.240	0.610	0.922	0.484

Table 4. **ID-Animator resolution comparison.**

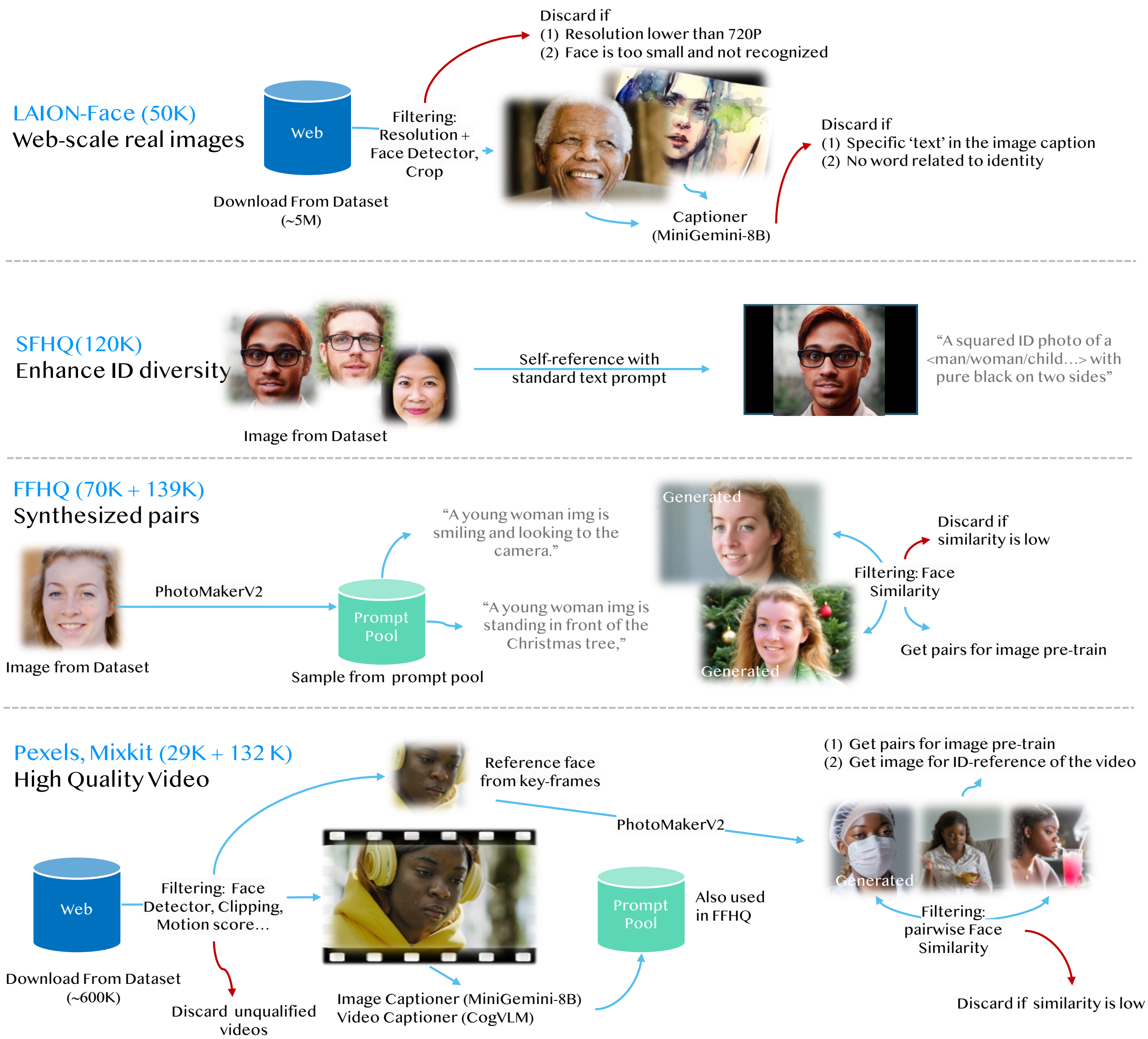


Figure 8. **Detailed training data processing pipeline.** Building upon Fig. 5, we illustrate comprehensive filtering criteria, prompt examples, and processing specifications. The data flow is indicated by blue arrows, while filtering rules leading to data exclusion are marked with red arrows.

ConsisID [69] We utilize the CogVideoX-5B[64] version. Its base inference settings are aligned with those of our model, and enhanced long prompts are employed to fully leverage its capabilities.

CogVideoX-5B-I2V [64] For this image-to-video variant, we first generate reference images using PhotoMakerV2 [34] for each prompt-identity pair. These images, combined with enhanced long prompts, serve as input for video generation.

EasyAnimate [63] We evaluate using the same PhotoMakerV2-generated reference images as in our CogVideoX-5B-I2V experiments.

DynamiCrafter [60] Due to model-specific resolution requirements, we create a dedicated set of reference images using PhotoMakerV2 that conform to the model's specifications.

In image-to-video baselines, through reference images generated by enhanced prompts, we deliberately use original short concise prompts for video generation. This choice stems from our empirical observation that image-to-video models exhibit a strong semantic bias when processing lengthy prompts. Specifically, these models tend to prior-

Figure 9. **Impact of prompt length on image-to-video generation.** We demonstrate how image-to-video models perform differently with concise versus enhanced prompts. Frames with large artifacts are marked in red. First frame images are generated from enhanced prompts.

itize text alignment over reference image fidelity, leading to degraded video quality and compromised identity preservation. This trade-off is particularly problematic for our face preservation objectives. We provide visual evidence of this phenomenon in Fig. 9.

A.4. Evaluation Metrics

Our evaluation framework combines standard video metrics with face-specific measurements. From VBench [25], we utilize Dynamic Degree for motion naturality and Overall Consistency for text-video alignment. Video quality is assessed using Inception Score from EvalCrafter [36]. For facial fidelity, we measure identity preservation using facial recognition embedding similarity [15] and temporal stability through frame-wise similarity decay.

We propose a novel facial dynamics metric to address the limitation of static face generation in existing methods. As shown in Fig. 10, we extract five facial landmarks using RetinaFace [10] and compute two motion scores: $\mathrm{FM}_{\mathrm{ref}}$ measures facial motion relative to the reference image (computed on aspect-ratio-normalized frames to eliminate positional bias), while $\mathrm{FM}_{\mathrm{inter}}$ quantifies maximized inter-frame facial motion (computed on original frames to preserve translational movements). This dual-score approach enables a comprehensive assessment of facial dynamics.

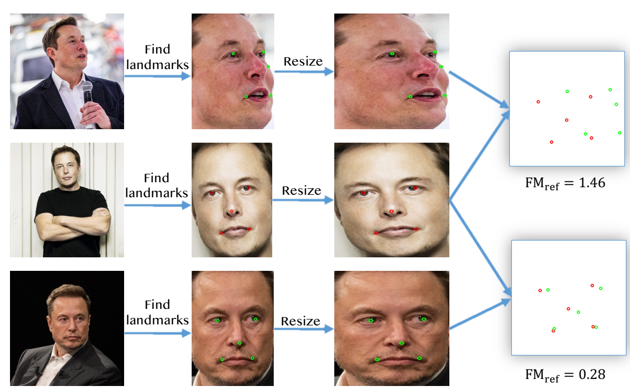


Figure 10. **Face Motion (FM) calculation.** $\mathrm{FM}_{\mathrm{inter}}$ follows a similar computation across consecutive video frames.

Success rate & failcase analysis. Success rate metrics better demonstrate reliability. Our additional experiments with

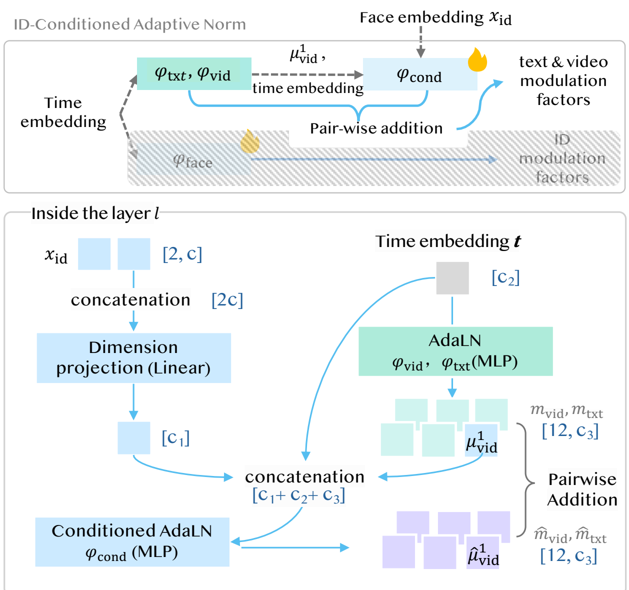


Figure 11. **Detailed implementation of Conditioned Adaptive Normalization.** We present the expanded architecture of φ_{cond} (illustrated in the unmasked region above) with comprehensive annotations of input-output tensor dimensions at each transformation.

200 videos (50 prompts × 4 seeds) compared MagicMirror with identity preservation baselines on success rates for face recognition, identity check, motion, and text alignment. MagicMirror achieves improved SR across most dimensions, though failcase analysis reveals motion quality remains the primary limitation, which is predominantly model-dependent.

Method / SR	motion quality	text align	face recongized	identity check	average
ID-Animator	11.5%	60.0%	93.5%	82.5%	61.9%
ConsisID	38.5%	73.5%	89.5%	74.5%	69.0%
MagicMirror	44.0%	75.5%	98.0%	81.0%	**74.6%**

Table 5. **Success rate comparison.**

A.5. Implementation Details

Decoupled Facial Embeddings. Our architecture employs two complementary branches: an ID embedding branch based on pre-trained PhotoMakerV2 [34] with two-token ID-embedding query q_{id}, and a facial structural embedding branch that extracts detailed features from the same ViT's penultimate layer. The latter initializes 32 token embeddings as facial query q_{face} input. We use a projection layer to align facial modalities before diffusion model input.

Conditioned Adaptive Normalization. This paragraph elaborates on the design details of the Conditioned Adaptive Normalization (CAN) module, complementing the overview provided in Sec. 3.3 and Fig. 4. For predicting facial modulation factors m_{face}, we employ a two-layer MLP architecture, following the implementation structure of the original normalization modules $\varphi_{\{vid, text\}}$. The detailed implementation of CAN is illustrated in Fig. 11. Given the facial ID embedding $x_{id} \in \mathcal{R}^{2\times c}$ containing two tokens, we first apply one global projection layer for dimensionality reduction, mapping it to dimension c_1. Subsequently, in each adapted layer, we concatenate this projected embedding with the time embedding $\mathbf{t}$ and the predicted shift factor μ^1_{vid} along the channel dimension. An MLP then processes this concatenated representation to produce the final modulation factors. To ensure stable training, all newly introduced modulation predictors are initialized with zero.

We also tried to directly use the prediction of CAN as the data distribution, this results in a bad initialization, comparing with the residual prediction, direct prediction leads to abnormal video generation quality.

Figure 12. **CAN speeds up the convergence.** Without the Conditioned Adaptive Normalization, the model cannot fit the simplest appearance features like hairstyle in the image pre-train stage.

B. Additional Discussions

B.1. Advantages of CAN

The benefits of CAN in facial condition injection are evident in its ability to enhance training convergence, particularly during the image pre-train stage. As illustrated in Fig. 12, models equipped with CAN achieve significantly improved identity information capture, enabling faster adaptation to appearance attributes. This acceleration in convergence highlights CAN's effectiveness in preserving identity consistency throughout the training process.

Furthermore, we specifically design CAN and related modules to be lightweight and avoid altering **any** pre-trained weights of the video DiT, thereby preserving the original model capacity. We evaluate GPU memory utilization, parameter count, and inference latency for generating a 49-frame 480P video. Compared to the baseline model, the additional parameters introduced by MagicMirror are primarily concentrated in the embedding extraction stage, which requires only a single forward pass. As summarized in Tab. 6, compared with ConsisID [69] and

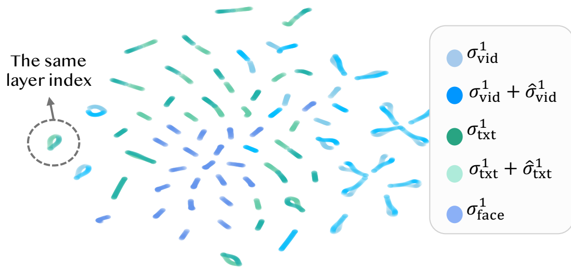


Figure 13. **Different modalities' scale distribution using t-SNE.** Each point represents the scale with a unique timestep-layer index. We also illustrate a shift variant on text and video's adaptive scale using different colors.

CogVideoX [64] baseline, MagicMirror introduces minimal computational overhead, with only a slight increase in GPU memory consumption and inference time.

Model	Video size	Memory	Params.	Latency	Batch×Iter.	Data (I+V)	GPU
ID-Animator	(16,512,512)	8.4 GiB	1.52B	11s	2×58K	0K+13K	A100*1
CogVideoX-5B	(49,480,720)	24.9 GiB	10.5B	204s	(0.1-2K)×750K	2B+35M	-
CogVideoX-I2V	(49,480,720)	25.9 GiB	10.6B	213s	-	-	-
ConsisID	(49,480,720)	41.5 GiB	11.1B	213s	80×1.8K	0K+130K	H100*40
MagicMirror	(49,480,720)	28.6 GiB	12.8B	209s	8×9K*	570K +29K	A800*8

Table 6. **Computation overhead of MagicMirror.** All computations are measured on the one A800 GPU.

B.2. Distribution Analysis and Its Impact

We begin by visualizing the predicted modulation scale factors σ using t-SNE [54] in Fig. 13. The results show that distinct modalities occupy characteristic distributions across different Transformer layers, and these distributions appear largely invariant to the specific timestep input. In particular, the face modality exhibits a unique pattern, while the conditioned residual $\hat{\sigma}$ introduces targeted shifts away from the baseline distribution. This shift empirically accelerates model convergence when incorporating ID conditions.

Beyond the t-SNE visualization, we further investigate the critical role of distribution alignment by examining how modality-aware data distributions affect generation quality. Specifically, we fine-tuned only the normalization layers $\varphi_{\text{vid}}, \varphi_{\text{txt}}$ of the CogVideoX base model on two distinct datasets—CelebV-Text [66] and our Pexels video collection [2]. As illustrated in Fig. 14, this distribution-specific fine-tuning exerts a substantial influence on the spatial fidelity of generated videos. These observations underscore the importance of aligning modality distributions during training, and they also validate the high quality of our curated video dataset.

Additionally, we conducted another experiment using our Pexels dataset. We found that by using a dataset with twice the frame rate and training only the modulation layers, we achieved an improvement in the VBench [25] dynamic motion score from 0.71 to 0.84. This result, similar to Experiments E-F in Tab. 3, further verifies the impact of the modulation module on dynamic facial motion.

Figure 14. **Modulation layers reflect data distribution.** Finetuning solely the modulation layer weights demonstrates adaptation to distinct data distributions, affecting both spatial fidelity and temporal dynamics.

B.3. Two-Stage Training Analysis

In Fig. 15, we present additional ablation results that clarify how each training phase addresses a distinct aspect of identity-preserving video generation. Specifically, the image pre-training phase prioritizes robust identity encoding, ensuring that facial features remain consistent and accurately captured. However, training exclusively on image data leads to color-shift artifacts during video inference, caused by modulation factor inconsistencies across different training stages. By combining these two stages, our final approach aligns both identity representation and color distribution, resulting in dynamic and high-fidelity ID-preserving videos without the artifacts observed in single-stage alternatives.

B.4. Limitation Analysis

As discussed in Sec. 5, our approach faces several limitations, particularly in handling multi-person scenarios and preserving fine-grained features. Fig. 16 illustrates two representative failure cases: incomplete transfer of reference character details (such as accessories) and motion artifacts caused by the base model. These limitations highlight critical areas for future research in controllable personalized

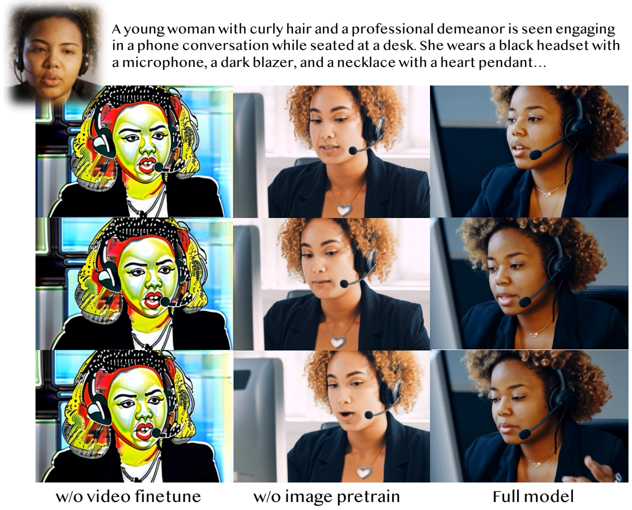
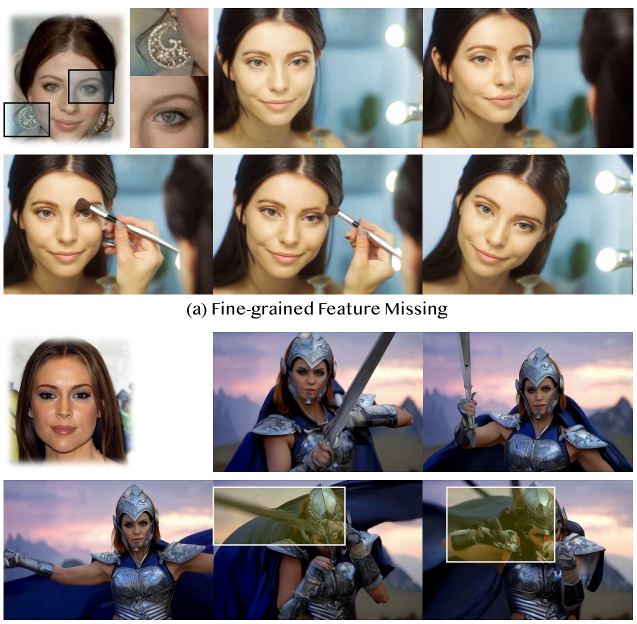


Figure 15. **Examples for ablation studies on training strategies.**



(a) Fine-grained Feature Missing

(b) Video Quality: Motion Artifacts

Figure 16. **Limitations of MagicMirror.** (a) Fine-grained feature preservation failure in facial details and accessories. (b) Motion artifacts in generated videos showing temporal inconsistencies.

video generation, particularly in maintaining temporal consistency and fine detail preservation.

C. Additional Results & Applications

C.1. Additional Applications

Fig. 17 demonstrates two extended capabilities of MagicMirror. First, beyond realistic customized video generation, our framework effectively handles stylized prompts, leveraging CogVideoX's diverse generative capabilities to produce identity-preserved outputs across various artistic styles and visual effects. Furthermore, we show that our method can generate high-quality, temporally consistent multi-shot sequences when maintaining coherent character and style descriptions. We believe these capabilities have significant implications for automated video content creation.

(a) Videos with style-specific prompts

(b) multi-shot videos with consistent character prompts

Figure 17. **Additional applications of MagicMirror.** We can generate identity-preserved videos across artistic styles and can generate multi-shot videos with consistent characters. More results are presented in the project page.

C.2. Image Generation Results

MagicMirror demonstrates strong capability in ID-preserving image generation with the image-pre-trained stage. Notably, it achieves even superior facial identity fidelity compared to video-finetuned variants, primarily due to optimization constraints in video training (e.g., limited batch sizes and dataset scope). Representative examples are presented in Fig. 18.

C.3. Video Generation Results

Additional video generation results and comparative analyses are provided in Figs. 19 and 20, highlighting our method's advantages. Fig. 19 specifically demonstrates the benefits of our one-stage approach over I2V, including superior handling of occluded initial frames, enhanced dynamic range, and improved temporal consistency during facial rotations. In Fig. 20, we provide more results with human faces on different scales.

D. Acknowledgments

Social Impact. MagicMirror is designed to facilitate creative and research-oriented video content generation while preserving individual identities. We advocate for responsible use in media production and scientific research, explicitly discouraging the creation of misleading content or violation of portrait rights. As our framework builds upon the DiT foundation model, existing diffusion-based AI-content detection methods remain applicable.

Data Usage. The training data we used is almost entirely sourced from known public datasets, including all image data and most video data. All video data was downloaded and processed through proper channels (i.e., download requests). We implement strict NSFW filtering during the training process to ensure content appropriateness.

Figures 18-20 are presented on the following pages ↓

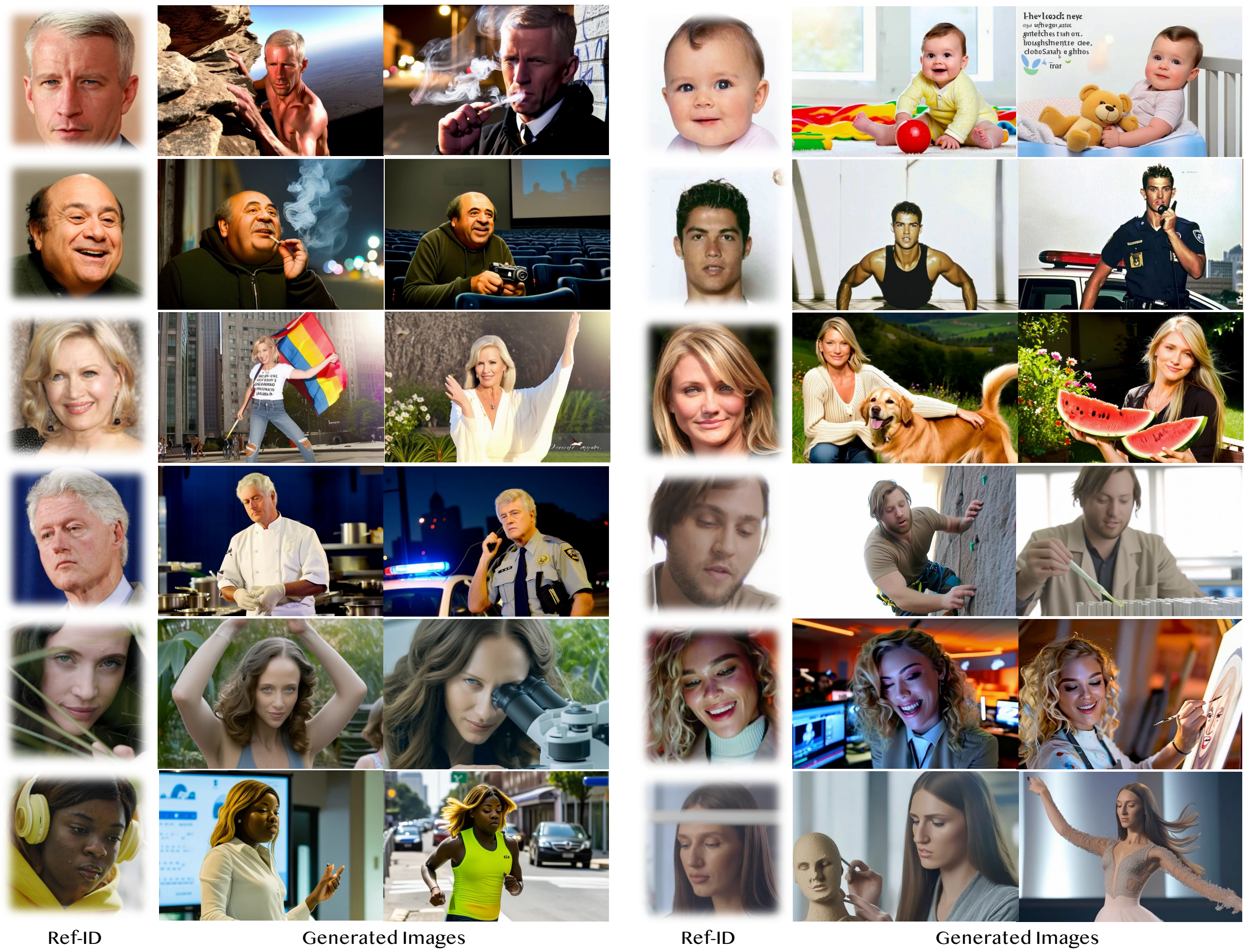


Figure 18. **Image generation using MagicMirror.** Model in the image pre-train stage captures ID embeddings of the reference ID (Ref-ID), yet over-fits on some low-level distributions such as image quality, style, and background.

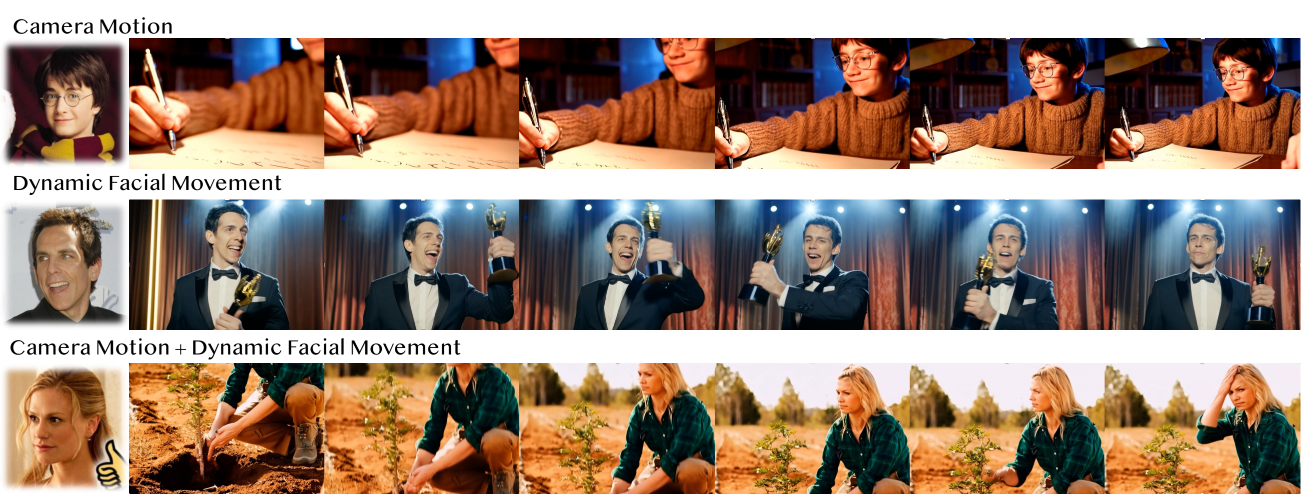


Figure 19. **Advantages over I2V generation.** MagicMirror successfully handles challenging scenarios including partially occluded initial frames and maintains identity consistency through complex facial dynamics, addressing limitations of traditional I2V approaches.

Figure 20. **Video generation results.** We demonstrate MagicMirror's capability across varying facial scales and compositions. Additional examples and comparative analyses are available in the project page.

References

[1] Mixkit: Free assets for your next video project. 5

[2] Pexels: Free 4k stock videos full hd video clips to download. 5, 13

[3] David Beniaguev. Synthetic faces high quality (sfhq) dataset, 2022. 5, 9

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. Technical report, OpenAI, 2023. 1

[5] Li Chen, Mengyi Zhao, Yiheng Liu, Mingxu Ding, Yangyang Song, Shizun Wang, Xu Wang, Hao Yang, Jing Liu, Kang Du, et al. Photoverse: Tuning-free image customization with text-to-image diffusion models. *arXiv preprint arXiv:2309.05793*, 2023. 1, 3

[6] Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated mri. *Medical image analysis*, 80: 102479, 2022. 2

[7] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *CVPR*, pages 12413–12422, 2022. 2

[8] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *ICLR*, 2023. 2

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 3, 5

[10] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, pages 5203–5212, 2020. 6, 11

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021. 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 2

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *ICLR*, 2023. 3

[15] Adam Geitgey. face_recognition, 2017. 6, 11

[16] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *ICLR*, 2022. 2

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2, 3

[18] 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. *ICLR*, 2024. 1, 3

[19] Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, and Qian He. Pulid: Pure and lightning id customization via contrastive alignment. *NeurIPS*, 2024. 1, 3, 4

[20] Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, Man Zhou, and Jie Zhang. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*, 2024. 1, 2, 3, 5, 6, 7, 8, 9

[21] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 3

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 1, 2

[23] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 35:8633–8646, 2022. 2

[24] Jiehui Huang, Xiao Dong, Wenhui Song, Hanhui Li, Jun Zhou, Yuhao Cheng, Shutao Liao, Long Chen, Yiqiang Yan, Shengcai Liao, et al. Consistentid: Portrait generation with multimodal fine-grained identity preserving. In *arXiv preprint arXiv:2404.16771*, 2024. 1, 3

[25] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, pages 21807–21818, 2024. 2, 6, 9, 11, 13

[26] Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. *NeurIPS*, 35:24240–24253, 2022. 2

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 2, 3, 5, 9

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 2

[29] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *NeurIPS*, 34:21696–21707, 2021. 2

[30] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *ICCV*, pages 365–372. IEEE, 2009. 6, 9

[31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 4

[32] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *NeurIPS*, 35:4328–4343, 2022. 2

[33] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 5, 9

[34] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *CVPR*, pages 8640–8650, 2024. 1, 2, 3, 4, 5, 6, 7, 9, 10, 12

[35] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, Tanghui Jia, Junwu Zhang, Zhenyu Tang, Yatian Pang, Bin She, Cen Yan, Zhiheng Hu, Xiaoyi Dong, Lin Chen, Zhang Pan, Xing Zhou, Shaoling Dong, Yonghong Tian, and Li Yuan. Open-sora plan: Open-source large video generation model, 2024. 3, 9

[36] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *CVPR*, pages 22139–22149, 2024. 6, 11

[37] LuChen. Video ocean - filmmaking for everyone. 2

[38] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 3

[39] Ze Ma, Daquan Zhou, Chun-Hsiao Yeh, Xue-She Wang, Xiuyu Li, Huanrui Yang, Zhen Dong, Kurt Keutzer, and Jiashi Feng. Magic-me: Identity-specific video customized diffusion. *arXiv preprint arXiv:2402.09368*, 2024. 1, 3, 5

[40] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *PMLR*, 2021. 2

[41] OpenAI. Video generation models as world simulators. 3

[42] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 2, 3, 5

[43] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024. 3

[44] Xu Peng, Junwei Zhu, Boyuan Jiang, Ying Tai, Donghao Luo, Jiangning Zhang, Wei Lin, Taisong Jin, Chengjie Wang, and Rongrong Ji. Portraitbooth: A versatile portrait model for fast identity-preserved personalization. In *CVPR*, pages 27080–27090, 2024. 1, 3

[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 *ICML*, pages 8748–8763. PMLR, 2021. 3

[46] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831. Pmlr, 2021. 2

[47] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, pages 2287–2296, 2021. 3

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 2, 4

[49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 3

[50] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 2, 3, 4

[51] 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. *NeurIPS*, 35:36479–36494, 2022. 1

[52] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *NeurIPS*, 29, 2016. 6

[53] Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*, number FZJ-2022-00923. Jülich Supercomputing Center, 2021. 5, 9

[54] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008. 2, 13

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 3

[56] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 1, 3

[57] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 6, 9

[58] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, pages 9168–9178, 2021. 3

[59] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *SIGGRAPH*, pages 1–11, 2024. 3

[60] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *ECCV*, 2024. 1, 3, 6, 7, 8, 10

[61] Jinbo Xing, Long Mai, Cusuh Ham, Jiahui Huang, Aniruddha Mahapatra, Chi-Wing Fu, Tien-Tsin Wong, and Feng Liu. Motioncanvas: Cinematic shot design with controllable image-to-video generation. *arXiv preprint arXiv:2502.04299*, 2025. 3

[62] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 2023. 2

[63] Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. Easyanimate: A high-performance long video generation method based on transformer architecture. *arXiv preprint arXiv:2405.18991*, 2024. 1, 3, 6, 7, 8, 10

[64] 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. 1, 2, 3, 4, 6, 7, 8, 10, 13

[65] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3, 4, 5

[66] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. Celebv-text: A large-scale facial text-video dataset. In *CVPR*, pages 14805–14814, 2023. 2, 13

[67] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *ICLR*, 2 (3):5, 2024. 2

[68] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 3

[69] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyuan Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. *arXiv preprint arXiv:2411.17440*, 2024. 1, 3, 6, 7, 8, 10, 12

[70] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. In *CVPR*, 2024. 3

[71] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 3

[72] Yuechen Zhang, Jinbo Xing, Bin Xia, Shaoteng Liu, Bohao Peng, Xin Tao, Pengfei Wan, Eric Lo, and Jiaya Jia. Training-free efficient video generation via dynamic token carving. *arXiv preprint arXiv:2505.16864*, 2025. 3

[73] 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, 2024. 3