CubeComposer: Spatio-Temporal Autoregressive 4K 360° Video Generation from Perspective Video

Supplementary Material

Due to the space limit of the main paper, we provide more details on our method design, dataset construction, training strategy, and more experimental analysis in this supplementary document. More visual results are presented in the demo video of the project page.

A. Dataset Details

We curated a high-quality 360° video dataset, named *4K360Vid*, with 360° videos of at least 4K resolution and detailed global and face-wise prompt for each video. These videos are from the public source used in [29]. We also adopt the high-resolution division of the ODV360 dataset in our experiments.

Specifically, to build the *4K360Vid* dataset, we implement a comprehensive filtering mechanism to exclude low-quality and anomalous videos that could adversely affect training. Initially, we exclude videos where the resolution falls below 4K. Subsequently, we employ Qwen3-VL [2, 3, 40] (235B-A22B-Instruct) to perform anomaly detection in the equirectangular view through a prompt-based analysis, checking for the following issues:

- If the video is not a proper 360° equirectangular panoramic video;
- If the video contains post-production watermarks;
- If the video exhibits obvious post-processing or compositing;
- If the video shows decoding errors or corrupted content;
- If the video lacks real scene content (e.g., entirely solid color or text-only).

Videos flagged as anomalous in any category are excluded from our dataset.

After filtering, we employ Qwen3-VL [2, 3, 40] (235B-A22B-Instruct) to generate captions for both the global scene based on the equirectangular view and each cubemap face of the 360° video. The global prompt focuses on the overall environment, atmosphere, and key elements across the full sphere, whereas the face-specific prompts target the respective directional views (front, right, back, left, up, down), detailing scene composition, objects and their spatial relationships, actions or movements, lighting conditions and atmosphere, colors and textures, notable features, and global camera movements. The generated captions (one global caption and six face-wise captions per clip) are incorporated into the training of CubeComposer. During training, we randomly apply the face-specific prompts with a probability of 0.5, enabling the model to adapt to inference scenarios using either a single global prompt for all faces or distinct prompts per face for enhanced controllability. In total, we obtain 11,832 high-quality 360° video scenes with a resolution at least of 4K and detailed global/face-wise textual captions, and each video contains 100 to 300 frames.

B. Training Details

Based on the 4K360Vid dataset we filtered and captioned, we train CubeComposer using the flow-matching training objective, with random perspective synthesis and context selection. Specifically, for each 360° video scene in our dataset, we first randomly select a time interval as our training target. Then, we randomly select 3 to 5 points on the camera sphere (roll, yaw, and pitch), and interpolate a bicubic curve along the selected points, resulting in a random but smooth camera trajectory that scans the current dynamic 360° scene. The perspective video is then extracted using the synthesized camera trajectory with a random selection of camera field-of-view from 60 to 120. This strategy ensures our model adapts to a large variety of input perspective conditions.

Then, for each training step, we plan the generation order of current perspective input according to the method described in Section 3.3 of the main paper, and randomly select a generation step along the planned order and simulate the context of that generation face according to the context strategy discussed in Section 3.4 of the main paper.

During training, we randomly apply a varying resolution from 2K to 4K to enhance the flexibility of CubeComposer (with single cubemap resolution ranging from 512×512 to 960×960). Training time window length is set to 9 and we randomly simulate the max length of 27 frames during training. We train the CubeComposer on 8 GPUs for 50 epochs on the training set of our filtered *4K360Vid* and ODV360 datasets.

C. Inference Details

Detailed Inference Process. As described in Algorithm 1, the inference begins by projecting the input perspective video into a masked cubemap representation. For camera rotation estimation, we follow the methods used in the previous method [29]. The sequence is then divided into temporal windows to facilitate autoregressive generation. For each window, we first determine a generation order by calculating the spatial coverage of the perspective input on each face, prioritizing faces with richer information to guide subsequent generation. Inside the generation loop for each

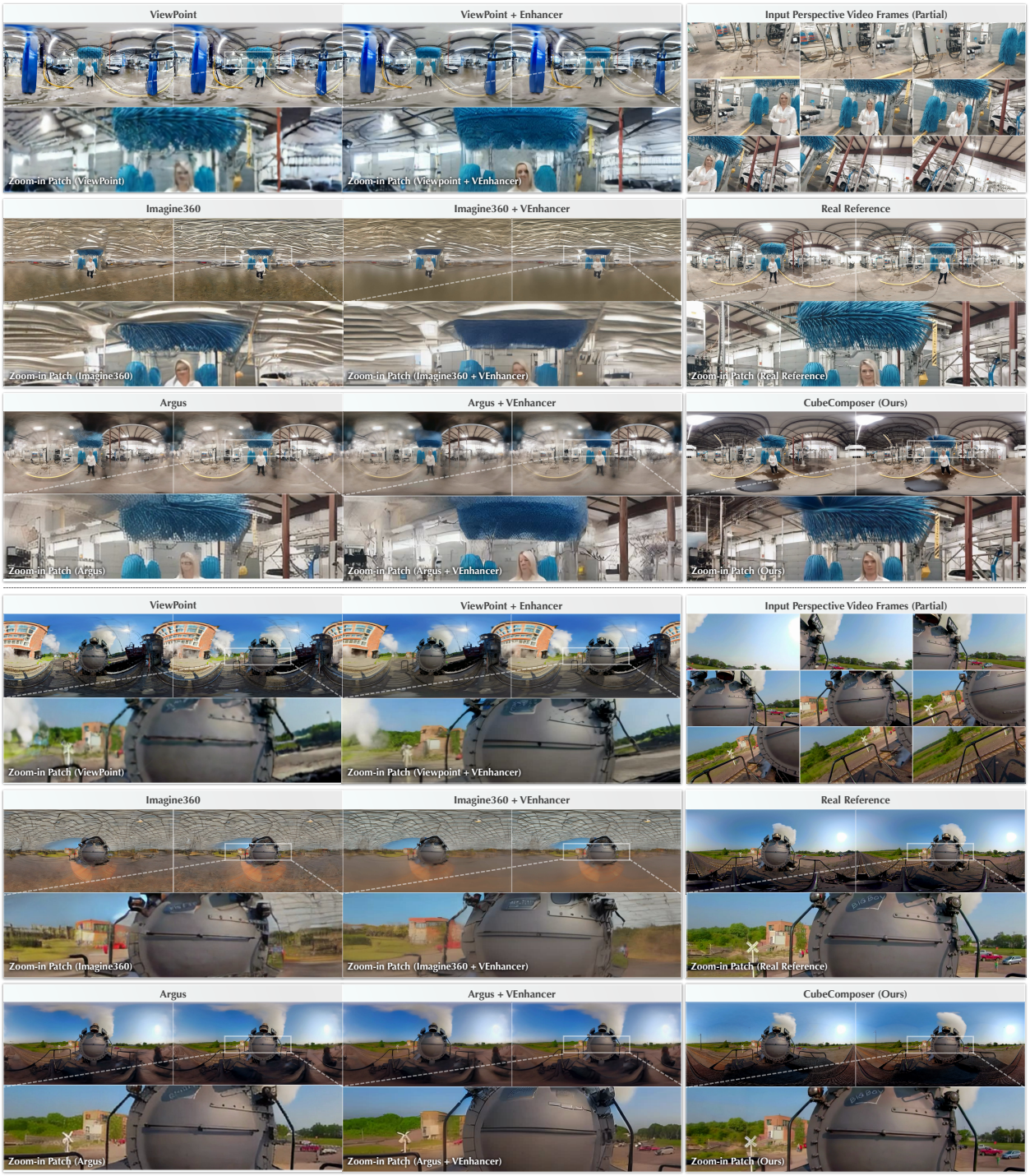

Figure 8. Visual comparison (two more scenes) between CubeComposer and previous perspective-to-360° video generation models ViewPoint [11], Imagine360 [35], and Argus [29]. Input perspective frames and real references are provided for better comparison. Zoom-in for a better view. More visual results and comparison are available in the ***supplementary video***.

face, we construct a comprehensive context comprising history tokens from previous windows, current tokens from already generated faces within the window, and future fragment tokens dynamically selected from the perspective input. This context is then fed into the diffusion model, where we apply our cube-aware padding and blending strategies to

Algorithm 1 Inference process of CubeComposer.

Require: Perspective video $\{I_t^{\text{pers}}\}_{t=1}^{N}$, field-of-view ϕ, camera rotations $\{R_t\}_{t=1}^{N}$, single window length T_{win}, history window budget H, future fragment length T_{frag}, coverage threshold r, prompts y

Ensure: 360° video $\{\tilde{X}_t^{\text{eq}}\}_{t=1}^{N}$

Compute masked cubemap condition $\{(X_{f,t}^{\text{cond}}, M_{f,t})\}$ via Eq. (1)
Divide into L windows of length T_{win} with boundaries (s_w, e_w) by Eq. (2)
Initialize context: $\mathbf{u}_w^{\text{hist}} \leftarrow \emptyset$; $\mathbf{u}_{w,f}^{\text{curr}} \leftarrow \{I_t^{\text{pers}}\}_{t=1}^{T_{\text{win}}}$;
for each window $w = 1..L$ **do**
    Compute coverage $c_{f,w}$ and order σ_w via Eq. (4)
    for $k = 1..6$ **do** ▷ faces in descending coverage
        $f \leftarrow \sigma_w(k)$
        Update $\mathbf{u}_{w,f}^{\text{fut}}$ by Eq. (7) using T_{frag}, r
        Build context $\mathbf{u}_{w,f} = [\mathbf{u}_w^{\text{hist}}; \mathbf{u}_{w,f}^{\text{curr}}; \mathbf{u}_{w,f}^{\text{fut}}]$
        Encode context, apply cube-aware padding;
        Diffusion denoise to generate $\{\tilde{\mathbf{z}}_{f,t}\}_{t=s_w}^{e_w-1}$
        Decode pixels and apply cube-aware blending;
        Update $\mathbf{u}_{w,f}^{\text{curr}}$ with the generated face;
    end for
    Append $\mathbf{u}_{w,f}^{\text{curr}}$ after $\mathbf{u}_w^{\text{hist}}$;
    Poll out earliest history if reached the budget H;
end for
Convert generated cubemap videos $\{\tilde{X}_t^{\text{cube}}\}_{t=1}^{N}$ to 360° video in equirectangular format $\{\tilde{X}_t^{\text{eq}}\}_{t=1}^{N}$;
return generated 360° video $\{\tilde{X}_t^{\text{eq}}\}_{t=1}^{N}$

ensure seamless transitions across face boundaries. Finally, the generated cubemap faces are assembled and converted back to the equirectangular format to produce the final 4K 360° video.

Spatial/Temporal Scale Settings. In our experiments we compare our CubeComposer with other four perspective-to-360° video models. Due to the different resolution/length setting of competitor video models, we apply the original settings that fits the model itself, and unifies the across-model settings as much as possible. The detailed settings are listed as follows:

- *CubeComposer*: We run our CubeComposer under the resolution of 2K (2048×1024, with cubemap resolution 512×512) and 4K (3840×1920, with single cubemap resolution 960×960). Since CubeComposer runs on multiple windows and each window's temporal length should satisfy the $4N + 1$ condition constrained by the VAE of the foundation model, we set the length of each window to 9 and run 3 windows for each inference, resulting in 27 frames for each generated video. Height-width aspect ratio of the input perspective video is $1 : 2$.

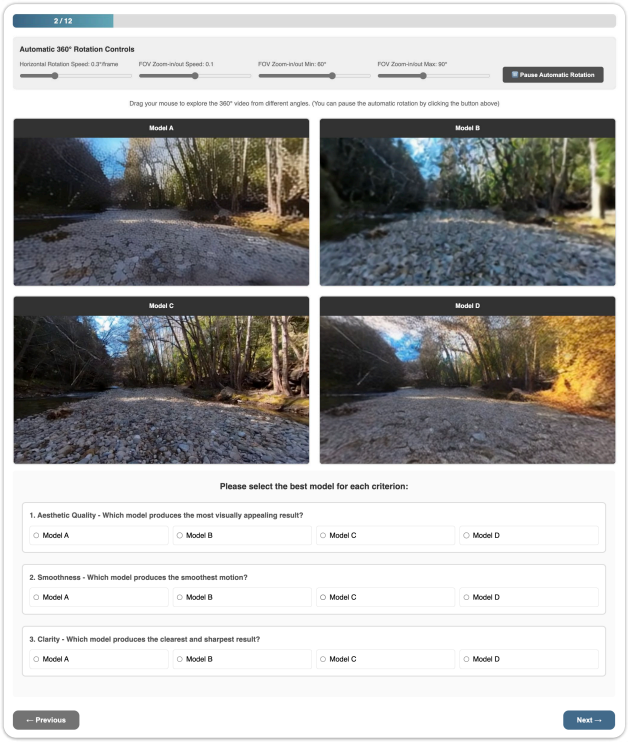

Figure 9. Screenshot of the interactive user study interface. The 360° video scene is displayed in a rendered perspective view and rotates automatically. Participants are instructed to explore the 360° video using their mouse, then select the model they perceive as superior in terms of aesthetic quality, smoothness, and clarity.

- *Argus*: We run Argus under the resolution of 1K (1024×512, equirectangular format) since this is the native resolution supported by Argus. For temporal length, we also run 27 frames for a fair comparison with CubeComposer. Height-width aspect ratio of the input perspective video is $1 : 2$.
- *Imagine360*: We run Imagine360 under its native trained resolution of 1K (1024×512, equirectangular format). For temporal length, we also run 27 frames to ensure fairness. Height-width aspect ratio of the input perspective video is $1 : 2$.
- *ViewPoint*: Since ViewPoint uses a special representation of the 360° video, we keep the default resolution settings of the pretrained model (with a squared cubeface-like input resolution of 256×256, which can be converted to equirectangular resolution of 1K, 1024×512). We run this model with 29 frames to satisfy the $4N+1$ constraint of the VAE encoder of its foundation model.

D. Experimental Analysis

More Visual Results. In addition to the comparison results provided in the main paper, we show more visual sam-

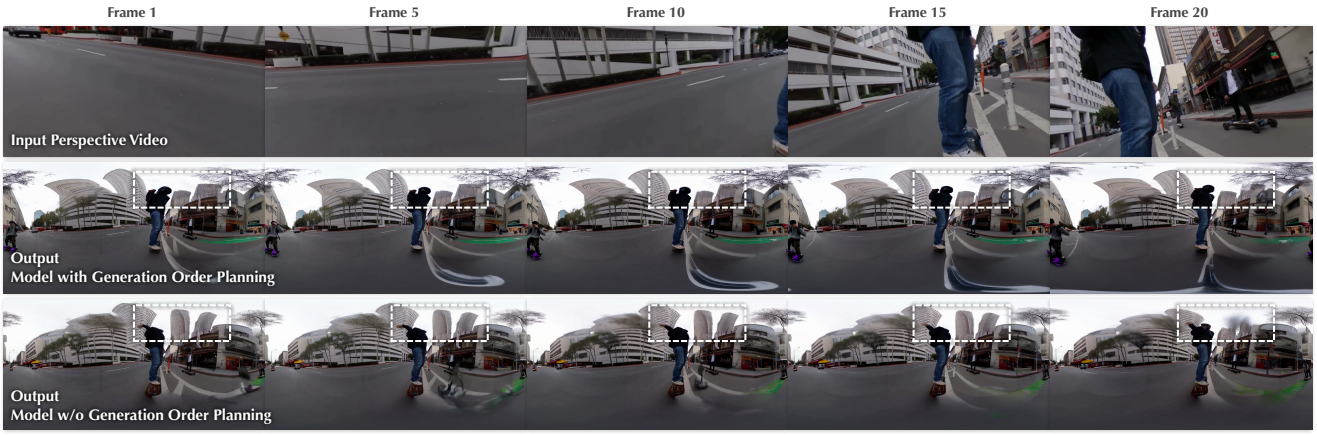

Figure 10. Ablation on autoregressive generation order planning. We compare outputs from our coverage-guided strategy (theh second row) against a random ordering baseline (the third row). The perspective input view (the first row) lacks coverage for certain regions, such as the skateboarder's head, leading to artifacts in the model without the order planning strategy.

Model	Aesthetic Qua.	Smoothness	Clarity
Argus	5.00%	5.83%	3.75%
ViewPoint	0.42%	2.50%	0.42%
Imagine360	3.75%	5.00%	4.17%
CubeComposer	**90.83%**	**86.67%**	**91.67%**

Table 4. Preference rates of aesthetic quality, smoothness, and clarity reported in user study. Our method outperforms other baselines in all three metrics, especially in clarity.

Model Type	Res.	FVD↓	FID↓	LPIPS↓	CLIP↑
Non-AR	1K	7.0106	174.6391	0.5459	0.8595
ST-AR (Ours)	1K	5.5168	158.7049	0.5093	0.8659
Non-AR (OOM)	4K	N/A	N/A	N/A	N/A
ST-AR (Ours)	4K	**3.5054**	**123.5605**	**0.4170**	**0.9061**

Table 5. Comparison between the vanilla diffusion manner and our spatio-temporal autoregressive (ST-AR) manner for 360° video generation. The non-AR model generates 360° video as a whole part, which reaches the computational bottleneck beyond 1K resolution (out-of-memory, OOM). In contrast, our ST-AR model achieve comparable performance with the vanilla baselines and can breakthrough to computational bottleneck and generates 4K 360° videos. The column *Res.* refers to the generation resolution.

ples in Figure 8. When zooming in, we can clearly observe that our method produces much better details than all competitors, while preserving more fine-grained details. A demo video for comparison is also provided on the project page.

Interactive User Study. In order to demonstrate the superior performance of our model compared to previous ones, we conducted a comprehensive user study. We construct an interactive page, which allows users to explore the 360° videos generated by each model from different angles (interact with mouse) and compare between them.

Specifically, we invited 20 participants to explore and evaluate 12 groups of generated 360° videos. The videos of different models were displayed anonymously to ensure fairness. Figure 9 shows the interactive interface of our user study. Participants were asked to select the best video based on three criteria: the aesthetic quality, the smoothness, and the clarity. The user preference rates in Table 4 demonstrate that CubeComposer is preferred over other methods in all three aspects, particularly in clarity, which benefits from our high-resolution generation capability.

Ablation on the Generation Order Planning. To validate the effectiveness of our order planning strategy, we conduct an experiment to compare the performance of our model with or without the autoregressive order planning strategy. Specifically, we benchmark our coverage-guided approach against a random ordering baseline. As illustrated in Figure 10, the model without order planning produces artifacts, particularly in regions not covered by the input perspective view. For instance, the upper body of the man on the skateboard is not captured from the perspective input, leading the model without order planning to omit his head in the generated output (noted in the white dashed box). Temporal inconsistencies of the model without order planning are also evident, such as a building vanishing in frame 20. In contrast, our order planning strategy prioritizes well-conditioned regions, promoting high-quality outputs and reducing uncertainty.

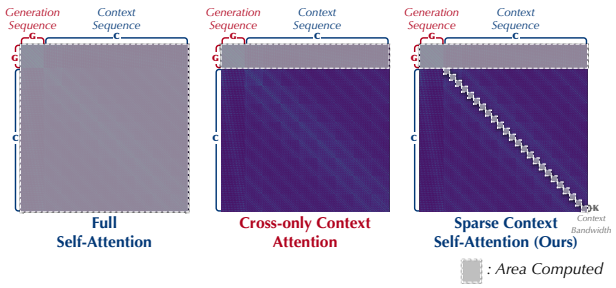

Figure 11. Illustration of the three attention designs in our ablation. The full attention computes the entire map, the cross-like variant keeps only the generation self interaction and context-to-generation cross interaction $G \times (G + C)$, and our sparse context attention further adds a diagonal $C \times K$ band that preserves constrained self-attention among context tokens. Areas that are computed during the forward pass are highlighted by white dashed boxes.

Comparison with Non-Autoregressive Baseline. To validate the effectiveness of our spatial-temporal AR manner, we conduct a comparison with a non-AR model that shares the same model architecture of our CubeComposer. This non-AR baseline shares the same foundation model (Wan 2.2 5B [38]), but is trained to generate the whole 360° video in a single diffusion denoising loop. This baseline model uses the in-context injection mechanism for perspective video input. The evaluation is done on the test set of the ODV360 dataset [6]. Table 5 shows the numerical results of the comparison. Our method is comparable in the 1K resolution but also works fine with 4K resolution, while the non-AR model fails to run (out-of-memory) with a GPU of 96 GigaBytes VRAM. This indicates that our spatial temporal autoregressive manner is necessary to achieve ultra high-resolution 360° video generation with conventional computational resources.

Analysis on Context Attention Design. For the sparse context attention mechanism described in Section 3.4, we analyze its effectiveness and efficiency against two alternatives: the full attention baseline and a cross-like variant. As illustrated in Figure 11, the full version densely couples the entire generation map, the cross-like design keeps only interaction between the generation tokens itself and between generation and the context tokens (i.e., $G \times (G + C)$), while our sparse context attention augments these crossings with an additional diagonal $C \times K$ band that enables constrained self-attention among context tokens. The quantitative comparison in Table 6 shows that our diagonal sparse context attention maintains the comparable quality of the full attention design while lowering the computational load, whereas removing the diagonal band (cross-like variant) noticeably harms fidelity and perceptual metrics. It is worth noting

Attention Type	FID↓	FVD↓	CLIP↑	LPIPS↓
Full Attention	183.1853	5.1025	0.8496	0.5286
Cross-like Attention	224.9359	5.0380	0.8151	0.6547
Sparse Context Attn. (Ours)	**157.1220**	**4.1961**	**0.8590**	**0.5142**

Table 6. Quantitative comparison of context attention designs. Our sparse context attention slightly surpasses the full attention baseline while remaining efficient, whereas the cross-like variant that lacks diagonal $C \times K$ interactions significantly degrades the performance.

that the full attention design even performs slightly worse than our sparse context attention with a controlled number of training steps. With the full-attention mechanism, the model must distinguish the informative context from all other contextual elements. This larger effective training space introduces optimization difficulty and requires more optimization steps to converge to a high-quality solution.

References

[1] Naofumi Akimoto, Yuhi Matsuo, and Yoshimitsu Aoki. Diverse plausible 360-degree image outpainting for efficient 3dcg background creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11441–11450, 2022. 3

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 6, 1

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6, 1

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 3

[6] Mingdeng Cao, Chong Mou, Fanghua Yu, Xintao Wang, Yinqiang Zheng, Jian Zhang, Chao Dong, Gen Li, Ying Shan, Radu Timofte, et al. Ntire 2023 challenge on 360deg omnidirectional image and video super-resolution: Datasets, methods and results. In *Proceedings of the IEEE/CVF con-*

ference on computer vision and pattern recognition, pages 1731–1745, 2023. 6, 8, 5

[7] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024. 3

[8] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2light: Zero-shot text-driven hdr panorama generation. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022. 3

[9] Yen-Chi Cheng, Chieh Hubert Lin, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, and Ming-Hsuan Yang. Inout: Diverse image outpainting via gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11431–11440, 2022. 3

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

[11] Zixun Fang, Kai Zhu, Zhiheng Liu, Yu Liu, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Panoramic video generation with pretrained diffusion models. *arXiv preprint arXiv:2506.23513*, 2025. 1, 2, 3, 6, 7

[12] Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. *arXiv preprint arXiv:2503.10589*, 2025. 3

[13] Jingwen He, Tianfan Xue, Dongyang Liu, Xinqi Lin, Peng Gao, Dahua Lin, Yu Qiao, Wanli Ouyang, and Ziwei Liu. Venhancer: Generative space-time enhancement for video generation. *arXiv preprint arXiv:2407.07667*, 2024. 1, 7, 8

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[17] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025. 2, 3

[18] Yukun Huang, Yanning Zhou, Jianan Wang, Kaiyi Huang, and Xihui Liu. Dreamcube: 3d panorama generation via multi-plane synchronization. *arXiv preprint arXiv:2506.17206*, 2025. 3

[19] 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. 7

[20] Nikolai Kalischek, Michael Oechsle, Fabian Manhardt, Philipp Henzler, Konrad Schindler, and Federico Tombari. Cubediff: Repurposing diffusion-based image models for panorama generation. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[21] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 3

[22] Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. *Advances in Neural Information Processing Systems*, 37:89834–89868, 2024. 2, 3

[23] Lingen Li, Guangzhi Wang, Zhaoyang Zhang, Yaowei Li, Xiaoyu Li, Qi Dou, Jinwei Gu, Tianfan Xue, and Ying Shan. Tooncomposer: Streamlining cartoon production with generative post-keyframing. In *The Fourteenth International Conference on Learning Representations*, 2026. 3

[24] Yaowei Li, Xiaoyu Li, Zhaoyang Zhang, Yuxuan Bian, Gan Liu, Xinyuan Li, Jiale Xu, Wenbo Hu, Yating Liu, Lingen Li, Jing Cai, Yuexian Zou, Yancheng He, and Ying Shan. ICcustom: Diverse image customization via in-context learning. In *The Fourteenth International Conference on Learning Representations*, 2026. 3

[25] Shanchuan Lin, Ceyuan Yang, Hao He, Jianwen Jiang, Yuxi Ren, Xin Xia, Yang Zhao, Xuefeng Xiao, and Lu Jiang. Autoregressive adversarial post-training for real-time interactive video generation. *arXiv preprint arXiv:2506.09350*, 2025. 2, 3

[26] Xin Lin, Xian Ge, Dizhe Zhang, Zhaoliang Wan, Xianshun Wang, Xiangtai Li, Wenjie Jiang, Bo Du, Dacheng Tao, Ming-Hsuan Yang, et al. One flight over the gap: A survey from perspective to panoramic vision. *arXiv preprint arXiv:2509.04444*, 2025. 2, 3

[27] Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025. 3

[28] Zhuqiang Lu, Kun Hu, Chaoyue Wang, Lei Bai, and Zhiyong Wang. Autoregressive omni-aware outpainting for open-vocabulary 360-degree image generation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, February 20-27, 2024, Vancouver, Canada*, pages 14211–14219. AAAI Press, 2024. 3

[29] Rundong Luo, Matthew Wallingford, Ali Farhadi, Noah Snavely, and Wei-Chiu Ma. Beyond the frame: Generating 360° panoramic videos from perspective videos. *arXiv preprint arXiv:2504.07940*, 2025. 1, 2, 3, 6, 7

[30] Jingwei Ma, Erika Lu, Roni Paiss, Shiran Zada, Aleksander Holynski, Tali Dekel, Brian Curless, Michael Rubinstein, and Forrester Cole. Vidpanos: Generative panoramic videos from casual panning videos. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 3

[31] Yongrui Ma, Shijie Zhao, Mingde Yao, Junlin Li, Li Zhang, Xiaohong Liu, Qi Dou, Jinwei Gu, and Tianfan Xue. Real-time video frame interpolation using one-step diffusion sampling. In *The Fourteenth International Conference on Learning Representations*, 2026. 3

[32] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2303.12345*, 2023. 3

[33] 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. 7

[34] 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. 3

[35] Jing Tan, Shuai Yang, Tong Wu, Jingwen He, Yuwei Guo, Ziwei Liu, and Dahua Lin. Imagine360: Immersive 360 video generation from perspective anchor. *arXiv preprint arXiv:2412.03552*, 2024. 1, 2, 3, 6, 7

[36] Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025. 3

[37] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 7

[38] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 3, 4, 6, 5

[39] Guangcong Wang, Yinuo Yang, Chen Change Loy, and Ziwei Liu. Stylelight: Hdr panorama generation for lighting estimation and editing. In *European conference on computer vision*, pages 477–492. Springer, 2022. 3

[40] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6, 1

[41] Qian Wang, Weiqi Li, Chong Mou, Xinhua Cheng, and Jian Zhang. 360dvd: Controllable panorama video generation with 360-degree video diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6913–6923, 2024. 3

[42] Songsong Wu, Hao Tang, Xiao-Yuan Jing, Haifeng Zhao, Jianjun Qian, Nicu Sebe, and Yan Yan. Cross-view panorama image synthesis. *IEEE Transactions on Multimedia*, 25: 3546–3559, 2022. 3

[43] Yifei Xia, Shuchen Weng, Siqi Yang, Jingqi Liu, Chengxuan Zhu, Minggui Teng, Zijian Jia, Han Jiang, and Boxin Shi. Panowan: Lifting diffusion video generation models to 360° with latitude/longitude-aware mechanisms. *arXiv preprint arXiv:2505.22016*, 2025. 3

[44] Kevin Xie, Amirmojtaba Sabour, Jiahui Huang, Despoina Paschalidou, Greg Klar, Umar Iqbal, Sanja Fidler, and Xiaohui Zeng. Videopanda: Video panoramic diffusion with multi-view attention. *arXiv preprint arXiv:2504.11389*, 2025. 2, 3

[45] Shuzhou Yang, Xiaoyu Li, Xiaodong Cun, Guangzhi Wang, Lingen Li, Ying Shan, and Jian Zhang. Gencompositor: Generative video compositing with diffusion transformer. In *The Fourteenth International Conference on Learning Representations*, 2026. 3

[46] 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. 2, 3

[47] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22963–22974, 2025. 2, 3

[48] Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *arXiv preprint arXiv:2504.12626*, 2025. 3

[49] Muyang Zhang, Yuzhi Chen, Rongtao Xu, Changwei Wang, JinMing Yang, Weiliang Meng, Jianwei Guo, Huihuang Zhao, and Xiaopeng Zhang. Panodit: Panoramic videos generation with diffusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10040–10048, 2025. 3

[50] 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. 7

[51] Junhao Zhuang, Shi Guo, Xin Cai, Xiaohui Li, Yihao Liu, Chun Yuan, and Tianfan Xue. Flashvsr: Towards real-time diffusion-based streaming video super-resolution. *arXiv preprint arXiv:2510.12747*, 2025. 3