
PanoWan: Lifting Diffusion Video Generation Models to 360° with Latitude/Longitude-aware Mechanisms

Supplementary Materials

Yifei Xia^{1,2,3*} Shuchen Weng^{4*} Siqi Yang⁵ Jingqi Liu^{1,2} Chengxuan Zhu⁶
Minggui Teng^{1,2} Zijian Jia⁷ Han Jiang³ Boxin Shi^{1,2†}

¹State Key Lab of Multimedia Info. Processing, School of Computer Science, Peking University

²Nat'l Eng. Research Ctr. of Visual Tech., School of Computer Science, Peking University

³OpenBayes Information Technology Co., Ltd. ⁴Beijing Academy of Artificial Intelligence

⁵Institute for Artificial Intelligence, Peking University

⁶Nat'l Key Lab of General AI, School of Intelligence Science and Technology, Peking University

⁷School of Artificial Intelligence, Beijing University of Posts and Telecommunications

{yfxia, shuchenweng, yousiki, peterzhu, minggui_teng, shiboxin}@pku.edu.cn

liujingqi@stu.pku.edu.cn jiazijian@bupt.edu.cn hahn@openbayes.com

7 Appendix

7.1 Proof of Noise Distribution Preservation

To align the initial noise with the spherical frequency distribution and avoid latitudinal distortion in polar regions of ERP, Sec. 4.2 proposes the *latitude-aware sampling* strategy. For clarity, we recall the noise at each coordinate after remapping the horizontal sample coordinate x based on latitude y , as stated in Eq. (5) of the main paper:

$$P'(x, y) = \text{Interp}_P \left(R + (x - R) \cos\left(\frac{2y + 1 - R}{2R}\pi\right), y \right), \quad (12)$$

To best exploit the diffusion prior, it is desired to have $\mathbb{E}[P'(x, y)] = 0$ and $\mathbb{E}[\text{Var } P'(x, y)] = 1$. First, we provide $\mathbb{E}[P'(x, y)] = 0$ as follows:

$$\mathbb{E}[P'(x, y)] = \mathbb{E}_{x, y} \left[\text{sign}(\text{BI}(P, x, y)) \sqrt{\text{BI}(P^2, x, y)} \right] \quad (13)$$

$$= \mathbb{E}_{P_{ij} \sim \mathcal{N}(0, 1)} \left[\text{sign} \left(\sum w_{ij} P_{ij} \right) \sqrt{\sum w_{ij} P_{ij}^2} \right] \quad (14)$$

$$\stackrel{\tilde{P}_{ij} := -P_{ij}}{=} \mathbb{E}_{(\tilde{P}_{ij}) \sim \mathcal{N}(0, 1)} \left[\text{sign} \left(\sum w_{ij} (-\tilde{P}_{ij}) \right) \sqrt{\sum w_{ij} (-\tilde{P}_{ij})^2} \right] \quad (15)$$

$$= - \mathbb{E}_{\tilde{P}_{ij} \sim \mathcal{N}(0, 1)} \left[\text{sign} \left(\sum w_{ij} (\tilde{P}_{ij}) \right) \sqrt{\sum w_{ij} \tilde{P}_{ij}^2} \right] = -\mathbb{E}[P'(x, y)]. \quad \square \quad (16)$$

*Equal contribution.

†Corresponding author.

After that, we prove $\mathbb{E}[\text{Var } P'(x, y)] = 1$ as follows:

$$\mathbb{E}[\text{Var } P'(x, y)] = \mathbb{E}[P'(x, y)^2] - (\mathbb{E}[P'(x, y)])^2 = \mathbb{E}[P'(x, y)^2] \quad (17)$$

$$= \mathbb{E}_{P_{ij} \sim \mathcal{N}(0,1)} \left[\sum_{i,j \in \{0,1\}} w_{ij} P_{ij}^2 \right] = \sum_{i,j \in \{0,1\}} w_{ij} \mathbb{E}_{P_{ij} \sim \mathcal{N}(0,1)} [P_{ij}^2] \quad (18)$$

$$= \sum_{i,j \in \{0,1\}} w_{ij} = 1. \quad \square \quad (19)$$

Note that the first equation in Eq. (19) is possible only because P_{ij} is independent and identically distributed.

7.2 Additional Experiment Results

In this section, we provide additional comparison results between PanoWan and existing text-based panoramic video generation methods [2, 4], where PanoDiT [8] is omitted as its code is unavailable. We also present additional examples showcasing applications such as long video generation, super-resolution, semantic inpainting, and video outpainting. Finally, we provide a detailed discussion of failure cases.

Additional comparison results. As shown in Fig. 5, DynamicScaler [2] suffers from a limited local denoising window, resulting in globally inconsistent content with repeated semantic elements appearing in different regions. 360DVD [4] often produces observable artifacts in high-latitude regions and exhibits relatively limited scene consistency. In contrast, our PanoWan achieves the most coherent and visually consistent results across diverse scenarios.

Additional application results. We provide additional examples across four application scenarios. As shown in Fig. 6, (a) PanoWan enables long video generation while maintaining consistent semantics throughout extended temporal durations. (b) For the super-resolution, directly applying Wan 2.1 [5] leads to severe artifacts in high-latitude regions, whereas PanoWan produces consistent and artifact-free results across all latitudes. (c) and (d) further demonstrate its effectiveness in semantic editing and video outpainting, respectively. These results highlight the strong potential of PanoWan as a versatile model for high-quality panoramic video generation and editing.

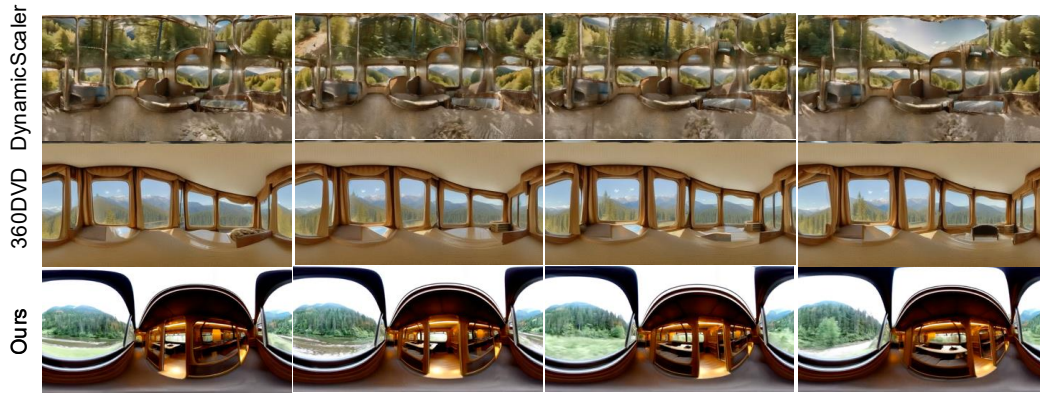
Failure cases. PanoWan can occasionally exhibit failures in certain scenarios. As illustrated in the top row of Fig. 7, the generated pet dog exhibits inconsistent features. In the bottom row, the book being read by the woman appears with two spines and three pages, deviating from real-world structure and common sense. We believe these failure cases are not primarily related to panoramic properties but are largely inherited from the pre-trained text-to-video model [5]. As the backbone model improves, these failure cases will be improved.

7.3 Technical Comparison with Prior Work

To clarify the novelty and design motivation of our proposed mechanisms, we provide a concise technical comparison between PanoWan and prior panoramic generation approaches, highlighting how our latitude-/longitude-aware modules extend existing ideas from the image domain to dynamic video generation while preserving pretrained generative priors.

Rotated semantic denoising. While existing panoramic text-to-image method PanoDiffusion [6] adopts a similar two-end alignment strategy, it relies on a fixed rotation angle, which can be inadequate for dynamic video generation because even minor spatial inconsistencies can accumulate temporally and degrade overall consistency. In contrast, our proposed rotated semantic denoising mechanism employs a step-dependent randomized circular shift that gradually disperses seam-related artifacts over time.

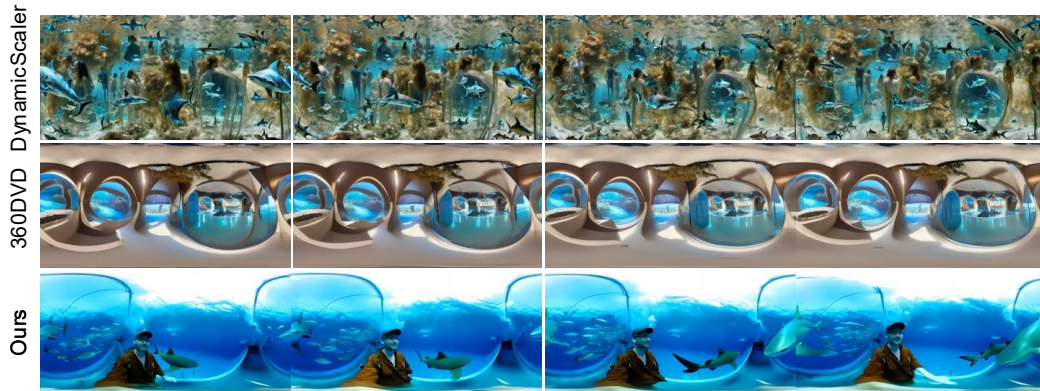
Padded pixel-wise decoding. Although the use of wrap-around or pixel-rotation context during VAE decoding has been previously explored in PanFusion [7] and StitchDiffusion [3] for panoramic images, our focus is the integration of this decoding-stage operation into a rectified-flow video generation backbone, coupling it with rotated semantic denoising to jointly handle latent- and pixel-level discontinuity while preserving the pretrained priors.



Wide-angle panoramic view from inside a luxurious vintage train wagon as it gently traverses through spectacular landscapes. Mountains, forests, and meandering rivers captured clearly through panoramic windows, accentuated by warm interior lighting and cozy compartments, providing a relaxing and nostalgic atmosphere.



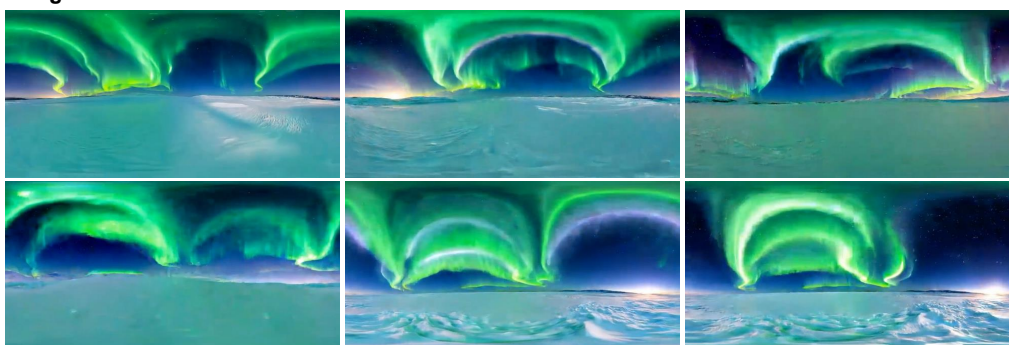
Stunning panoramic underwater shot of a vibrant coral reef ecosystem brimming with marine life. Colorful fish dart effortlessly among intricate coral formations, soft rays of sunlight filter through the crystal-clear waters, creating mesmerizing patterns on the ocean floor. Wide-angle capturing vivid hues and abundant biodiversity.



Immersive panoramic view inside an elongated aquarium tunnel, visitors walking beneath a transparent underwater canopy surrounded by vibrant fish, graceful manta rays, and large sharks moving serenely through clear azure waters, producing a compelling sense of underwater wonder and tranquility.

Figure 5: Additional comparison results with existing text-based panoramic video generation methods.

Long Video



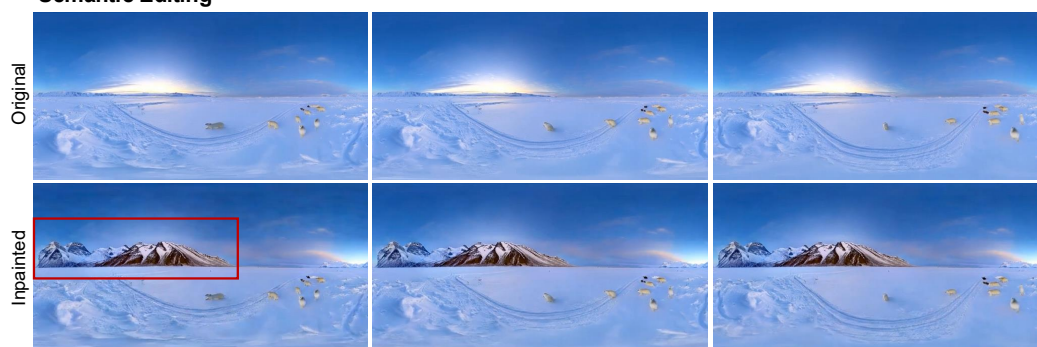
Majestic panoramic shot capturing vivid green and violet northern lights gracefully illuminating a quiet, snow-covered tundra beneath a star-studded night sky. Gentle, fluid ribbons of color dance overhead in mesmerizing motion, creating an awe-inspiring spectacle. Ultra-wide landscape shot emphasizing grandeur and serenity.

Super-resolution



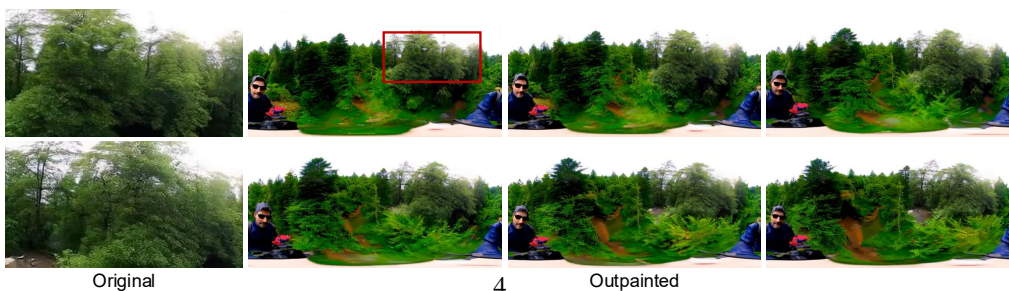
360-degree panoramic interior view inside a charming artisan bakery bustling with activity, bakers carefully preparing handcrafted breads, pastries, and desserts. Shelves stocked with warm baked goods, aromatic scents filling the air, creating feelings of warmth, comfort, and culinary delight.

Semantic Editing



Add a snowy mountain.

Video outpainting



Original

4

Outpainted

Figure 6: Additional application results, showcasing the zero-shot capabilities for downstream tasks.

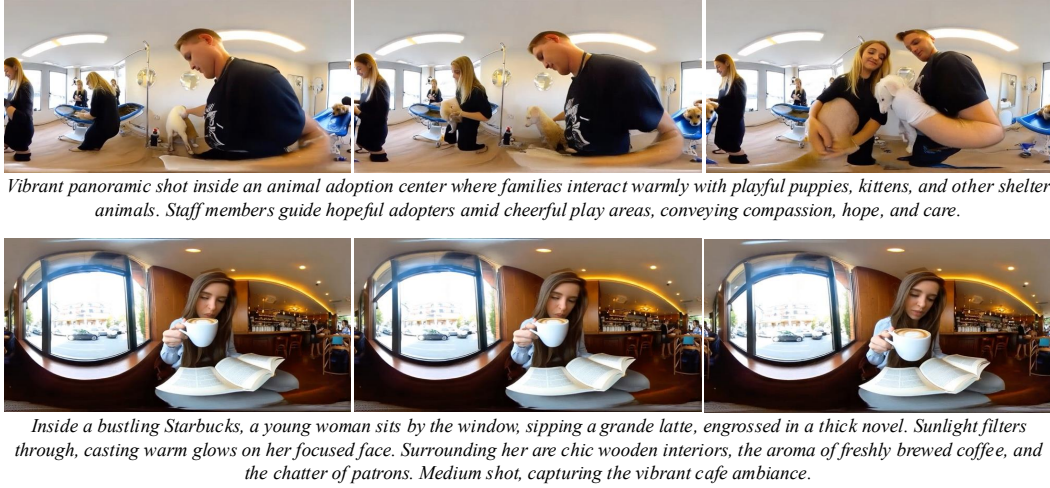


Figure 7: Visualization of failure cases.

7.4 Dataset Details

For transparency and reproducibility, we provide the detailed prompt used for Qwen-2.5-VL [1] during the vision-language annotation stage. Each 10-second panoramic clip is analyzed with the following instruction, where the model responses are automatically validated for JSON integrity and ERP format consistency:

Please analyze this video and provide your response in JSON format with the following fields:

1. 'caption': A detailed description of what's happening in the video. Do not show your analysis, just describe what you see. Do not start with "The video shows", describe the video itself as a whole. Include the panoramic statement like "Panoramic view of ..." or "360 degree view of ..." if it is a panoramic video. Here are some examples:
 - 'Panoramic shot of colorful hot air balloons gracefully ascend, floating over lush green fields, their vibrant hues contrasting against a vast, cloud-dappled blue sky. Gentle breezes propel them in a serene dance, casting dynamic shadows on the verdant landscape below. Wide shot from ground level, capturing the expansive scene.'
 - 'Panoramic shot of an active volcano spewing smoky plumes against a fiery sunset sky, majestic mountains shrouded in misty clouds in the foreground, creating a breathtaking contrast. Camera pans slowly, capturing the vastness and awe-inspiring beauty of nature.'
 - 'Aerial perspective of vibrant fireworks blossoming in the ink-black sky, casting shimmering lights over a sprawling urban landscape below. Mesmerizing pyrotechnics burst in various colors and patterns against the starless expanse, illuminating cityscapes with transient brilliance. Wide shot from a plane window, capturing the nocturnal city alive under the grand firework spectacle.'
2. 'is_panorama': A boolean (true or false) indicating whether this is a 360-degree ERP projected video.
3. 'poi_category': A list of strings indicating the points of interest in the video. If there are no points of interest, set this to an empty list. If there are multiple points of interest, describe each of them in a string, sorted by their importance. You should use the available POI categories. In case none of the provided POI categories can describe the video, you may return a succinct word in the similar pattern as given categories. For example:
 - ['Coffee-Shop']
 - ['Mountains', 'Lakes']

Your response should be valid JSON string, like this:

```
““json
{
  "caption": "Panoramic shot of colorful hot air balloons gracefully ascend,
             floating over lush green fields, their vibrant hues contrasting
             against a vast, cloud-dappled blue sky. Gentle breezes propel them in
             a serene dance, casting dynamic shadows on the verdant landscape below.
             Wide shot from ground level, capturing the expansive scene.",
  "is_panorama": true,
  "poi_category": ["Mountains"]
}
””
```

Note that the video may be compressed with limited fps to reduce uploaded file size. Your response in ‘caption’ should not include any description of the video quality or compression. Just focus on the content of the video.

Here are the available POI categories: "Restaurant, Coffee-Shop, Bars-and-Pubs, Residential-area, Hotels-Motels, Vacation-Rentals, Hospitals-Clinics, Pharmacies, Dentists, School-Universities, Library, Supermarkets, Shopping-Malls, Clothing-Stores, Shoe-Stores, Bookstores, Flowerstore, Furniture-Stores, Electorical-Store, Pet-Store, Toy-Shop, Airports, Train-Stations, Bus-Stops, Gas-Station, Car-Rental-Agencies, Theaters, Concert-Halls, Sports-Stadiums, Parks-and-Recreation-Areas, Museums, Art-Galleries, Zoos-Aquariums, Botanical-Gardens, Landmarks, Cultural-Centers, Post-Offices, Police-Stations, Courthouses, City-Halls, Banks-ATMs, Events-Conferences-halls, Beaches, Hiking-Trails, Campgrounds, Lakes, Mountains, Forest-Mountains, Farms, Street-View, Square, Business-Centers, Tech-Companies, Co-working-Spaces, Gyms-and-Fitness-Centers, Sports-Clubs, Swimming-Pools, Tennis-Courts, Auto-Repair-Shops, Car-Washes, Parking-Lots, Churches, Mosques, Temples, Graveyards."

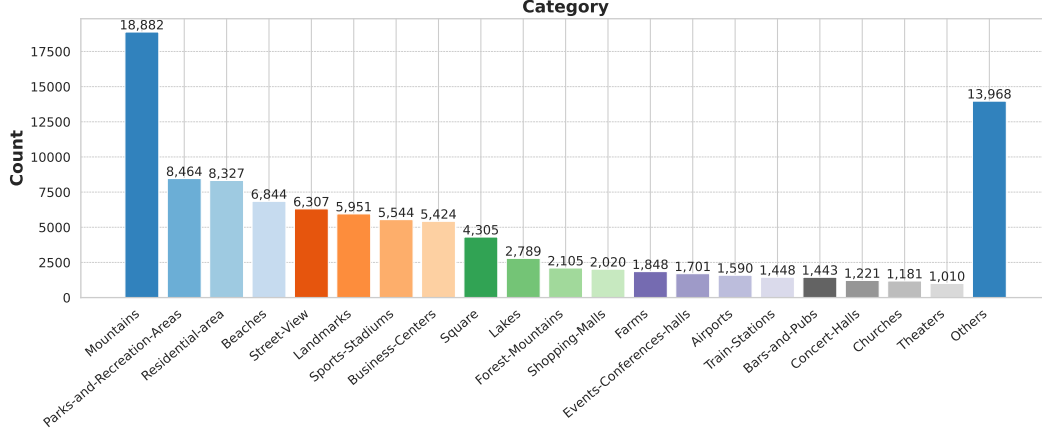


Figure 8: Category distribution of PANOVID dataset before balancing the semantics.

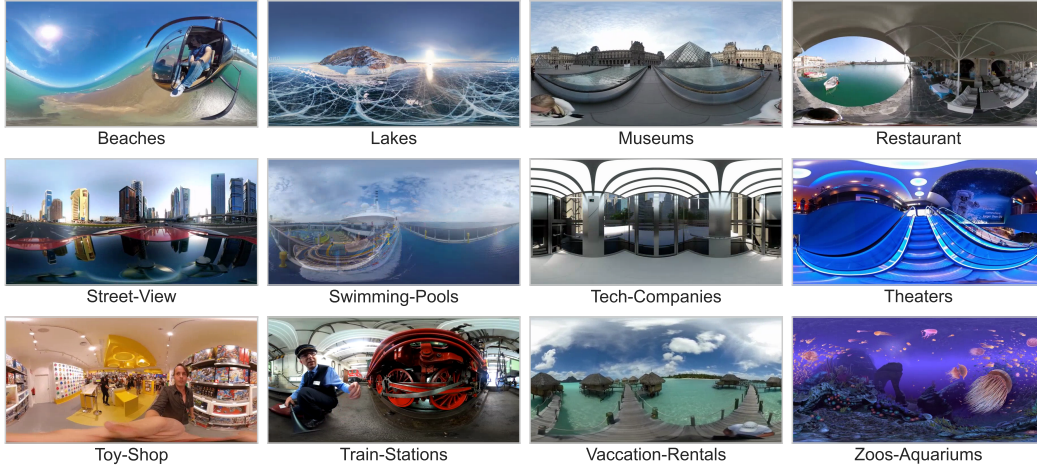


Figure 9: Representative samples from the PANOVID dataset.

Category distribution analysis. We present the POI category distribution before the balancing step described in the main paper Fig. 8. A strong long-tail behavior is observed, with natural-scene categories dominating the dataset. This motivates our semantic balancing procedure that caps each category to 200 top-quality clips.

Representative examples. We provide additional qualitative examples from PANOVID to illustrate scene diversity and caption richness, as shown in Fig. 9. Each sample includes a panoramic frame and its generated caption.

References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [2] Liu Jinxiu, Lin Shaoheng, Li Yinxiao, and Yang Ming-Hsuan. DynamicScaler: Seamless and scalable video generation for panoramic scenes. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [3] Hai Wang, Xiaoyu Xiang, Yuchen Fan, and Jing-Hao Xue. Customizing 360-degree panoramas through text-to-image diffusion models. In *Proc. of IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

- [4] Qian Wang, Weiqi Li, Chong Mou, Xinhua Cheng, and Jian Zhang. 360DVD: Controllable panorama video generation with 360-degree video diffusion model. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [5] WanTeam, 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.
- [6] Tianhao Wu, Chuanxia Zheng, and Tat-Jen Cham. PanoDiffusion: 360-degree panorama out-painting via diffusion. *arXiv preprint arXiv:2307.03177*, 2023.
- [7] Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360° panorama image generation. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [8] Muiyang 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 *Proc. of the AAAI Conference on Artificial Intelligence*, 2025.