

# VIDEO4SPATIAL: Towards Visuospatial Intelligence with Context-Guided Video Generation

Zeqi Xiao<sup>1,2\*</sup> Yiwei Zhao<sup>1✉†</sup> Lingxiao Li<sup>1</sup> Yushi Lan<sup>3</sup>,  
 Ning Yu<sup>4</sup> Rahul Garg<sup>1</sup> Roshni Cooper<sup>1</sup> Mohammad H. Taghavi<sup>1</sup> Xingang Pan<sup>2✉</sup>

<sup>1</sup>Netflix <sup>2</sup>Nanyang Technological University,

<sup>3</sup>University of Oxford <sup>4</sup>Netflix Eyeline Studios

{zeqi001, xingang.pan}@ntu.edu.sg

{yiweiz, lingxiaol, rahulgarg, mtaghavi}@netflix.com

yushi.lan@eng.ox.ac.uk ning.yu@scanlinevfx.com

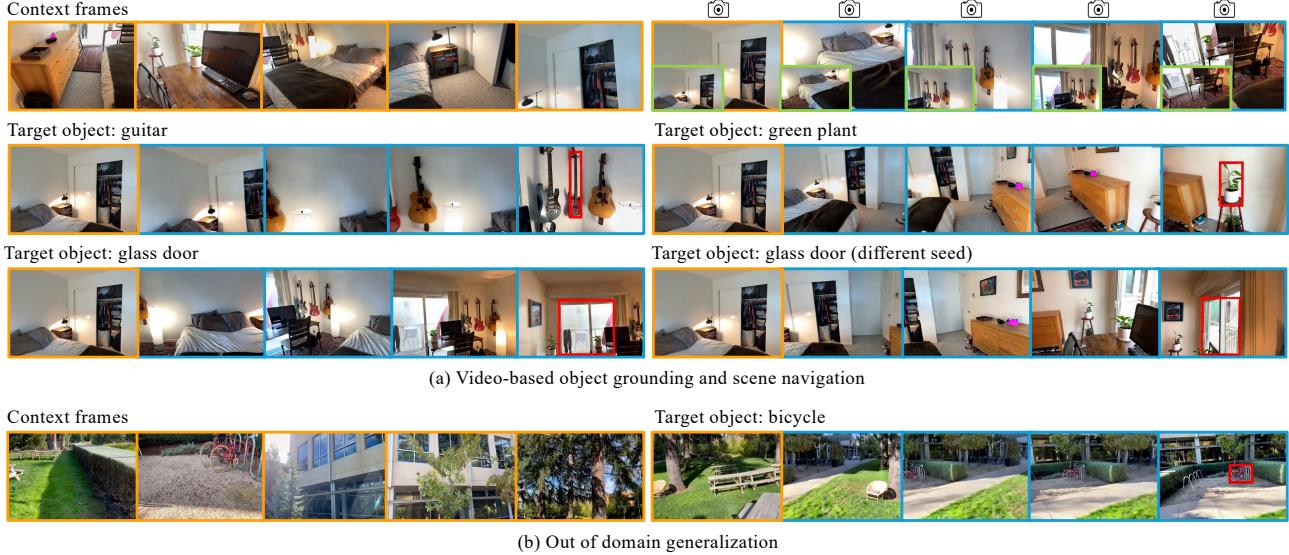


Figure 1. Our method generates videos that fulfill instructional spatial tasks while remaining geometrically consistent with the provided video context. (a) We demonstrate two tasks: **video-based object grounding**, where the model follows text instructions (the second and the third row) to navigate and locate a target object (e.g., “a green plant” or “a guitar”). Navigation paths and final locations are implicitly planned with diversity by model during generation to achieve the same goal (third row); and **scene navigation** (top-right), where the model generates a video that adheres to a specified camera trajectory (denoted by camera logos) and aligns with ground truth videos (green boxes). (b) Although trained on indoor datasets, our model generalizes well to out-of-domain scenarios, such as an outdoor park. Orange boxes denote context frames. Blue boxes denote generated frames. Please note that red bounding boxes are generated by the video model itself. Frames are subsampled for visualization. Project page at <https://xiaoqu.github.io/video4spatial>.

## Abstract

We investigate whether video generative models can exhibit visuospatial intelligence, a capability central to hu-

man cognition, using only visual data. To this end, we present VIDEO4SPATIAL, a framework showing that video diffusion models conditioned solely on video-based scene context can perform complex spatial tasks. We validate on two tasks: scene navigation - following camera-pose instructions while remaining consistent with 3D geometry of the scene, and object grounding - which requires se-

\* Work done during the internship at Netflix.

✉ Corresponding Author. † Project Lead.

*mantic localization, instruction following, and planning.* Both tasks use video-only inputs, without auxiliary modalities such as depth or poses. With simple yet effective design choices in the framework and data curation, VIDEO4SPATIAL demonstrates strong spatial understanding from video context: it plans navigation and grounds target objects end-to-end, follows camera-pose instructions while maintaining spatial consistency, and generalizes to long contexts and out-of-domain environments. Taken together, these results advance video generative models toward general visuospatial reasoning.

## 1. Introduction

Humans excel at remembering, understanding, and acting within spatial environments—a hallmark of high-level spatial reasoning capability. A long line of research has sought to endow neural networks with similar capabilities. Recently, rapidly advancing video generative models [22, 48, 63] have begun to address this challenge, exhibiting general-purpose perceptual and reasoning abilities [74], echoing the trajectory of Large Language Models [3, 46]. Because video generation is inherently sequential and temporally coherent, these models are particularly well-suited to visuospatial-related tasks: by rolling out future frames that preserve scene geometry and appearance, they can simulate plausible futures. For example, in maze navigation, a model can “imagine” progress by generating frames that maintain the maze’s topology while ensuring smooth transitions [61, 74].

Humans primarily perceive spatial information visually. We aim to endow models with the same capability by using video as a simple, general, and scalable modality. However, learning spatial understanding solely from RGB video is challenging, and many recent methods still rely on auxiliary signals (*e.g.*, depth maps, camera poses, point clouds) for supervision or augmentation [73, 75, 80, 92, 103]. In this paper, we present VIDEO4SPATIAL and show that *video generative models that rely solely on video context of a spatial environment can exhibit strong spatial intelligence*. Our framework is intentionally simple: a standard video diffusion architecture trained only with the diffusion objective. The inputs are a video context (several frames from the same environment) and instructions; the output is a video that completes the instructed spatial task while maintaining scene geometry and temporal coherence.

The core spatial abilities our framework aims to achieve include (i) implicitly inferring 3D structure [34, 102], (ii) controlling viewpoint with geometric consistency (scale, occlusion, layout) [27, 72], (iii) understanding semantic layout [38], (iv) planning toward goals [2], and (v) maintaining long-horizon temporal coherence. We introduce two tasks to probe these abilities: *video-based scene navigation*

*and object grounding*. In scene navigation, the model follows camera-pose instructions and produces a video whose trajectory is consistent with both the instructions and the scene geometry. In object grounding, a text instruction asks the model to reach and visibly localize a target object, thereby testing semantic layout understanding and plausible planning (Fig. 1). Together, the tasks evaluate whether VIDEO4SPATIAL can infer and act on scene structure using video alone.

Beyond the base architecture, our central design choice is how to model conditions on *video context*. In line with DiT [50], we process context and target frames through the same transformer stack, while setting the diffusion timestep of context frames to  $t = 0$ . Inspired by History Guidance [55], we extend classifier-free guidance [28] to video context, which we find significantly improves contextual coherence. To reduce redundancy in continuous footage in context, we subsample non-contiguous frames during training and inference and apply non-contiguous RoPE [57] over the corresponding subsampled indices. This index sparsity also enables extrapolation to much longer contexts at inference. For visuospatial tasks, we find that explicit reasoning patterns [20] help task completion; accordingly, we train the model to predict videos with visual bounding boxes to strengthen object grounding.

Since no large public datasets are purpose-built for these spatial task settings, we repurpose indoor scanning datasets—ScanNet++ [87] and ARKitScenes [7], which provide long trajectories through diverse rooms and enable sampling of context–target video clip pairs. We use an off-the-shelf vision-language model (VLM) [59] to (i) identify clips that end with a centered object and (ii) generate natural-language grounding instructions; for navigation, we convert the annotated camera poses into pose-following instructions.

Experimental results demonstrate that the proposed framework effectively addresses challenging visuospatial tasks. From video context alone, the model acquires scene-level geometry and semantics, and achieves substantially strong performance on object grounding and scene navigation. Trained with short contexts, VIDEO4SPATIAL extrapolates at inference to substantially longer temporal windows, further improving performance. Meanwhile, though being trained on indoor scenes, it generalizes to outdoor environments and to object categories unseen during training.

In summary, we: (1) introduce VIDEO4SPATIAL, a simple yet effective video generation framework that relies solely on video-based scene context to perform visuospatial tasks; (2) highlight key design choices—joint classifier-free guidance over context and instruction, auxiliary bounding boxes as a reasoning prior to improve grounding accuracy, and non-continuous context sampling for efficient context understanding; and (3) instantiate and evaluate two spatial

reasoning tasks: scene navigation and object grounding, which require consistency with 3D geometry and semantic layout without auxiliary modalities (*e.g.*, depth, point clouds). These results advance video-only spatial reasoning, and we hope our work will inspire future research on leveraging video generation models for visuospatial intelligence.

## 2. Related Work

**Generative Video Models: from Renderer to Reasoner.** Building on advances in diffusion models [10, 50, 56], video generation has progressed rapidly [11, 25, 35, 47, 64, 71, 90]. Video diffusion models can function as high-fidelity, controllable renderers, with generation steered by control signals such as camera control [5, 6, 27, 72, 79], sketches [70], depth maps [70], trajectories [16, 18, 19, 60, 89], point cloud [51, 92, 93] and human poses [97, 104].

Powered by the rapid growth of web-scale training, video models [1, 22, 37, 48] have begun to exhibit surprising world priors [21, 39, 44, 98] beyond visual fidelity, like physical awareness [44] and reasoning ability [74]. In particular, Wiedemer et al. [74] demonstrate the potential of generative video models on reasoning tasks such as maze solving and robot navigation, marking a transition from mere rendering to genuine reasoning. Our work focuses on video model’s visuospatial reasoning capability with visual context.

**Video-to-Video Generation.** Beyond generating videos solely from text or other multimodal controls, video-to-video generation [43, 68, 69] conditions on an input video to produce a new video. It has been explored across tasks such as video editing [42, 49, 83], video outpainting [12], video super-resolution [66, 81, 101], and camera control [5, 24, 79]. These settings typically adopt frame-to-frame transformations, which constrain the output’s content and duration. A more general paradigm treats the input video as context while allowing free-form generation, requiring stronger long-range modeling and attention. Recent efforts in this direction include multi-shot movie generation [9, 26, 77] and autoregressive video generation [31, 94]. However, how to effectively leverage scene-level video context for spatial tasks remains underexplored.

**Visuospatial Intelligence.** Visuospatial Intelligence (VSI) [82] is the ability to perceive [65, 67], represent [36, 45], and act [2, 15, 73] spatially from visual input. Prior works can be roughly categorized by output modality. Many studies probe visuospatial intelligence with vision-language models (VLMs) [38, 88, 95], where the output is text. Conditioning language models on visuals alone often falls short, so several methods add explicit 3D signals (*e.g.*, point clouds, depth) to the context window [75, 80, 103]. Recent work also brings visual generation into the reasoning loop and shows promise on spatial tasks [84, 86]. Al-

ternatively, many works study VSI through video generation [13, 34], where the output is video. For spatially consistent video generation, some reconstruct external 3D memories (*e.g.*, point clouds) [40, 76] extracted by 3D visual foundation models [65, 67]. Other approaches rely solely on visual input but still require camera-pose annotations [14, 78, 91, 96, 100]. In contrast, we show that conditioning video generative models solely on raw video sequences suffices for spatial understanding, enabling an end-to-end, scalable approach without explicit 3D signals like depth or pose.

## 3. Methods

This section formalizes video generation as spatial reasoning (Section 3.2) and describes the evaluation tasks (Section 3.3) and key design choices (Section 3.4).

### 3.1. Preliminaries

**Video Diffusion Models (VDMs).** Diffusion models [29, 54, 56] learn a score function  $s_\theta(\mathbf{x}, t, c) \approx \nabla_{\mathbf{x}} \log p_t(\mathbf{x} | c)$ , where  $p_t$  is the data distribution smoothed by Gaussian noise at level  $t$ , and  $c$  denotes conditioning. At inference, a sample from the standard Gaussian is iteratively denoised using  $s_\theta$  to recover the data distribution. VDMs [30] produce high-fidelity, temporally consistent videos, benefiting from transformer backbones [50, 62] and latent diffusion [52]. Typical conditioning includes the first frame [8, 85] and/or a text instruction [30, 64]. History-conditioned video generation [55] shows that  $c$  can include preceding video frames treated as part of the generation sequence inside the transformer. We adopt this architectural template and extend  $c$  to *general video context* (Section 3.2).

**Classifier-Free Guidance (CFG).** CFG [28] improves conditional sampling by learning with randomly dropped conditioning ( $c = \emptyset$ ) and, at sampling, using

$$s_\theta(\mathbf{x}, t, \emptyset) + \omega(s_\theta(\mathbf{x}, t, c) - s_\theta(\mathbf{x}, t, \emptyset)), \quad (1)$$

where  $\omega \geq 1$  is the guidance scale.

**Rotary Positional Embeddings (RoPE).** RoPE [57] injects relative spatial-temporal position by rotating queries and keys, enabling attention by displacement rather than absolute indices. In video diffusion, RoPE improves alignment across frames and stabilizes motion [63, 99]. In practice, attention heads can be partitioned over spatial/temporal axes with per-axis frequency scaling to match resolution and clip length.

### 3.2. Video Generation Models as Spatial Reasoners

Prior work often frames video generative models as *neural renderers* [24, 51], prioritizing visual fidelity while overlooking the reasoning potential afforded by large-scale pretraining on web-scale corpora rich in structural spatial

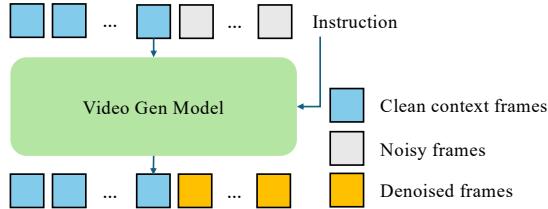


Figure 2. We frame video diffusion models as spatial reasoners, treating context and target frames equally in the model architecture except that context frames are noise-free.

knowledge. We instead cast video generation as *spatial reasoning*: given video-only context, the model produces coherent, goal-directed outcomes following instructions.

Concretely, given a context video  $\mathbf{x}_{\text{ctx}}$  (frames from the observed environment) and an instruction  $\mathbf{g}$  (textual, visual, or other structured forms), the model synthesizes

$$\mathbf{x}_{\text{out}} \sim p_{\theta}(\cdot | \mathbf{x}_{\text{ctx}}, \mathbf{g}) \approx p(\cdot | \mathbf{x}_{\text{ctx}}, \mathbf{g}), \quad (2)$$

where  $p_{\theta}$  is the model distribution and  $p$  the true posterior over videos consistent with  $\mathbf{x}_{\text{ctx}}$  and  $\mathbf{g}$ . We implement context condition by concatenating  $\mathbf{x}_{\text{ctx}}$  with the noisy target frames along the temporal axis (Fig. 2); context frames are fixed at noise-free ( $t = 0$ ) during training and inference, while the target frames share the same diffusion noise level. The instruction tokens (text or structured signals such as poses) can be injected via cross-attention [63], learned embedding addition [27], or token-wise concatenation [5, 17].

### 3.3. Solving Spatial Tasks with Videos

We explore video model’s spatial reasoning capability in realistic environments (homes, offices, factories) using the following two tasks: *object grounding* (text instruction) and *scene navigation* (pose instruction). Both tasks assess how well a model can infer 3D structure, maintain contextual consistency, and generate coherent motion that aligns with a given instruction. Both tasks use the same format described in Eq. 2; unless stated otherwise, the first target frame is also provided as part of  $\mathbf{x}_{\text{ctx}}$ .

**Video-based object grounding.** Given a context video and a natural-language instruction specifying a target object, the model generates a sequence that moves from the initial viewpoint and ends with the target prominently localized in the final frame. We evaluate this task through outputs’ 3D geometric consistency with the context and accuracy to find the object, details in Sec. 4.1.

**Video-based scene navigation.** Given a context video and a sequence of *egocentric* 6-DoF pose waypoints  $\{(\Delta\mathbf{t}_k, \Delta\mathbf{r}_k)\}_k$  ( $\Delta\mathbf{t} \in \mathbb{R}^3$  in meters;  $\Delta\mathbf{r}$  as yaw–pitch–roll in radians), the model synthesizes novel viewpoints along the specified trajectory (details in Sec. 4.3). While related to



Figure 3. **Auxiliary bounding box.** The model is trained to generate a bbox (pixel values) located at the target object in the final frames as an explicit reasoning pattern [20]. This auxiliary output improves object grounding accuracy.

novel view synthesis, we focus on navigation via continuous video generation without explicit 3D information (e.g., camera pose or depth of the context, or reconstructed geometry).

### 3.4. Key Design Choices

**Joint CFG [28].** Inspired by History Guidance [55], we use the following CFG formulation on the joint conditioning of instruction  $\mathbf{g}$  and  $\mathbf{x}_{\text{ctx}}$ :

$$s_{\theta}(\mathbf{x}, t, \emptyset, \mathbf{z}_{\text{ctx}}) + \omega(s_{\theta}(\mathbf{x}, t, \mathbf{g}, \mathbf{x}_{\text{ctx}}) - s_{\theta}(\mathbf{x}, t, \emptyset, \mathbf{z}_{\text{ctx}})),$$

where  $\mathbf{z}_{\text{ctx}} \sim \mathcal{N}(0, \mathbf{I})$  is pure noise. As shown in Table 1, such joint CFG significantly improves the quality and consistency of the output.

**Auxiliary bounding box for object grounding.** Injecting explicit reasoning patterns have been shown to improve language models’ capabilities [20]. By exploiting the fact that video is a flexible medium, for the object grounding task, we annotate and augment the training videos at the end with a red bounding box (bbox) centered at the target object (Fig. 3). This encourages the model to draw a bbox, reinforcing the grounding objective, leading to substantially improved accuracy (Table. 1).

**Non-contiguous context sampling.** A context video of contiguous frames exhibits substantial redundancy. To increase information content while keeping computation tractable, we train by sampling non-contiguous context frames from source videos. Crucially, we apply RoPE to these sparse frames using their original source indices (Fig. 4). This encourages the model to learn temporal coherence and improves robustness when extrapolating to longer context lengths at inference.

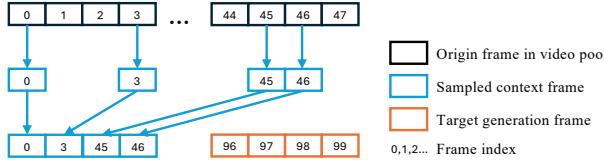
Additionally, we must also distinguish context frames from the target frames being generated, especially to prevent information leakage when both are sampled from the same source video. Therefore, inspired by [58], we set the RoPE indices for the generated frames to begin after the final context frame’s index, separated by a fixed interval of 50. It ensures a clear separation between the context and the target, and mitigates potential information leakage.

## 4. Experiments

**Implementation Details.** We fine-tune the attention

**Table 1. Evaluation on video-conditioned object grounding.** Spatial Distance (SD) evaluates 3D geometric consistency with the context. Instruction Following Rate (IF) measures the model’s ability to follow instructions. IF ( $SD < \delta$ ) measures the overall ability to ground the target object while maintaining spatial geometry consistency. Imaging Quality (IQ) and Dynamic Degree (DD) evaluate overall video quality. We resize all results to the same resolution for evaluation.

Methods	Resolution	# Context	# Generation	SD ↓	IF ↑	IF ( $SD < 0.2$ ) ↑	IF ( $SD < 0.1$ ) ↑	IQ ↑	DD ↑
GT*	416×256	-	161	0.0739	1.0	1.0	0.8260	0.6539	0.9783
Wan2.2-5B [63]	1280×704	1	121	0.5341	0.9439	0.2242	0.0934	0.6101	0.9277
Veo3 [22]	1280×720	1	192	0.2211	0.9532	0.4599	0.3821	0.7056	0.7487
FramePack [94]	704×544	104+1	180	0.3672	0.3	0.1	0.1	0.6315	0.9678
Ours	416×256	336+1	161+20	0.1099	0.7327	0.6486	0.4941	0.6435	0.9859
Ours w/o pretraining	416×256	336+1	161+20	0.4317	0.4990	0.2186	0.1214	0.5845	0.9803
Ours w/ only first frame	416×256	1	161+20	0.4053	0.7700	0.2616	0.1084	0.6219	0.9842
Ours w/o CFG	416×256	336+1	161+20	0.3890	0.4042	0.2031	0.1113	0.5299	0.9985
Ours w/o context CFG	416×256	336+1	161+20	0.8294	0.8841	0.0373	0.0112	0.4909	0.9937
Ours w/o auxiliary bbox	416×256	336+1	161	0.1102	0.6191	0.5401	0.3551	0.6456	0.9883
Ours w/ vanilla RoPE	416×256	336+1	161+20	0.2079	0.7383	0.4719	0.2570	0.6239	0.8906



**Figure 4. RoPE indexing for non-contiguous context.** Context frames(blue) are sampled sparsely to reduce redundancy but retain their original temporal indices. Target generation frames (orange) are given new indices with a fixed offset.

layers of video generation model Wan2.2 [63] with a flow-matching objective [41]. We use a 1e-4 learning rate and a batch size of 160 and train for 20K steps for object grounding and 10K steps for scene navigation. Unless stated otherwise, we use 337 context frames and generate 161 frames. Conditioning is provided via (i) text instructions injected through cross-attention or (ii) camera poses encoded with Plücker coordinates [53] and added to the frame tokens. See the supplementary materials for additional details.

**Datasets and Training.** Datasets are curated from ScanNet++ [87] and ARKitScenes [7]. For object grounding, we use Qwen3-VL [59] to identify video clips with an object centered in the last frame and generate text instructions, resulting in 400K text-video pairs (Fig. 6). For scene exploration, we directly use ScanNet++ with 1K videos and camera pose annotations and randomly sample short clips for training. For both tasks, we evaluate on randomly selected 18 scenes from ScanNet++ that are not included in the training set.

#### 4.1. Video-based Object Grounding

Prior spatial-consistency metrics typically focus on static-background consistency [32] or on the geometric consistency of generated videos [4]; and to our knowledge, there are no simliar setting on object grounding as ours.

Accordingly, we propose two metrics tailored to this task. **Spatial Distance (SD).** SD measures whether the generated scene is contained in the ground-truth(GT) point cloud and is an effective measure against out-of-context hallucination. We use VGGT [65] to reconstruct a point cloud  $\mathcal{P}^{\text{gen}} = \{\mathcal{P}_i^{\text{gen}}\}_{i \in \mathcal{I}}$  from each generated video, where  $\mathcal{P}_i^{\text{gen}}$  is the point cloud of frame  $i \in \mathcal{I}$ . We calculate the maximum per-frame one-sided Chamfer distance of  $\mathcal{P}^{\text{gen}}$  to the ground-truth point cloud  $\mathcal{P}^{\text{gt}}$  as:

$$d(\mathcal{P}^{\text{gen}}, \mathcal{P}^{\text{gt}}) = \max_{i \in \mathcal{I}} \frac{1}{|\mathcal{P}_i^{\text{gen}}|} \sum_{x \in \mathcal{P}_i^{\text{gen}}} \min_{y \in \mathcal{P}^{\text{gt}}} \|x - y\|_2, \quad (3)$$

and we report the mean of  $d$  over the test set. We use the maximum to penalize severe single-frame geometric inconsistencies, which we regard as critical failures for navigation. Figure 7 shows that the point cloud reconstructed from VGGT using generated frames from our method align well with the ground-truth point cloud.

**Instruction Following (IF).** To measure if the model is capable of navigating and locating the target object, we define IF score: for a total of  $N$  generated videos, we use Qwen3-VL to check if the last frame of each video  $k \in [N]$  contains an object matching the target category, and if so, we further prompt Qwen3-VL to output bbox parameters  $(x_k, y_k, W_k, H_k)$  for that object. We consider the grounding successful if

$$|x_k - x^*| \leq \frac{\gamma}{2} W_k \quad \text{and} \quad |y_k - y^*| \leq \frac{\gamma}{2} H_k,$$

where  $(x^*, y^*)$  is the center of the frame, and  $\gamma$  is a tolerance ratio we set to 0.6. The IF score is then defined to be the success rate of grounding across  $N$  videos.

Since there might be scenarios where the object gets hallucinated in close to the final frames, achieving high IF score becomes meaningless if SD is high. To evaluate spatial consistency and grounding jointly, we also propose SD-thresholded IF, denoted IF( $SD < \delta$ ), which we calculate

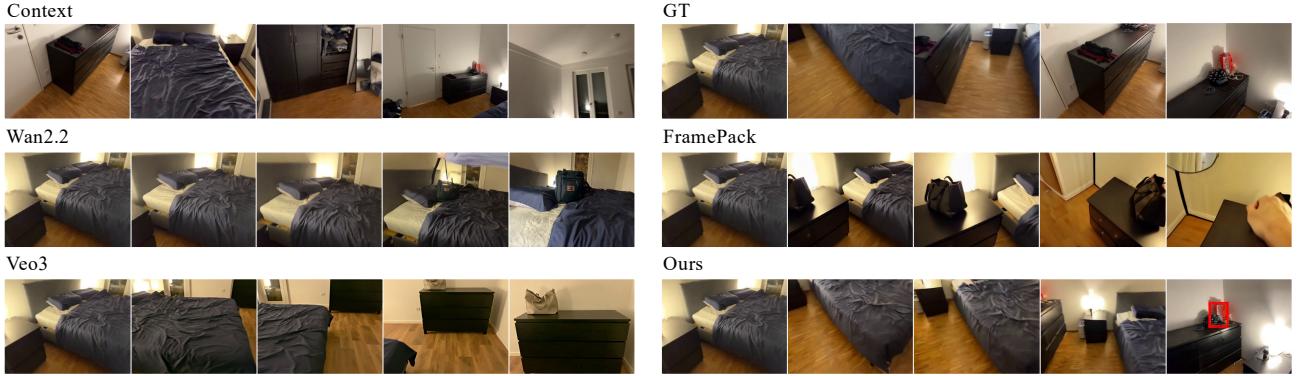


Figure 5. **Qualitative comparison across methods.** The instruction is: “The camera moves smoothly through a bedroom. Finally, it focuses on a bag on top of the dresser in the center of the frame.” Our method faithfully grounds the object present in the context, whereas other methods hallucinate the target object. Notably, the path to locate the target from our generated result is different from the GT. For the best experience, see the supplementary videos.

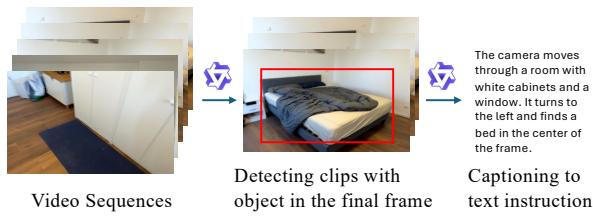


Figure 6. Object grounding data collection pipeline.

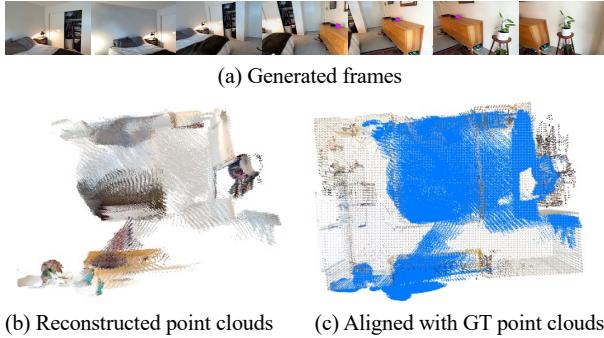
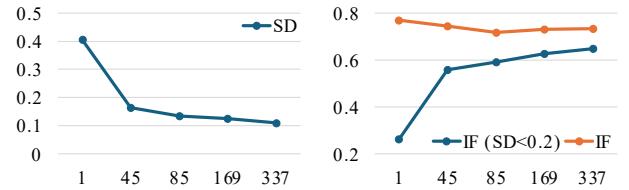


Figure 7. **Visualization of Spatial Distance.** We use generated frames (a) to reconstruct the point cloud (b), which is compared with the GT point cloud (c): Reconstructed point cloud (blue) is mostly overlapped with the GT point cloud, indicating our generated frames are spatially consistent. Distance between point clouds are used to measure the spatial consistency quantitatively.

among videos of which SD is less than  $\delta$ . By default we evaluate using  $\delta = 0.2$  and  $\delta = 0.1$

Additionally, we report VBench Imaging Quality (IQ) and Dynamic Degree (DD) [32] to assess frame-wise quality and motion dynamics as complementary metrics.

We evaluate object grounding on 107 prompts



(a) Ablations on different context frame number for inference

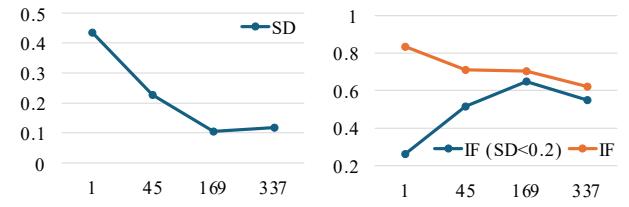


Figure 8. **Ablations on condition frame numbers for training and inference.** In the inference ablation (a), we fix the training number at 169; in the training ablation (b), we fix the inference number at 337.

from the 18 validation scene in the form of: “The camera moves across <SCENE DESCRIPTION>. Finally, it focuses on <OBJECT DESCRIPTION> in the center of the frame.” Table 1 reports results for our method alongside Wan2.2-5B [63], Veo3 [22] as SOTA open and closed-source image-to-video model, and FramePack [94] due to its capability to handle long context input. Since those methods use different video-conditioned settings, generation resolutions and the numbers of generated frames, we enumerate configuration differences in Table 1 for reference. Notably, Wan2.2-5B [63] and Veo3 [22] only support single image input; thus, the context frame count is 1. For FramePack [94], we adopt its default setting of

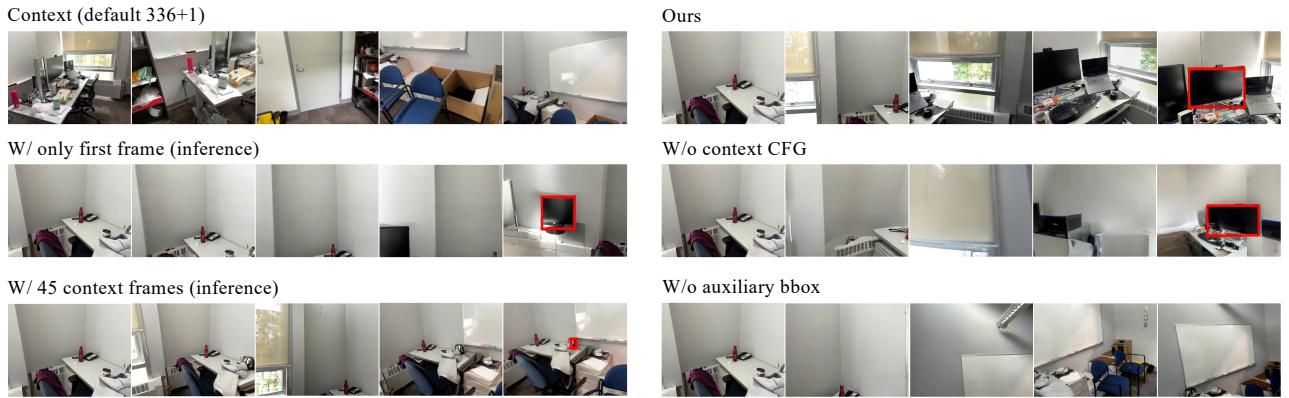


Figure 9. **Visualization for ablations on condition frame number, CFG, and auxiliary bbox.** The instruction is “The camera moves through an office space. Finally, it focuses on a **monitor** in the center of the frame.” For the best experience, see the supplementary videos.

105 context frames. Our default setting uses 337 context frames (including the first frame of the target video) of the generated video as context frames. By default, we generate 161 frames; when generating auxiliary bounding boxes at the video end, we add 20 frames (161+20). For reference, we also report the same metrics on ground-truth (GT) videos for these 18 scenes.

We present quantitative results in Table 1 and qualitative examples in Fig. 5. Our evaluation shows that though Wan2.2 [63] and Veo3 [22] exhibit strong instruction following capability, their generation is often hallucinatory: without video conditioning, their SD and  $\text{IF}(\text{SD} < \delta)$  degrade markedly. FramePack [94] conditions on history frames, yet still hallucinates due to heavily compressed context tokens that discard critical spatial cues. In contrast, our method delivers faithful grounding while preserving scene consistency: our SD approaches that of GT, indicating strong geometric alignment, and we achieve competitive IQ alongside high motion dynamics on DD.

## 4.2. Ablations on Object Grounding

We further ablate several key design choices. The main results are in the bottom part of Table 1 and Fig. 8 and 9.

**Context frame length.** Context length is a key factor for both training and inference. Our default uses 169 frames (43 latent frames) for training and 337 frames (85 latent frames) for inference. As shown in Table 1, using only the first frame without other context yields a much higher SD (0.4043) than using context (0.1099), indicating substantially greater inconsistency relative to the contextual environment. Conversely, IF improves with only the first frame because generation is no longer constrained and may hallucinate the target object for grounding.

To determine the optimal context length, we conduct ablations study shown in Fig. 8. In (a), we train with 169 context frames and vary inference from 1–337 frames. Both SD and  $\text{IF}(\text{SD} < 0.2)$  improves monotonically as the context

frame number increases. Interestingly, IF first decreases, then rises, and finally saturates. With insufficient context, the model tends to hallucinate, producing poor spatial consistency (higher SD) but faked grounding (higher IF). As more context frames are provided, generation becomes constrained; IF initially drops when the model is constrained yet lacks sufficient evidence for grounding, then increases once context is rich enough to support accurate scene understanding.

In (b), we vary the training context from 1–337 frames while fixing inference at 337 frames. All models share identical settings and sampling steps, differing only in training context frame length. Increasing training context generally improves spatial consistency (lower SD), whereas IF ( $\text{SD} < 0.2$ ) first increases, then declines, with a pronounced drop at 337 frames. We hypothesize that excessive training context biases the model toward adhering to visual context at the expense of following text instructions. Training with 169 context frames provides a strong balance.

Notably, models trained with shorter context frames (e.g., 169) generalize well to longer inference context frames (e.g., 337).

**Auxiliary bounding box.** Auxiliary bbox design plays a significant role in improving grounding accuracy. As shown in Table 1, training the model to predict the target object’s bbox raises  $\text{IF}(\text{SD} < 0.2)$  from 0.5401 to 0.6486. In Fig. 9 we show example output with and without auxiliary bbox: for the “monitor” target, prediction without the bbox leads the model to drift to irrelevant objects, whereas providing the bbox performs correct grounding of the monitor.

**CFG.** As shown in Table 1, with no CFG (pure conditional generation), both SD and  $\text{IF}(\text{SD} < \delta)$  degrade markedly. If we retain text CFG but remove context CFG, the model tends to hallucinate without enforcing contextual consistency: IF appears high, but SD and  $\text{IF}(\text{SD} < \delta)$  are significantly worse.

**Non-continuous RoPE for non-contiguous context.** We



Figure 10. **Qualitative results on scene navigation of a bedroom(left) and a kitchen(right).** Our method delivers the highest perceptual quality and good camera controllability. Please refer to the supplementary videos for clearer visual comparisons.

find that non-continuous RoPE helps the model reason over non-contiguous context. As shown in Table 1, using vanilla (continuous) RoPE over non-contiguous context substantially worsens SD and IF ( $SD < \delta$ ), indicating poorer spatial understanding of the contextual environment.

**Pretraining.** Pretraining confers clear benefits; without it (*i.e.*, training from scratch), SD, IF, IF( $SD < \delta$ ), and IQ degrade markedly (Fig. 1).

### 4.3. Video-based Scene Navigation

For scene navigation, we report PSNR, LPIPS and IQ for generation fidelity and adherence to the requested camera trajectory.

We evaluate video-based scene navigation on 18 ScanNet++ [87] scenes, each with 5 camera trajectories (90 cases in total). By default, we condition on 337 context frames (85 latent) and generate 161 output frames (41 latent). We compare against the feed-forward 3D reconstruction method AnySplat [33] and the video generation model Gen3C [51] and TrajectoryCrafter [92]. Notably, all these methods rely on external estimators (*e.g.*, VGGT [65]) for extra 3D information. For our method, it only takes context frames and no 3D information is needed. See the supplementary material for detailed settings.

In Table 2, we report comparative results. Because PSNR and LPIPS emphasize structural similarity, AnySplat [33] attains the highest scores on these metrics, though with significant artifacts it achieves lower perceptual IQ score. Our method delivers competitive PSNR and LPIPS while achieving substantially better IQ. Moreover, we outperform the video-based Gen3C [51] and TrajectoryCrafter [92] across all metrics; unlike these video-based methods, which heavily depends on external estimators, our approach achieves strong end-to-end performance. Fig. 10 further visualizes the results: although AnySplat achieves high PSNR, it exhibits noticeable artifacts, whereas our outputs are visually cleaner but not perfectly aligned with GT. We further supply qualitative results for visualization in Fig. 10

Table 2. **Evaluation on video-conditioned scene navigation.** “3D info.” indicates explicit 3D information required for each method. Our method does not need any explicit 3D information.

Method	3D info.	PSNR $\uparrow$	LPIPS $\downarrow$	IQ $\uparrow$
3D reconstruction				
AnySplat [33]	Cam. pose	<b>16.40</b>	<b>0.4402</b>	0.4568
Video generation				
TrajectoryCrafter [92]	Cam. pose, depth	13.56	0.4920	<u>0.5630</u>
Gen3C [51]	Cam. pose, depth	12.76	0.5695	0.3628
Ours	-	14.27	<u>0.4674</u>	<b>0.6320</b>

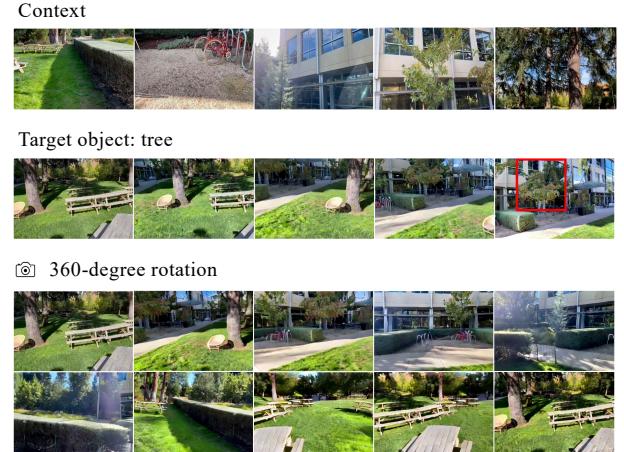


Figure 11. **Out-of-domain result.** Our models can generalize to OOD scenarios like outdoor scenes and new categories, performing object grounding and scene navigation.

### 4.4. Out-of-domain generalization.

Although trained on only indoor scenes, our model generalizes well to out-of-domain environments for both tasks. In Fig. 11, we show results on a real-captured outdoor park: the model reliably grounds object categories that were rarely encountered during indoor training (*e.g.*, trees), and supports free-form navigation, including 360° rotations.

## 5. Conclusions and Future Work

We introduced VIDEO4SPATIAL, a video-only framework that probes visuospatial intelligence in generative models by conditioning solely on scene context. Across scene navigation and object grounding, VIDEO4SPATIAL executes geometry-consistent camera trajectories and good target grounding, while generalizing to long contexts and out-of-domain environments.

Looking ahead, we note several limitations and avenues for future work. Our current approach operates at a modest  $416 \times 256$  resolution due to the absence of context compression (*e.g.*, [23, 96]), which we identify as a key lever for improving visual fidelity at higher resolutions. In addition, stronger temporal modeling, targeted data augmentation, and improved grounding objectives could mitigate artifacts such as temporal discontinuities and incorrect grounding on long-tail categories. Finally, extending this framework beyond static scenes to complex, dynamic environments and a broader range of tasks is an exciting direction.

## 6. Supplementary Materials

### 6.1. More Qualitative Results

It is highly recommended to refer to the webpage for visualization.

### 6.2. Implementation Details.

**Data Curation.** We sample frames from ScanNet++ [87] and ARKitScenes [7] at an interval of three to reduce temporal redundancy. To curate datasets for object grounding, for each sampled frame, we use Qwen3-VL [59] to perform object grounding with the following prompt template:

#### Object Grounding Prompt Template

You are a helpful assistant. Identify ALL objects that belong to these categories: YOUR\_CATEGORY.

Requirements:

1. Return ALL instances of objects from these categories (can be multiple).
2. Each object must be CLEARLY VISIBLE with SHARP, DISTINCT boundaries (not blurry or pixelated).
3. If the image is blurry, low quality, or no clear objects exist, return an empty array: [].
4. Format: [{"label": "DETECTED\_CATEGORY", "box": [x1, y1, x2, y2], ...}]

We then filter detections using a center-region threshold (center ratio) and select the instance closest to the image center as the grounding target. For each clip, we take a frame where the target is centered as the final frame and extract the preceding 161 frames (inclusive) to form the target video clip. We also use Qwen3-VL to caption the clip with the following prompt:

#### Video Caption Prompt Template

You are a helpful assistant. Describe a video sequence in which the camera moves through the environment and, at the end, a {centered\_object\_label} appears centered in the frame. Use simple sentences. Do NOT use complex grammar. Good example: “The camera moves through a room. It pans left and right. At the end, a {centered\_object\_label} is centered in the frame.” Bad example: “The camera pans around, revealing various objects and eventually discovering a {centered.object.label} positioned in the center, where it comes to rest in the frame.” Provide a concise description in 1–2 simple sentences.

**Training with context.** During training, we sample both context frames and target generation frames from source

videos. Target clips are 161 frames long. Context frames are randomly drawn but always in contiguous groups of four to satisfy the Wan2.2 [63] VAE encoder. For efficiency, we pre-encode frames into latents during preprocessing and sample latents directly, yielding roughly 2 $\times$  faster training. A key consideration is ensuring sufficient overlap between context and target frames so the model can learn geometric consistency. ScanNet++ sequences are long with ample overlap, so we sample context and targets from the same clip. ARKitScenes sequences are shorter with less view overlap but include multiple videos per environment; therefore, we sample context and targets from different videos within the same environment.

**Spatial Distance.** To measure the distance between the point cloud reconstructed by VGGT [65] and the ground-truth point cloud, we first register the two coordinate frames. ScanNet++ [87] provides frames with calibrated camera poses and depth maps. When reconstructing point clouds from generated videos, we append 40 ground-truth frames as anchors and use the estimated point clouds of these anchors to perform registration. After alignment, we compute the point-cloud distance.

### 6.3. Benchmark setting.

Since it is not feasible to make the setting for different methods perfectly fair, we list the detailed setting for each methods for reference.

**Wan2.2-5B [63].** We generate 121-frame videos at a resolution of  $1280 \times 704$ . The model takes the first frame and a text instruction as input, utilizing 50 inference steps and a classifier-free guidance scale of 5.0.

**Veo 3 [22].** We utilize the Veo 3 image to video model via its official API, generating videos at  $1280 \times 720$  with 192 frames.

**FramePack [94].** We employ FramePack to generate videos at a resolution of  $704 \times 544$ . The inference process utilizes 50 sampling steps with a classifier-free guidance scale of 3.0. The model is conditioned on a context window of 105 frames, which are encoded into a multi-scale latent representation comprising 16 latent frames at 4x compression, 2 latent frames at 2x compression, and 9 latent frames at 1x compression.

**Anysplat [33].** We employ Anysplat for 3D reconstruction and rendering for video outputs. The model utilizes 84 context frames along with the first frame serving as an anchor for reconstruction, which correspond to 84 latent frames plus the first frame for the video generation setting. We estimate camera poses using VGGT [65] and use them for coordinate regularization.

**Gen3C [51].** We follow the official implementation of Gen3C and its instruction on Multiview Images Input. First, we run VGGT [65] to get the depth information, camera intrinsics, and extrinsics of the context frame. We also include



Figure 12. Failure and success cases.

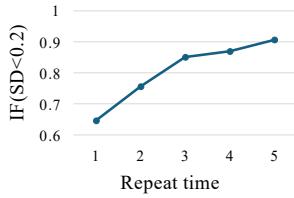


Figure 13. Ablation on grounding success rate over repeat time.

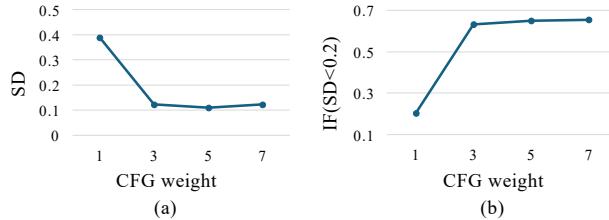


Figure 14. Ablation on different CFG weights .

the first frame of the generation as the first context frame. During the VGGT inference, we include context frames and ground truth frames as a single run to get the camera poses of the ground truth. During the Gen3C inference, we use the camera poses of ground truth as camera control. We input 85 context frames and output 121 frames, all at resolution 576x320.

**TrajectoryCrafter [92].** We use the same preprocessing algorithm as Gen3C, *i.e.* VGGT, to get the point cloud from 85 context frames and camera pose control from 121 ground truth frames, all in one pass to make sure they are in the same coordinate system. Then we project the point cloud based on ground truth camera and get masks. Since the

model is trained to generate 49 frames, we run three inferences to get the complete output.

#### 6.4. Experiments.

**Repeat time.** Consistent with prior findings that Veo3 [22] improves with more repeats [74], we observe the same trend. For object grounding, we compute IF ( $SD < 0.2$ ) across repeated runs and deem a case successful if any repeat succeeds. As shown in Fig. 13, increasing the number of repeats from 1 to 5 raises IF ( $SD < 0.2$ ), indicating that repeated sampling substantially boosts performance.

**CFG strength.** We investigate the impact of different CFG scales in Fig. 14. While a weight of 1 (no CFG) yields sub-optimal results, performance improves significantly once the scale reaches a sufficient level (e.g., 3) and remains robust across a wide range (e.g., 3–7). This suggests that the use of CFG itself is more critical than fine-tuning the exact weight.

**Non-continuous context sampling.** We investigate the effectiveness of non-continuous context sampling during training, as shown in Table 3. Our results demonstrate that non-continuous sampling significantly improves geometric consistency with the context and achieves highly faithful grounding. We hypothesize that this improvement stems from the model developing a more robust spatial understanding through such diverse and non-continuous context modeling.

**Random zero camera pose.** Existing camera control methods typically normalize the camera pose of the first frame to zero during both training and inference. We find that, particularly with limited datasets, fixing the first frame to the origin restricts spatial exploration, thereby limiting generalization in complex scenarios such as  $360^\circ$  rotations. To ad-

Table 3. Ablation on context sampling.

	SD $\downarrow$	IF $\uparrow$	IF(SD < 0.2) $\uparrow$
Non-continuous	0.1099	0.7327	0.6486
Continuous	0.3246	0.8497	0.4600

dress this, we propose randomly selecting a reference frame to serve as the zero pose, which effectively alleviates this issue and improves generalization.

**Computational analysis.** Although our model processes relatively long contexts (default 337 frames) and generates extended sequences (default 181 frames), we maintain computational costs within an acceptable range. Inference requires approximately 34GB of VRAM and takes 2 minutes on a single A100 GPU using CFG and 50 denoising steps, without optimizations such as VAE slicing, tiling, or dynamic model loading [94]. This efficiency stems primarily from two factors: (1) the use of a moderate resolution ( $416 \times 256$ ), which we find sufficient for high visual quality and experimental validation; and (2) the Wan2.2 VAE, which achieves high spatial and temporal compression ratios while preserving visual fidelity.

**Failure cases.** Our method still suffers from artifacts such as temporal discontinuities (Fig. 12(a)) and incorrect grounding (Fig. 12(b)) for some cases.

## 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] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 2018. 2, 3
- [3] Anthropic. Claude 3.5 sonnet model card addendum, 2024. 2
- [4] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6034–6044, 2025. 5
- [5] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. ReCamMaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025. 3, 4
- [6] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *CVPR*, 2025. 3
- [7] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 2, 5, 10
- [8] 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
- [9] Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, et al. Mixture of contexts for long video generation. *arXiv preprint arXiv:2508.21058*, 2025. 3
- [10] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *NeurIPS*, 37:24081–24125, 2025. 3
- [11] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. VideoCrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 3
- [12] Qihua Chen, Yue Ma, Hongfa Wang, Junkun Yuan, Wen-zhe Zhao, Qi Tian, Hongmei Wang, Shaobo Min, Qifeng Chen, and Wei Liu. Follow-your-canvas: Higher-resolution video inpainting with extensive content generation. *arXiv preprint arXiv:2409.01055*, 2024. 3
- [13] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *CVPR*, 2025. 3
- [14] Taiye Chen, Xun Hu, Zihan Ding, and Chi Jin. Learning world models for interactive video generation. *arXiv preprint arXiv:2505.21996*, 2025. 3
- [15] Xiao Chen, Tai Wang, Quanyi Li, Tao Huang, Jiangmiao Pang, and Tianfan Xue. Gleam: Learning generalizable exploration policy for active mapping in complex 3d indoor scenes. *arXiv preprint arXiv:2505.20294*, 2025. 3
- [16] Yufan Deng, Ruida Wang, Yuhao Zhang, Yu-Wing Tai, and Chi-Keung Tang. DragVideo: Interactive drag-style video editing. *arXiv preprint arXiv:2312.02216*, 2023. 3
- [17] 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. 4
- [18] Xiao Fu, Xian Liu, Xintao Wang, Sida Peng, Menghan Xia, Xiaoyu Shi, Ziyang Yuan, Pengfei Wan, Di Zhang, and Dahua Lin. 3DTrajMaster: Mastering 3D trajectory for multi-entity motion in video generation. *ICLR*, 2025. 3
- [19] Xiao Fu, Xintao Wang, Xian Liu, Jianhong Bai, Runsen Xu, Pengfei Wan, Di Zhang, and Dahua Lin. Learning video generation for robotic manipulation with collaborative trajectory control. *arXiv preprint arXiv:2506.01943*, 2025. 3
- [20] Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025. 2, 4
- [21] Google. Genie 3: A new frontier for world models. <https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/>, 2025. 3
- [22] Google. Veo 3. <https://aistudio.google.com/models/veo-3>, 2025. 2, 3, 5, 6, 7, 10, 11
- [23] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024. 9
- [24] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3D-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Papers*, 2025. 3
- [25] 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. In *ICLR*, 2024. 3
- [26] Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhi-jie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. *arXiv preprint arXiv:2503.10589*, 2025. 3

- [27] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 2, 3, 4
- [28] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 3, 4
- [29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 3
- [30] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 2022. 3
- [31] 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. 3
- [32] 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*, 2024. 5, 6
- [33] Lihang Jiang, Yucheng Mao, Lining Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. AnySplat: Feed-forward 3d gaussian splatting from unconstrained views. *SIGGRAPH Asia*, 2025. 8, 10
- [34] Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4D: Leveraging video generators for geometric 4D scene reconstruction. In *ICCV*, 2025. 2, 3
- [35] 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. In *ICLR*, 2025. 3
- [36] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 2023. 3
- [37] Kling. Kling ai: Next-generation ai creative studio. <https://klingai.com/global/>, 2025. 3
- [38] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. LLava-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 3
- [39] Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. WorldModelBench: Judging video generation models as world models. In *CVPR WorldModelBench Workshop*, 2025. 3
- [40] Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. VMem: Consistent interactive video scene generation with surfel-indexed view memory. *arXiv preprint arXiv:2506.18903*, 2025. 3
- [41] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 5
- [42] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jia-aya Jia. Video-P2P: Video editing with cross-attention control. In *CVPR*, 2024. 3
- [43] Arun Mallya, Ting-Chun Wang, Karan Sapra, and Ming-Yu Liu. World-consistent video-to-video synthesis. In *ECCV*, 2020. 3
- [44] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quan-feng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024. 3
- [45] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3
- [46] OpenAI. Hello GPT-4o, 2024. 2
- [47] OpenAI. Video generation models as world simulators. <https://openai.com/research/video-generation-models-as-world-simulators>, 2024. 3
- [48] OpenAI. Sora 2. <https://openai.com/zh-Hant-HK/index/sora-2/>, 2025. 2, 3
- [49] Wenqi Ouyang, Yi Dong, Lei Yang, Jianlou Si, and Xinggang Pan. I2VEdit: First-frame-guided video editing via image-to-video diffusion models. In *SIGGRAPH Asia*, 2024. 3
- [50] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 2, 3
- [51] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-informed world-consistent video generation with precise camera control. In *CVPR*, 2025. 3, 8, 10
- [52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3
- [53] Vincent Sitzmann, Semon Reznikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *NeurIPS*, 34:19313–19325, 2021. 5
- [54] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 3
- [55] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. *ICML*, 2025. 2, 3, 4
- [56] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 3
- [57] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024. 2, 3
- [58] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *ICCV*, 2025. 4
- [59] Qwen team. Qwen3-vl. <https://github.com/QwenLM/Qwen3-VL>, 2025. 2, 5, 10
- [60] Yao Teng, Enze Xie, Yue Wu, Haoyu Han, Zhenguo Li, and Xihui Liu. Drag-a-video: Non-rigid video editing with point-based interaction. *arXiv preprint arXiv:2312.02936*, 2023. 3

- [61] Jingqi Tong, Yurong Mou, Hangcheng Li, Mingzhe Li, Yongzhuo Yang, Ming Zhang, Qiguang Chen, Tianyi Liang, Xiaomeng Hu, Yining Zheng, et al. Thinking with video: Video generation as a promising multimodal reasoning paradigm. *arXiv preprint arXiv:2511.04570*, 2025. 2
- [62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3
- [63] 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, 5, 6, 7, 10
- [64] Juniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 3
- [65] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, 2025. 3, 5, 8, 10
- [66] Jianyi Wang, Zhijie Lin, Meng Wei, Yang Zhao, Ceyuan Yang, Chen Change Loy, and Lu Jiang. Seedvr: Seeding infinity in diffusion transformer towards generic video restoration. In *CVPR*, 2025. 3
- [67] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUST3R: Geometric 3D vision made easy. In *CVPR*, 2024. 3
- [68] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *NeurIPS*, 2018. 3
- [69] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. In *NeurIPS*, 2019. 3
- [70] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Juniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. VideoComposer: Compositional video synthesis with motion controllability. *NeurIPS*, 2024. 3
- [71] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. LAVIE: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 3
- [72] 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*, 2024. 2, 3
- [73] Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, et al. StreamVLN: Streaming vision-and-language navigation via slowfast context modeling. *arXiv preprint arXiv:2507.05240*, 2025. 2, 3
- [74] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025. 2, 3, 11
- [75] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025. 2, 3
- [76] Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *arXiv preprint arXiv:2506.05284*, 2025. 3
- [77] Junfei Xiao, Ceyuan Yang, Lvmin Zhang, Shengqu Cai, Yang Zhao, Yuwei Guo, Gordon Wetzstein, Maneesh Agrawala, Alan Yuille, and Lu Jiang. Captain cinema: Towards short movie generation. *arXiv preprint arXiv:2507.18634*, 2025. 3
- [78] Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025. 3
- [79] Zeqi Xiao, Wenqi Ouyang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan. Trajectory attention for fine-grained video motion control. In *ICLR*, 2025. 3
- [80] Runsen Xu, Zhiwei Huang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. VLM-Grounder: A VLM agent for zero-shot 3D visual grounding. In *Conference on Robot Learning*, 2024. 2, 3
- [81] Yiran Xu, Taesung Park, Richard Zhang, Yang Zhou, Eli Shechtman, Feng Liu, Jia-Bin Huang, and Difan Liu. VideoGigaGAN: Towards detail-rich video super-resolution. In *CVPR*, 2025. 3
- [82] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, 2025. 3
- [83] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023*, 2023. 3
- [84] Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. MindJourney: Test-time scaling with world models for spatial reasoning. *arXiv preprint arXiv:2507.12508*, 2025. 3
- [85] 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
- [86] Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*, 2025. 3
- [87] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023. 2, 5, 8, 10
- [88] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. *arXiv preprint arXiv:2506.21458*, 2025. 3
- [89] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. DragNUWA: Fine-grained

- control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 3
- [90] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast causal video generators. In *CVPR*, 2025. 3
- [91] Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *SIGGRAPH Asia*, 2025. 3
- [92] Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. *ICCV*, 2025. 2, 3, 8, 11
- [93] 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. *TPAMI*, 2025. 3
- [94] Lvmín Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *NeurIPS*, 2025. 3, 5, 6, 7, 10, 12
- [95] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 3
- [96] Tianyuan Zhang, Sai Bi, Yicong Hong, Kai Zhang, Fujun Luan, Songlin Yang, Kalyan Sunkavalli, William T Freeman, and Hao Tan. Test-time training done right. *arXiv preprint arXiv:2505.23884*, 2025. 3, 9
- [97] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. MimicMotion: High-quality human motion video generation with confidence-aware pose guidance. In *ICML*, 2025. 3
- [98] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. 3
- [99] 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
- [100] Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025. 3
- [101] Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *CVPR*, 2024. 3
- [102] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 2
- [103] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. LLaVA-3D: A simple yet effective pathway to empowering lmms with 3D-awareness. *arXiv preprint arXiv:2409.18125*, 2024. 2, 3
- [104] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3D parametric guidance. In *ECCV*, 2024. 3