

CamContextI2V: Context-aware Controllable Video Generation

Luis Denninger¹, Sina Mokhtarzadeh Azar¹, Juergen Gall^{1,2}

¹University of Bonn, ²Lamarr Institute for Machine Learning and Artificial Intelligence

Abstract

Recently, image-to-video (I2V) diffusion models have demonstrated impressive scene understanding and generative quality, incorporating image conditions to guide generation. However, these models primarily animate static images without extending beyond their provided context. Introducing additional constraints, such as camera trajectories, can enhance diversity but often degrades visual quality, limiting their applicability for tasks requiring faithful scene representation. We propose CamContextI2V, an I2V model that integrates multiple image conditions with 3D constraints alongside camera control to enrich both global semantics and fine-grained visual details. This enables more coherent and context-aware video generation. Moreover, we motivate the necessity of temporal awareness for an effective context representation. Our comprehensive study on the RealEstate10K dataset demonstrates improvements in visual quality and camera controllability. We make our code and models publicly available at: <https://github.com/LDenninger/CamContextI2V>

1. Introduction

Diffusion models have become a prominent approach for video generation producing high-quality videos based on user inputs. To make such approaches attractive for digital content creation, controllability achieved through specific conditioning of the generations, like human poses [18, 26], style [10, 29], motion [19, 22] or camera trajectories [7, 25, 32] have been a widely studied topic.

While text-to-video (T2V) diffusion models like VideoCrafter [3] or CogVideoX [28] have full freedom over the visual design, more recent image-to-video (I2V) models employ an image to convey style and scene context. Due to the typically short duration (< 2 seconds) of the generated videos, the image provides sufficient context to define the scene to render. With the ultimate objective of matching the generative quality and capabilities of traditional rendering engines, these approaches still require further development to achieve a fine-grained control over style, motion and scene composition, to allow for fully customizable video

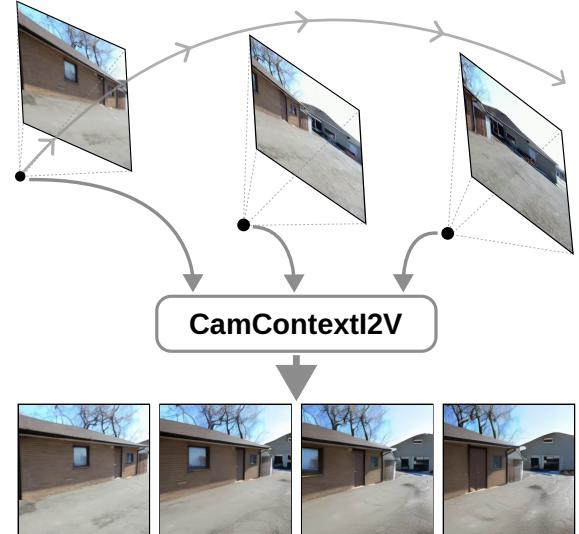


Figure 1. CamContextI2V performs context-aware generation provided a reference frame representing the initial frame and [1-4] additional views providing crucial context to the diffusion process missing in the reference frame.

creation.

As illustrated in Fig. 1, the initial reference frame alone provides only limited context for the diffusion process. Once the camera pans, the visual quality degrades and arbitrary interpretations of the scene by the diffusion model become evident. To address this, we introduce CamContextI2V, a novel conditioning mechanism that allows users to supply multiple context views, ensuring a comprehensive definition of the scene in which the video is generated.

Our proposed *context-aware encoder* integrates these context views into two complementary streams: a high-level semantic stream and a 3D-aware visual stream. This dual-stream approach provides the diffusion model with both a global semantic context and a detailed pixel-level visual embedding. By inserting 3D geometric constraints in the feature aggregation, we effectively retrieve important features from the context while filtering out irrelevant ones. This allows our method to considerably enhance the visual coherence of existing approaches. In summary, our key con-

tributions are as follows.

- We propose CamContextI2V, a camera-controllable context-aware diffusion model, which conditions the diffusion process on multiple context frames through a dual-stream encoder retrieving high-level semantic features and low-level visual cues from the context.
- We introduce a 3D-aware cross-attention mechanism leveraging epipolar constraints to effectively retrieve context from posed images.
- Our temporally-aware embedding strategy better aligns the context at different frame timesteps.
- Our method achieves a 24.09% improvement in visual quality over the state-of-the-art methods on RealEstate10K dataset.

2. Related Works

Diffusion-based Video Generation. Originally developed for image generation [8, 15], diffusion models have since demonstrated great success synthesizing high-quality videos [2, 9]. Models such as SVD [1], LAVIE [20] or VideoCrafter [3] have shown great success in distilling text-to-video (T2V) diffusion models from text-to-image (T2I) diffusion models by inserting temporal attention blocks modeling the added time dimension. Building on top, models like DynamiCrafter [24], Seine [4] or I2vgen-XL [31] further fine-tune these models for image-to-video (I2V) generation showing impressive results.

Camera-controllable Video Generation. Concurrent work also focuses on adding camera control to diffusion models allowing the user to define the trajectory along a video is generated. While initial work such as MotionCtrl [22], AnimateDiff [6] or Direct-a-Video [27] model camera movements through camera-motion primitives, recent approaches such as CameraCtrl [7], CamCo [25] or CamI2V [32] directly insert the camera poses showcasing fine-grained camera control. A key is the dense supervisory signal, such as pixel-wise camera rays represented through Plücker coordinates, which are encoded and inserted into the diffusion model in a ControlNet-like fashion [30].

CamCo and CamI2V further demonstrate that epipolar geometry can serve as an effective constraint in the information aggregation of vanilla attention mechanism. While CamCo employs cross-attention to constraint the feature aggregation from the condition frame, CamI2V constrain the temporal self-attention itself to guide the diffusion process and thus improving the 3D consistency and camera trajectory.

Multi-Image Condition. Large camera movements or longer generations result in multiple scenes being generated in one video which is insufficiently represented through a

singular reference image typically employed in concurrent image-to-video diffusion models [4, 24, 28, 31]. Models like Gen-L-Video [17], MEVG [13] or VideoStudio [11] explore the insertion of multiple text prompts to give a broader context across the temporal domain for longer video generation. This is achieved by generating distinct short videos with different text conditions and optimizing the noise between them either in a divide-and-conquer or auto-regressive setup to generate long consistent videos. Other approaches like 4DiM [23] or Seine [4] explore the insertion of multiple image conditions to induce motion cues or as key-frames to interpolate between.

3. Preliminaries

Before we describe in Section 4 our novel method, which enhances the context-awareness of pre-trained diffusion models by conditioning on multiple context views rather than a single reference frame, we briefly describe components of our baseline model, CamI2V, which extends DynamiCrafter [24], a latent image-to-video diffusion model with camera pose conditioning.

Latent Video Diffusion Models. Latent video diffusion models learn a latent video data distribution by gradually reconstructing noisy latents z_t sampled from a Gaussian distribution:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1 - \beta_t}, \beta_t \mathbf{I}), \quad (1)$$

where hyperparameters β_t determine the level of noise added at each timestep. The latent space is defined through a pre-trained auto-encoder, e.g. a pre-trained VQGAN [5] for DynamiCrafter, consisting of an encoder \mathcal{E} and a decoder \mathcal{D} . Conditioned on a text condition c_{text} and a reference image c_{img} , the diffusion model ϵ_Θ is then trained to predict the noise ϵ at timestep $t \in \mathcal{U}(0, T)$ using a simple reconstruction loss:

$$\min_{\theta} \mathbb{E}_{t, \mathbf{x} \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \|\epsilon - \epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t)\|_2^2. \quad (2)$$

The diffusion model itself is typically implemented as a UNet, e.g. a 3D-UNet [34] in DynamiCrafter, where θ denotes the neural network’s parameters.

Camera Conditioning. To incorporate camera control, CamI2V employs a dense supervisory signal using pixel-wise embeddings of camera rays, represented via Plücker coordinates. Specifically, for each pixel (u, v) the Plücker coordinates $P = (o \times d', d')$ are computed using the normalized ray direction $d' = \frac{d}{\|d\|}$ and the ray origin o (the camera focal point).

The ray direction relative to a reference coordinate frame—such as the camera coordinate system of the initial frame—is derived from the intrinsics \mathbf{K} and extrinsics

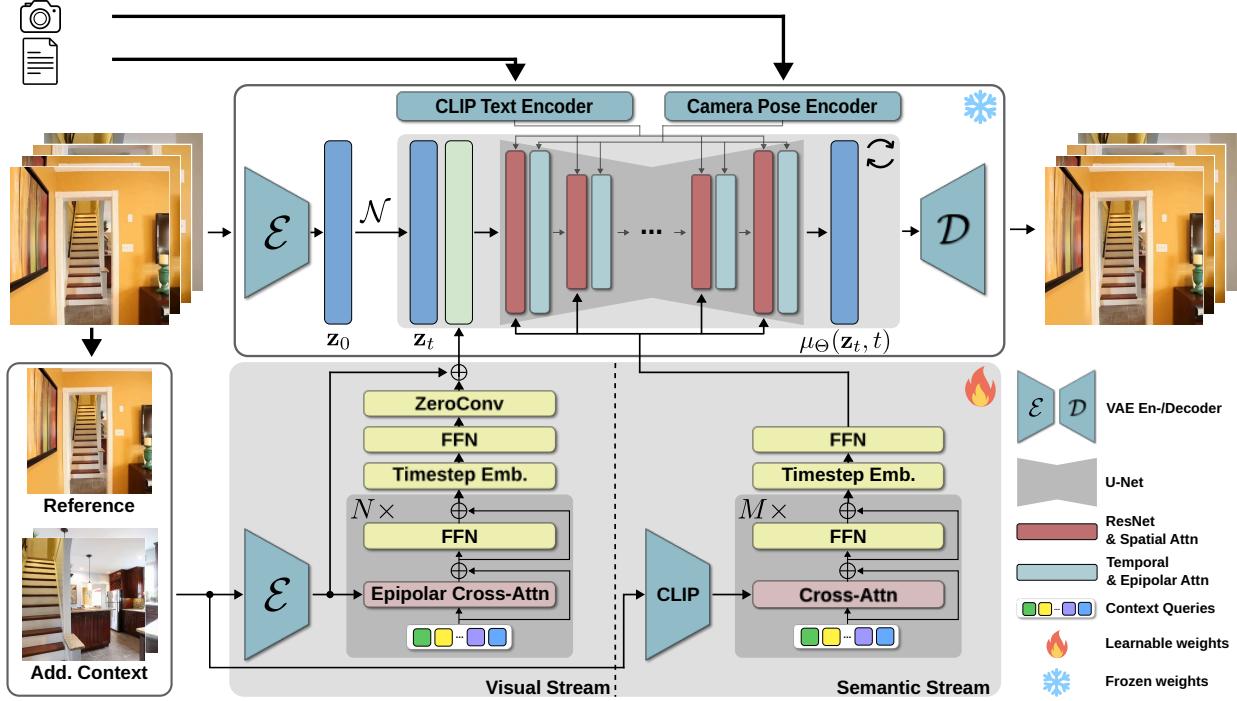


Figure 2. **CamContextI2V pipeline.** Our pipeline generates videos conditioned on a reference image, an optional text description and a camera trajectory encoded through a camera pose encoder conditioning. Additionally, frames are encoded in two parallel streams, one providing pixel-level visual cues and the other a global context. The pixel-level stream employs epipolar attention to enforce 3D consistent feature aggregation. Finally, both stream are augmented with a timestep embedding to ensure timestep-wise conditioning of the diffusion process.

$E = [\mathbf{R}|\mathbf{t}]$ as:

$$d = \mathbf{RK}^{-1} + \mathbf{t}. \quad (3)$$

These embeddings are further encoded at multiple resolutions and integrated into the epipolar attention blocks inserted into the U-Net.

4. Method

Image-to-video diffusion models generate videos based on a single reference frame c_{img} and an optional text condition c_{txt} . Additionally, camera-controlled diffusion models are conditioned on a camera trajectory $[P_{cam}^0, \dots, P_{cam}^{16}]$ allowing precise control of the camera view at each timestep. This enforces generations beyond the context conveyed through the reference frame degrading the visual quality. To counteract this, we provide the model with additional context frames $c_{ctx}^0, \dots, c_{ctx}^N$ and their poses $P_{ctx}^0, \dots, P_{ctx}^N$ that span a rich context utilized in the diffusion process.

Our *Context-aware encoder*, shown in Fig. 2, extends DynamiCrafter’s *Dual-stream Image Injection* to support multiple image conditions. Natively, it conditions the model at the pixel level by concatenating reference latents z_{img} with noisy latents z_t along the channel dimension, which restricts the generations to the narrow context provided by

the reference image. Additionally, to better guide the diffusion process, semantic features aggregated from CLIP-embedded image and text conditions are integrated layer-wise through spatial cross-attention. To utilize the pre-trained generative capabilities of the diffusion model and refrain from fine-tuning large parts of the U-Net, we chose to inject our condition in those streams.

Semantic Stream. We adopt DynamiCrafter’s query transformer \mathcal{E}_{sem} to integrate cross-modal information from the CLIP-embedded reference image \mathbf{F}_{img} , the text condition \mathbf{F}_{txt} , and additional context frames $\mathbf{F}_{ctx} = [F_{ctx}^1, \dots, F_{ctx}^N]$. Specifically, \mathcal{E}_{sem} employs learnable latent query tokens \mathbf{T}_{sem} to gather context across multiple layers of cross-attention and feed-forward networks, yielding a global representation:

$$\mathbf{F}_{sem} = \mathcal{E}_{sem}([\mathbf{F}_{img}, \mathbf{F}_{txt}, \mathbf{F}_{ctx}], \mathbf{T}_{sem}). \quad (4)$$

To preserve strong cross-modal context aggregation, we initialize \mathcal{E}_{sem} from DynamiCrafter’s *Dual-stream Image Injection* module and fine-tune it to handle multiple image conditions.

Visual Stream. While the semantic stream provides a well-suited global context representation, it lacks fine-grained visual details due to CLIP’s inherent training on visual-language alignment, which favors high-level representations of single entities.

To enhance context-aware generation, we integrate our visual condition directly into DynamiCrafter’s image conditioning. Specifically, we embed the context frames $c_{ctx}^0, \dots, c_{ctx}^N$ into the latent space $\mathbf{Z}_{ctx} = [z_{ctx}^0, \dots, z_{ctx}^N]$ and introduce pixel-wise learnable context tokens $\mathbf{T}_{vis} \in \mathbb{R}^{T \times h \times w \times D}$. The context tokens serve as queries in a query transformer, similar to the semantic stream, to aggregate timestep- and pixel-wise features from the latent context frames.

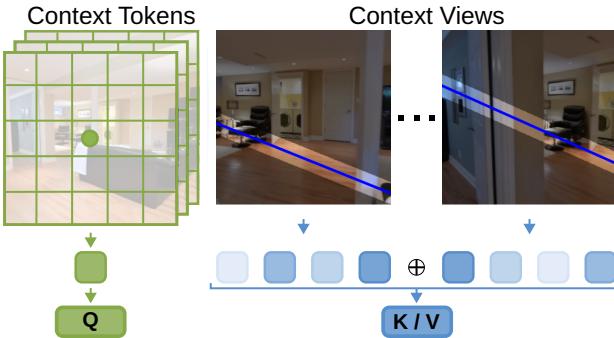


Figure 3. **Epipolar cross-attention.** Learnable context tokens act as queries to retrieve pixel-level features for each timestep from context views, masked according to epipolar lines to incorporate 3D geometric constraints.

3D Awareness. To introduce 3D awareness, we employ an epipolar cross-attention mechanism which guides the feature aggregation to only consider potentially relevant features. Specifically, each token $t_i \in \mathbf{T}_{vis}$, illustrated in Fig. 3, describes a pixel (u, v) at timestep t . Employing the provided camera pose P_{cam}^t at the given timestep, we can compute the epipolar line $l_{ij} = Ax + Bx + C$ in each context view c_{ctx}^j . Using the point-to-line distance:

$$d(u', v') = \frac{[A, B, C]^\top \cdot [u', v', 1]}{\sqrt{A^2 + B^2}}, \quad (5)$$

we produce the epipolar mask $m \in \mathbb{R}^{Thw \times Nhw}$ masking out pixels (u', v') with a distance larger than a threshold δ , set to half of the diagonal of the latent feature space, in the cross-attention mechanism:

$$\text{EpiCrossAttn}(q, k, v, m) = \text{softmax}\left(\frac{qk^\top}{\sqrt{d}} \odot m\right)v, \quad (6)$$

where $q \in \mathbb{R}^{Thw \times D}$ describes the learnable context queries and $k, v \in \mathbb{R}^{Nhw \times D}$ the latent embedded context frames.

Temporal Awareness. The native pixel-level embedding of DynamiCrafter is agnostic to the timestep within the video as each timestep is provided with the same condition. Thus, to further enforce the diffusion model to attend to context provided at specific timesteps, we found it advantageous to employ a sinusoidal timestep embedding. In practice, we concatenate the timestep embedding to our context embeddings before forwarding it through a feed-forward network.

Finally, the visual stream of our *context-aware encoder* maps a spatially distributed embedding represented through the latent embedding of posed views to a timestep-wise embedding:

$$\mathbf{F}_{vis} = \mathcal{E}_{vis}(\mathbf{Z}_{ctx}, \mathbf{T}_{vis}, m). \quad (7)$$

To retain the reference image as a strong anchor to the generation and smoothly insert the new condition, we employ a 3D zero-convolution which weighs the usage of DynamiCrafter’s native condition z_{ref} and ours \mathbf{F}_{vis} before adding them together.

5. Experiments

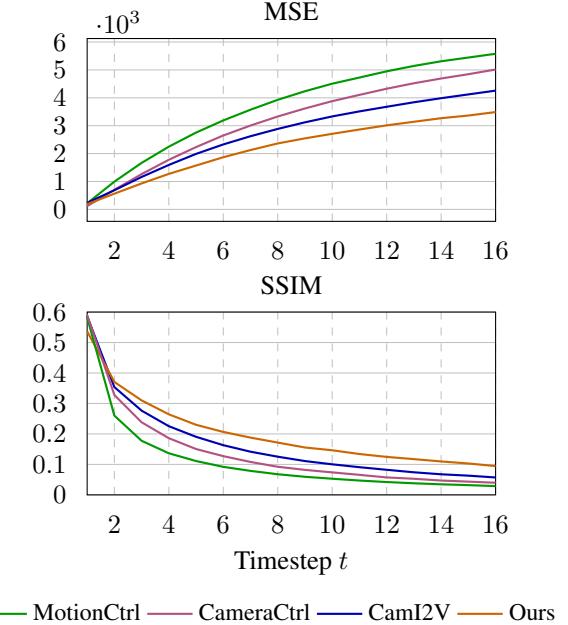


Figure 4. **Frame-wise quantitative comparison.** We compare the quality of each frame depending on the timestep in the video against state-of-the-art methods in terms of MSE and SSIM. While the reference frame provides sufficient visual cues for the initial frames, the visual quality is degrading logarithmically as time progresses and the diffusion model is forced to generate scenes beyond the provided context.

In the following, we thoroughly analyze our method in the setup described in Sec. 5.1. Sec. 5.2 and Sec. 5.3 com-

Method	FVD ↓			TransErr ↓	RotErr ↓	CamMC ↓
	VideoGPT	StyleGAN	MSE ↓			
MotionCtrl	78.30	64.47	3654.54	2.89	2.04	4.34
CameraCtrl	71.22	58.05	3130.63	2.54	1.84	3.85
CamI2V	<u>71.01</u>	<u>57.90</u>	<u>2692.84</u>	<u>1.79</u>	<u>1.16</u>	<u>2.58</u>
Ours	53.90	45.36	2579.96	1.53	1.09	2.29

Table 1. **Quantitative comparison.** We compare our method against state-of-the-art camera-controlled diffusion models. The context-awareness of our model improves the visual quality by 24.09% in terms of FVD and MSE over the baseline methods. Moreover, the model follows the camera trajectory more precisely achieving improvements in RotErr, TransErr and CamMC. The results were obtain using 25 DDIM steps with CFG set to 7.5, except for our method performing best with CFG set to 3.5

pares our method to baseline camera-controlled methods and argue the effectiveness of the provided context. Finally, in Sec. 5.4, we investigate our design choices and motivate the importance of 3D and temporal awareness in our method and the complementary effect of our two-stream design. Additionally, we demonstrate the impact of different sampling strategies on our method.

5.1. Setup

Dataset The RealEstate10K [33] comprises approximately 70K video clips at 720p of static scenes depicting indoor and outdoor house tours. The clips are annotated with camera extrinsic and intrinsic values obtained through the ORB-SLAM2 [12] pipeline. Additionally, we use the captions provided by the authors of CameraCtrl [7]. The video clips are then center-cropped to a size of 256×256 and clipped to short frames of length 16 with a stride sampled between 1 and 10.

Metrics We evaluate our method with respect to generative quality, the faithfulness to the provided context and the camera trajectory. Firstly, to ensure improved visual quality we report the Frechet Video Distance (FVD) [16]. To ensure the faithfulness with respect to the additional context, we evaluate the pixel-wise mean squared error (MSE) and the Structural Similarity Index (SSIM) [21] independently for each timestep.

Finally, to examine the generated camera trajectory we follow the evaluation paradigm proposed by CameraCtrl and CamI2V. Using the structure-from-motion pipeline GLOMAP [14], we estimated the camera rotation \tilde{R}_i and translation \tilde{T}_i for each camera i and compute the independent rotation and translation errors, $RotErr$ and $TransErr$ respectively, as well as the combined element-wise error

CamMC:

$$RotErr = \sum_{i=1}^n \cos^{-1} \frac{\text{tr}(\tilde{R}_i R_i^T) - 1}{2}, \quad (8)$$

$$TransErr = \sum_{i=1}^n \|\tilde{T}_i - T_i\|_2, \quad (9)$$

$$CamMC = \sum_{i=1}^n \|[\tilde{R}_i | \tilde{T}_i] - R_i | T_i]\|_2. \quad (10)$$

To counteract the randomness of GLOMAP, we follow the scheme to average the metrics over successful runs out of 5 trials. All metrics are computed on a subset consisting of videos extending over a duration of over 30 seconds to ensure sufficient additional context to be sampled from and prevent avoid sampling to close to the 16 frame clip.

Implementation Details Our model builds on DynamicCrafter with inserted camera control from CamI2V. Training is initialized from CamI2V checkpoints at 50K iterations, freezing all parameters except for our *Context-aware Encoder*. We train for 50K iterations at a resolution of 256×256 , using the Adam optimizer with a fixed learning rate of 1×10^{-4} and a batch size of 64. Following the baseline, we chose LIGHTNING as our training framework with mixed-precision using DeepSpeed ZeRo-1 on 4 NVIDIA A100 GPUs for approximately 7 days. For comparison to the baseline methods we use the re-implementations of MotionCtrl [22] and CameraCtrl [7] provided by the authors of CamI2V. We sample 1-4 context frames uniformly from the complete videos during training. For fairness, during evaluation we only sample from outside the 16 frames window to be generated.

5.2. Quantitative Comparison

To show the effectiveness of the additional context provided by our method, we compare against several camera-controlled methods, namely MotionCtrl [22], CameraCtrl [7] and CamI2V [32]. Tab. 1 presents the comparison of our method against the baseline methods. Our model achieves an improvement of 24.09% in terms of the FVD

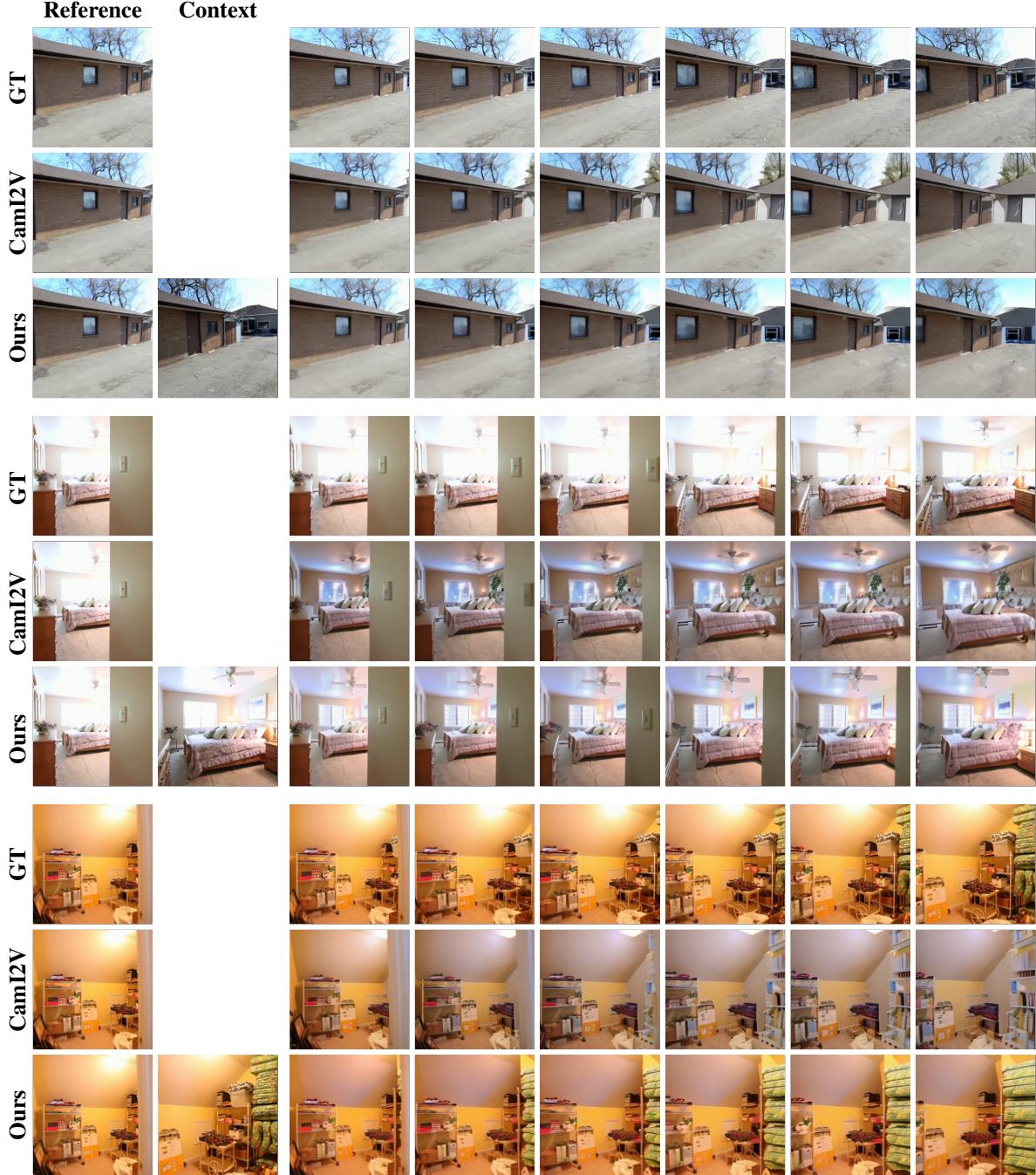


Figure 5. **Qualitative comparison.** The reference frame only provides an insufficient context, leading to visually degrading video quality and 3D consistency in the baseline. Our method, provided with an additional context frame, is able to faithfully represent the provided scene, improving visual quality beyond the reference frame. Zoom in for more details.

score highlighting the effectiveness of added context for video generation.

To further evaluate the context-awareness of our method, we report the MSE in Fig. 4 between the generated videos and the ground-truth videos on a per-frame basis to assess the improvement especially for later frames that typically

lack sufficient context from the reference frame. Additionally, to assess the visual quality of each frame, we compute the SSIM metric on each frame.

It is visible that the visual quality degrades logarithmically with the video length as the diffusion model lacks sufficient context. Our method outperforms the baseline meth-

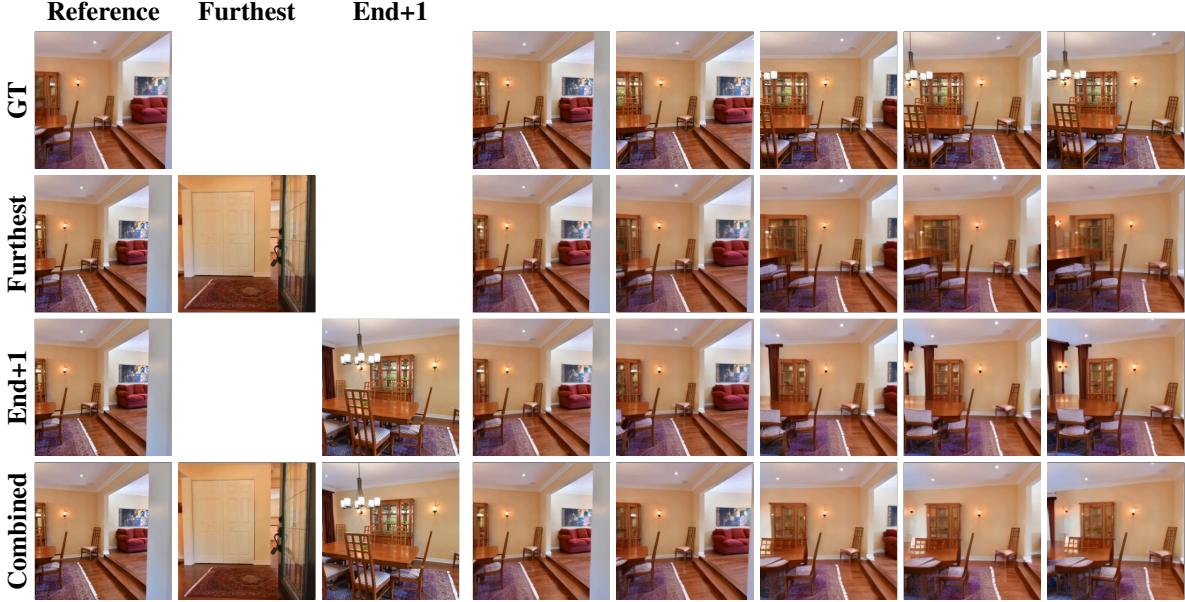


Figure 6. **Qualitative results of different sampling strategies.** We generate samples conditioned on the furthest frame providing minimal context and the frame immediately following the video providing maximal context. Our method is able to reject unrelated features from the *furthest* frame and only aggregate features from the *end + 1* frame providing additional information to the diffusion process.

ods in both MSE and SSIM, especially for later frames. This shows that providing the diffusion process with additional context can stabilize the generative quality over time.

Additionally, we investigate the accuracy of the generated camera trajectory with respect to the RotErr, TransErr and CamMC. We observe a slightly improved rotational error compared to CamI2V’s, indicating an improved camera trajectory of our method. As the evaluation pipeline, GLOMAP, used for estimating the camera trajectory matches keypoint features to simultaneously estimate the camera trajectory and reconstruct a 3D scene using bundle adjustment and we do not train the camera encoder, nor the diffusion model itself, this improved camera trajectory is mainly linked to an improved 3D consistency and visual quality of the generated scene. This demonstrates that the additionally provided context enforces more faithful representation of the 3D scene.

5.3. Qualitative Comparison

Fig. 5 shows different samples from our method compared against CamI2V. It is evident that the reference frame does not provide sufficient context for the generation past the first few frames. This results in visually degrading image quality and unrealistic generations of the baseline method.

In contrast, our method is provided with an additional context frame sampled from a later timestep past the 16 window frame that shows entities outside of the field of view of the reference frame or obstructed by obstacles. Our method is able to comprehend the position of these enti-

ties in space and effectively embed it into the timestep-wise embedding resulting in these objects being placed at correct locations in later frames. Moreover, it is visible, while the baseline method produces artifacts not visible in its condition, the extended context provides an additional constraint preventing unwanted artifacts.

5.4. Ablation Studies

To thoroughly evaluate the impact of our design choices, we conducted several ablation studies. The results are summarized in Tab. 2.

Semantic and visual stream. First, we examined the individual contributions of the semantic and visual streams to the diffusion process. We trained two model variants, each utilizing only one stream to inject additional context. Despite both variants being provided with an extended context, neither improved upon the baseline results. This limited improvement likely stems from DynamiCrafter being originally trained under matching conditions. In contrast, combining both semantic and visual streams significantly enhanced performance, highlighting their complementary interaction.

3D awareness. Next, we evaluated the effectiveness of our method’s 3D awareness, achieved through the epipolar cross-attention mechanism. Replacing epipolar cross-attention with standard (vanilla) cross-attention, allowing unrestricted feature aggregation from all tokens, still

Pixel	Multi-Cond. Sem.	Epipolar	Time	FVD ↓		MSE ↓		
				VideoGPT	StyleGAN	Total	t=2	t=16
✓		✓	✓	76.00	63.40	2622.32	632.94	4141.67
	✓	✓	✓	70.44	59.56	2810.75	862.84	4225.31
✓	✓		✓	61.61	52.04	2678.45	782.86	4102.77
✓	✓	✓		58.15	47.73	2642.69	753.36	4014.67
✓	✓	✓	✓	53.90	45.36	2579.96	668.60	4076.78

Table 2. **Ablation studies.** We compare our design choices in different studies showing that our two-stream design complementarily embeds the context and guides the diffusion process. Moreover, the 3D and temporal awareness induced into the embedding by our method is beneficial for an effective context-aware conditioning.

yielded a considerable improvement of 9.5 FVD points over the baseline. This model variant, still, demonstrates a significant improvement on the baseline by 9.5 points in the FVD score but fails to match the performance of the 3D-aware model variant. This can be contributed to the model still leveraging the additional context for the generation but failing to reject features from invalid positions, as seen in Fig. 6, especially when context frames provide minimal additional information due to them being sampled from distant regions.

Temporal awareness. Further, we assess the effect of temporal embeddings integrated into semantic and visual streams. Removing temporal embeddings results in a performance decline, although still outperforming CamI2V considerably. The temporal embeddings, particularly within the visual stream, explicitly guide the temporal attention of the U-Net to properly interpret timestep-specific context. Without this guidance, the epipolar cross-attention timestep-wise embedded context may be interpreted freely, resulting in a impaired performance.

Sample Range	FVD (VideoGPT)	MSE
(end, -1]	45.63	2579.96
end + 1	44.21	2474.28
Furthest	48.52	2668.91

Table 3. **Condition Sampling Study.** To investigate the impact of different context views, we condition our method using different context sampling strategies. *(end, -1]* represents the sampling strategy used through our evaluations, while *end+1* provides context with the maximal amount of information and *furthest* with the minimal amount of information.

Context sampling. Lastly, Tab. 3 compares different sampling strategies for additional context views. Our default method samples context frames from the interval $(\text{end}, -1]$ following the generated video. Further, we investigate two extremes: First, sampling a completely unrelated frame, the *furthest* frame, as can be seen in Fig. 6.

Our results show that this only slightly degrades the visual quality, indicating that our method effectively rejects unrelated features through the induced 3D awareness of the epipolar cross-attention. Second, sampling a frame directly following the video, providing a maximal amount of information to the diffusion process. This only slightly improves our method, showing that it can effectively gather context from loosely placed context views. The qualitative results in Fig. 6 show that our *context-aware encoder* effectively sorts out unrelated information and provides the diffusion process only with the necessary context.

6. Conclusion

This paper introduces CamContextI2V, a novel conditioning mechanism that provides the diffusion process with extensive contextual information derived from multiple context views. Unlike conventional image-to-video diffusion models, which typically rely on a single reference image, our proposed method employs a *context-aware encoder* that encodes additional context through a high-level semantic stream and a 3D-aware visual stream, generating a global semantic representation and a dense, pixel-wise visual embedding from context views. This encoder effectively aggregates relevant features complementing the generation while filtering out unrelated information. As a result, our approach significantly improves visual quality and enhances the accuracy of the generated camera trajectories compared to existing baseline methods.

References

- [1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 2
- [2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 2
- [3] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu,

- Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 2023. [1](#) [2](#)
- [4] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction, 2023. [2](#)
- [5] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2021. [2](#)
- [6] 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, 2024. [2](#)
- [7] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation, 2024. [1](#), [2](#), [5](#)
- [8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. [2](#)
- [9] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. [2](#)
- [10] Gongye Liu, Menghan Xia, Yong Zhang, Haoxin Chen, Jinbo Xing, Yibo Wang, Xintao Wang, Yujiu Yang, and Ying Shan. Stylecrafter: Enhancing stylized text-to-video generation with style adapter, 2024. [1](#)
- [11] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videostudio: Generating consistent-content and multi-scene videos, 2024. [2](#)
- [12] Raul Mur-Artal and Juan D. Tardos. Orb-slam2: An open-source slam system for monocular, stereo, and rgbd cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017. [5](#)
- [13] Gyeongrok Oh, Jaehwan Jeong, Sieun Kim, Wonmin Byeon, Jinkyu Kim, Sungwoong Kim, and Sangpil Kim. Mevg: Multi-event video generation with text-to-video models, 2024. [2](#)
- [14] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L. Schönberger. Global structure-from-motion revisited, 2024. [5](#)
- [15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. [2](#)
- [16] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges, 2019. [5](#)
- [17] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising, 2023. [2](#)
- [18] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation, 2024. [1](#)
- [19] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingen Zhou. Videocomposer: Compositional video synthesis with motion controllability, 2023. [1](#)
- [20] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. Lavie: High-quality video generation with cascaded latent diffusion models, 2023. [2](#)
- [21] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. [5](#)
- [22] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation, 2024. [1](#), [2](#), [5](#)
- [23] Daniel Watson, Saurabh Saxena, Lala Li, Andrea Tagliasacchi, and David J. Fleet. Controlling space and time with diffusion models, 2024. [2](#)
- [24] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors, 2023. [2](#)
- [25] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation, 2024. [1](#), [2](#)
- [26] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model, 2023. [1](#)
- [27] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24*, page 1–12. ACM, 2024. [2](#)
- [28] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazhen Xu, Yuanming Yang, Wenqi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihai Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2024. [1](#), [2](#)
- [29] Zixuan Ye, Huijuan Huang, Xintao Wang, Pengfei Wan, Di Zhang, and Wenhan Luo. Stylemaster: Stylize your video with artistic generation and translation, 2024. [1](#)
- [30] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. [2](#)
- [31] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingen Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models, 2023. [2](#)
- [32] Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model, 2024. [1](#), [2](#), [5](#)
- [33] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images, 2018. [5](#)

- [34] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation, 2016. [2](#)

CamContextI2V: Context-aware Controllable Video Generation

Supplementary Material

7. Additional Qualitative Results

We provide additional qualitative results of our method in Fig. 7.

8. Camera Evaluation

To ensure comparability with the evaluation paradigm proposed by CamI2V, we follow their configuration of the GLOMAP pipeline. The adjusted configuration parameters are listed in Tab. 4. The remaining parameters are set to the default parameters as provided by GLOMAP.

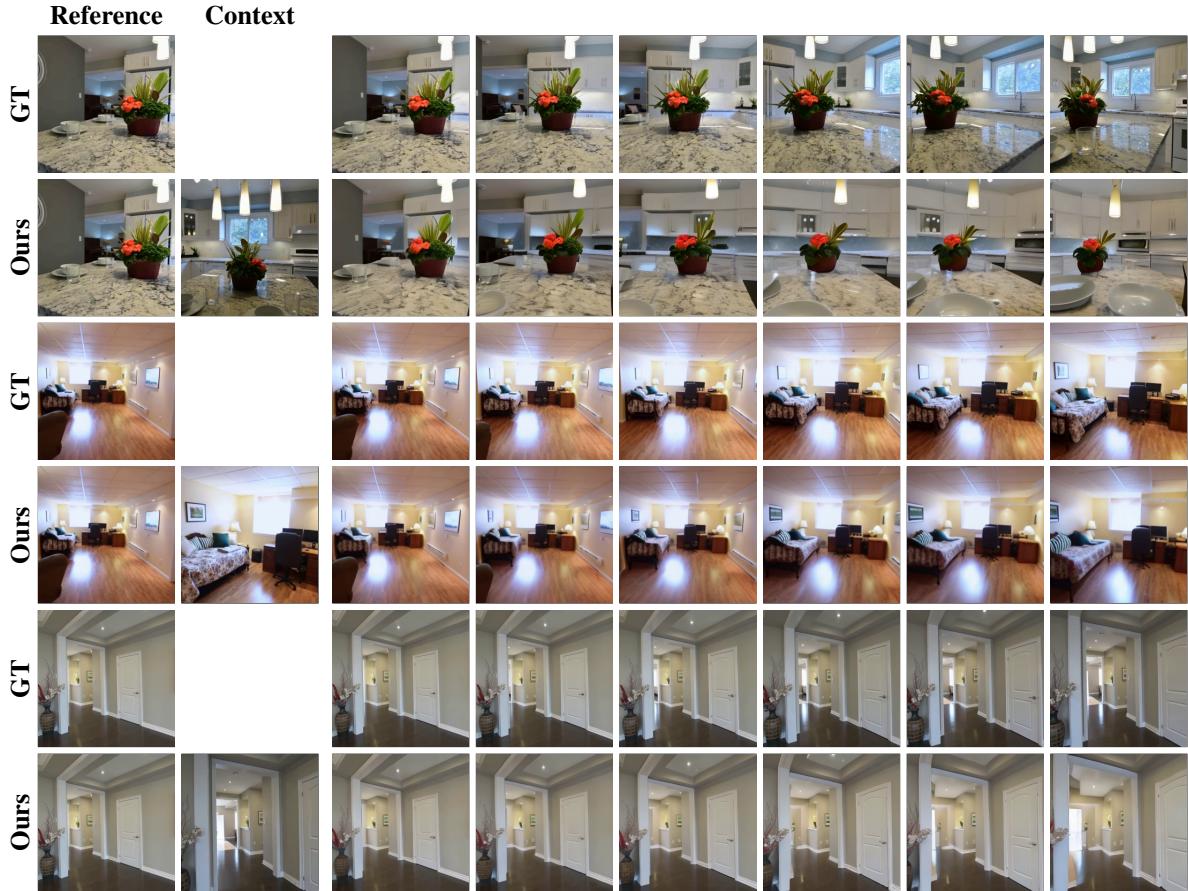


Figure 7. **Additional Qualitative results.** Zoom in for more details

Stage	Parameter	Value
Feature Extraction	ImageReader.single_camera	1
	ImageReader.camera_model	SIMPLE_PINHOLE
	ImageReader.camera_params	$\{f\}, \{cx\}, \{cy\}$
	SiftExtraction.estimate_affine_shape	1
	SiftExtraction.domain_size_pooling	1
Sequential Matcher	SiftMatching.guided_matching	1
	SiftMatching.max_num_matches	65536
Mapper	RelPoseEstimation.max_epipolar_error	4
	BundleAdjustment.optimize_intrinsics	0

Table 4. Changed parameters of the Glomap pipeline used in our evaluation.