CamC2V: Context-aware Controllable Video Generation

Luis Denninger[1] Sina Mokhtarzadeh Azar[1,2] Juergen Gall[1,2]
[1]University of Bonn [2]Lamarr Institute for Machine Learning and Artificial Intelligence
l_denninger@uni-bonn.de mokhtarzadeh@iai.uni-bonn.de gall@iai.uni-bonn.de

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 degrade visual quality, limiting their applicability for tasks requiring faithful scene representation. We propose CamC2V, a context-to-video (C2V) model that integrates multiple image conditions as context 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 a 24.09% *(FVD) improvement in visual quality and camera controllability. Our code is publicly available at: https://github.com/LDenninger/CamC2V.*

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 [21, 28], style [12, 32], motion [22, 25] or camera trajectories [7, 27, 36] have been a widely studied topic.

While text-to-video (T2V) diffusion models like VideoCrafter [3] or CogVideoX [31] 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 creation.

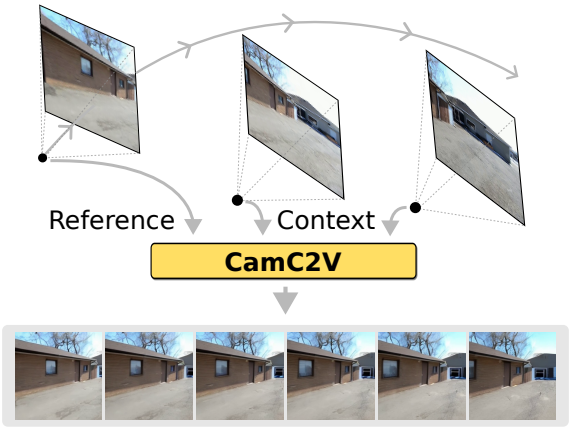

Figure 1. CamC2V performs context-aware generation provided a reference frame representing the initial frame c_{img} and [1-4] additional views c^i_{ctx} providing crucial context to the diffusion process missing in the reference frame.

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 CamC2V, 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 contributions are as follows.

• We propose CamC2V, 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 the RealEstate10K dataset.

2. Related Works

Diffusion-based Video Generation. Originally developed for image generation [8, 17], diffusion models have since demonstrated great success synthesizing high-quality videos [2, 9]. Models such as SVD [2], LAVIE [23] 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 [26], Seine [4] or I2vgen-XL [35] 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 [25], AnimateDiff [6] or Direct-a-Video [30] model camera movements through camera-motion primitives, recent approaches such as CameraCtrl [7], CamCo [27] or CamI2V [36] directly insert the camera poses showcasing fine-grained camera control. A key is the dense supervisory signal provided by pixel-wise camera rays represented as Plücker coordinates, which are encoded and inserted into the diffusion model in a ControlNet-like fashion [34].

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, 26, 31, 35]. Models like Gen-L-Video [20], MEVG [15] or VideoStudio [13] 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 the noise between them either in a divide-and-conquer or auto-regressive setup to generate long consistent videos.

Recently approaches like ReCamMaster [1] or TrajectoryCrafter [33] focus on conditioning the generative process on complete videos to recreate the video from another camera trajectory. While these methods effectively leverage the context of multiple images, their approaches heavily rely on the one-to-one mapping from condition and target frames. In contrast, our method only relies on loosely placed images that do not explicitely correspond to a timestamp in the target sequence.

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 [26], 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:

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

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:

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

The diffusion model itself is typically implemented as a UNet, *e.g.* a 3D-Unet [38] 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

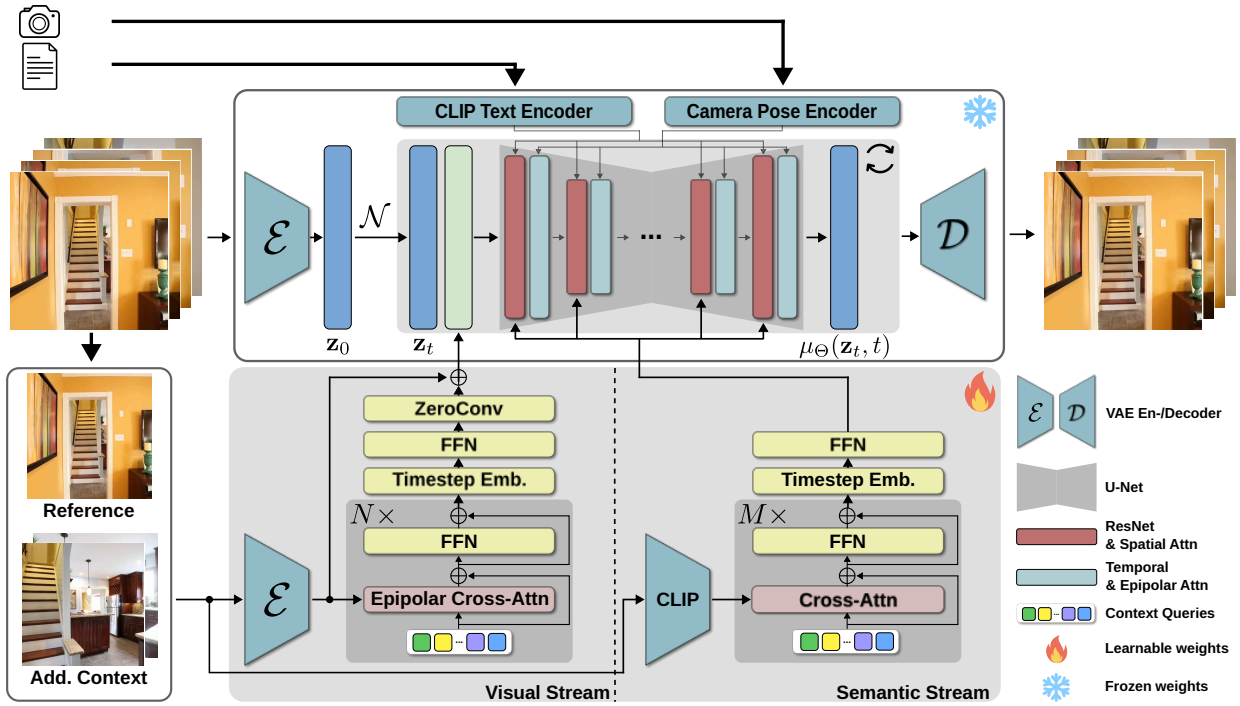

Figure 2. **CamC2V pipeline.** Our pipeline generates videos conditioned on a reference image, an optional text description and a camera trajectory encoded through a camera pose encoder conditioning. Additionally, the *Context-aware Encoder* processes frames in two parallel streams, one providing pixel-level 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.

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

$$d = \mathbf{R}\mathbf{K}^{-1} + \mathbf{t}. \tag{3}$$

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 $[P^0_{cam}, \ldots, P^T_{cam}]$ allowing precise control of the camera view at each timestep. The reference frame does not always provide the necessary context corresponding to the camera trajectory. This can lead to insufficient visual quality of the generated frames. In contrast, we propose a new scheme coined context-to-video which enhances the generation process with a rich context conveyed through additional context frames $c^0_{ctx}, \ldots, c^N_{ctx}$ and their poses $P^0_{ctx}, \ldots, P^N_{ctx}$

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 $\mathbf{F}_{ctx} = [F^0_{ctx}, \ldots, F^N_{ctx}]$. 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, yield-

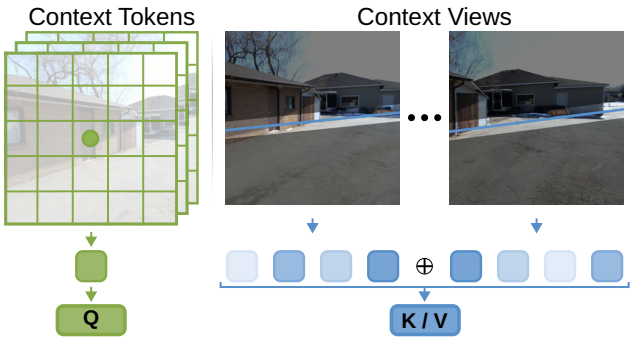

Figure 3. **Epipolar cross-attention.** Learnable context tokens act as queries to retrieve pixel-level features for each timestep from context views, masked according to epipolar lines to incorporate 3D geometric constraints.

ing 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 $c^0_{ctx}, \dots, c^N_{ctx}$ into the latent space $\mathbf{Z}_{ctx} = [z^0_{ctx}, \dots, z^N_{ctx}]$ 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.

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^t_{cam} at the given timestep, we can compute the epipolar line $l_{ij} = Ax + Bx + C$ in each context view c^j_{ctx}. Using the point-to-line distance:

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

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:

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

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.

Log-weighted Loss. Our *Context-aware Encoder* injects crucial information, especially to later frames, that have limited context in the baseline methods leading to degrading visual quality. To force the training process to focus on such frames that rely most heavily on our context-aware conditioning, we apply a logarithmic re-weighting to the standard reconstruction loss along the time axis. Specifically, for each frame k in the sequence, we define:

$$\mathcal{L} = \frac{\sum_{k=0}^{15} \log_{10}(k+1) \cdot ||\varepsilon_k - \varepsilon_{\theta,k}(\mathbf{x_t}, \mathbf{c}, t)||_2^2}{\sum_{k=0}^{15} \log_{10}(k+1)}. \quad (8)$$

This not only improves generative quality but also stabilizes the training, mitigating divergence in later stages of training.

5. Experiments

5.1. Setup

Dataset The RealEstate10K [37] 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 [14] pipeline. Additionally, we use the

Method	FVD ↓		MSE ↓	TransErr ↓	RotErr ↓	CamMC ↓
	VideoGPT	StyleGAN				
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	71.01	57.90	2692.84	1.79	1.16	2.58
Ours	**53.90**	**45.36**	**2579.96**	**1.53**	**1.09**	**2.29**

Table 1. **Quantitative comparison.** Compared against state-of-the-art camera-controlled diffusion models, our method achieves an improved video fidelity of 24.09% in terms of FVD 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

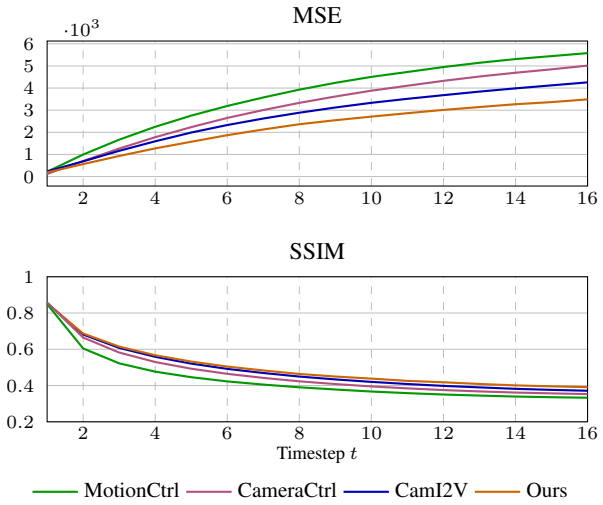

Figure 4. **Frame-wise quantitative comparison.** We compare the per-timestep MSE and SSIM against state-of-the-art methods. Due to the insufficient context provided by the reference frame, the visual quality degrades logarithmically as time progresses.

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 faithfullness to the provided context and the camera trajectory. Firstly, to ensure improved video fidelity we report the Frechet Video Distance (FVD) [19] using the evaluation protocols from VideoGPT [29] and StyleGAN [10]. 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) [24] independently for each timestep.

Finally, to examine the generated camera trajectory we follow the evaluation paradigm proposed by CameraCtrl and CamI2V. Using GLOMAP [16], we estimate the camera rotation $\tilde{R}_i$ and translation $\tilde{T}_i$ for each camera i over 5 trials and compute the independent rotation and translation errors, *RotErr* and *TransErr* respectively, as well as the combined element-wise error *CamMC*:

$$\text{RotErr} = \sum_{i=1}^{n} \cos^{-1} \frac{\text{tr}(\tilde{R}_i R_i^T) - 1}{2}, \tag{9}$$

$$\text{TransErr} = \sum_{i=1}^{n} ||\tilde{T}_i - T_i||_2, \tag{10}$$

$$\text{CamMC} = \sum_{i=1}^{n} ||[\tilde{R}_i|\tilde{T}_i] - R_i|T_i]||_2. \tag{11}$$

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 avoid sampling to close to the 16 frame clip.

Implementation Details Initialized from CamI2V checkpoints, 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. Using LIGHTNING as our training framework with mixed-precision using DeepSpeed ZeRo-1 on 4 NVIDIA A100 GPUs, training takes approximately 7 days. For comparison, we use the re-implementations of MotionCtrl [25] and CameraCtrl [7] provided by the authors of CamI2V. We sample 1-4 context frames uniformly from the complete videos during training.

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 [25], CameraCtrl [7] and CamI2V [36]. 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 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.

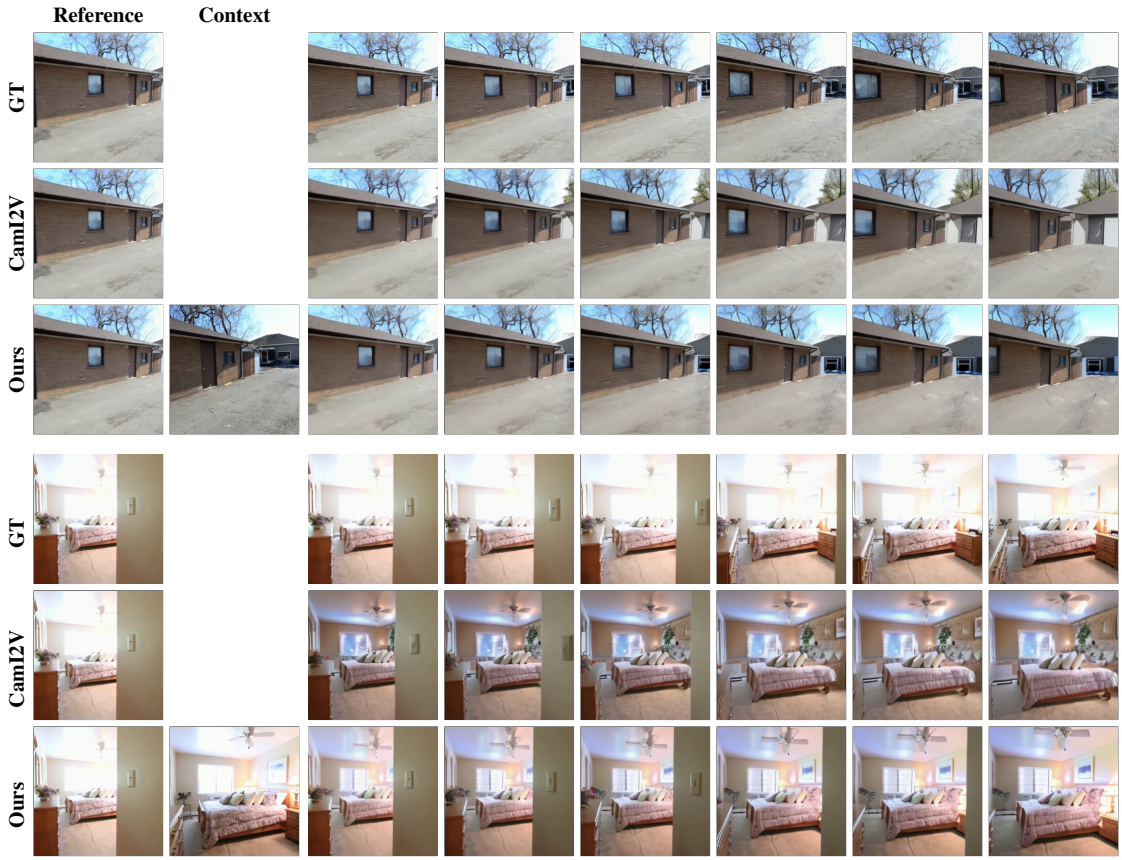

Figure 5. **Qualitative comparison**. Our method, provided with an additional context frame, overcomes the limited context of a single reference frame, improving visual quality beyond the reference frame. Zoom in for more details.

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 methods 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 are 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 entities 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.

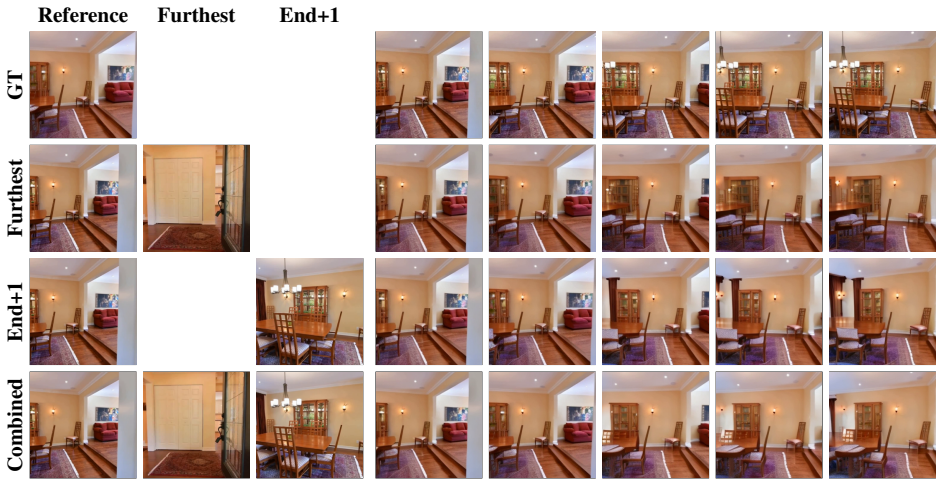

Figure 6. **Qualitative results of different sampling strategies.** We generate samples conditioned on the furthest frame providing minimal context and the frame immediately following the video providing maximal context. Our method is able to reject unrelated features from the *furthest* frame and only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

Multi-Cond.		Epipolar	Time	FVD ↓		MSE ↓		
Pixel	Sem.			VideoGPT	StyleGAN	Total	t=2	t=16
				71.01	57.90	2792.84	758.38	4101.71
✓		✓	✓	76.00	63.40	2622.32	**632.94**	4141.67
	✓	✓	✓	70.44	59.56	2810.75	862.84	4225.31
✓	✓			63.817	54.13	2701.28	791.49	4127.12
✓	✓		✓	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. Adding epipolar attention and temporal embeddings to the *Context-aware Encoder* equips it with explicit 3D and temporal awareness, improving context retrieval and further boosting performance.

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 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 attributed 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 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 (end, -1] following the generated video. Furthermore, we investigate two extremes: first, sampling a completely unrelated frame, the *furthest* frame, as shown 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.

5.5. Comparison to Novel View Synthesis Approach

We finally compare our method against FrugalNeRF [11], a state-of-the-art novel view synthesis approach for sparse views. Such NeRF-based methods typically rely on an initial sparse 3D reconstruction step, limiting their applicability to diverse scenes. In a two-view setup, the COLMAP pipeline [18], typically employed in this step, only achieves a reliable registration in over 80% of the cases if the frames distance is between $\sim 1s$ and $\sim 8.5s$, heavily limiting such approaches if the context frames are too close or too distant.

We train FrugalNeRF on $\sim$ 100 test scenes from the RealEstate10K dataset using the first and 17th frame as training frames and the intermediate ones as test frames, similar to the *End+1* evaluation setup of the context sampling study in Sec. 5.4. Accordingly, we define the reference and context frame for our approach.

Tab. 4 shows that our method slightly outperforms FrugalNeRF while being more broadly applicable, as it does not require a prior reconstruction step. Moreover, because NeRF-based methods typically employ a test-time optimization training scheme, the time needed to learn and render the 3D representation ($1265.87s$) is substantially larger than the pure rendering time for our approach ($8.01s$), which can be directly applied to novel scenes after training. See the supplementary material for additional details.

Method	SSIM ↑	LPIPS ↓	PSNR ↑	Time (s)
FrugalNeRF	0.737	0.156	23.76	1265.87
Ours	**0.741**	**0.128**	**24.32**	**8.01**

Table 4. **Novel view synthesis comparison.** We evaluate our model against a state-of-the-art sparse-view 3D reconstruction method that is geometrically grounded to faithfully represent the scene. All methods are tested at a resolution of 512×512 using the first and 17th frames as conditioning signals. Our approach achieves slightly better image quality while requiring only a fraction of the compute time.

6. Conclusion and Limitations

This paper introduces CamC2V, 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 results in significantly improved video fidelity and adherence to scene context, bringing video diffusion models closer to traditional rendering engines.

Still, our method treats the context views as an instantaneous scene snapshot which limits its applicability in highly dynamic scenes. Moreover, its generative capability is limited by the baseline diffusion model. Applying our method on more dynamic scenes and on novel DiT-based diffusion models may be basis to future work.

Acknowledgment

This work has been supported the ERC Consolidator Grant FORHUE (101044724). The authors gratefully acknowledge the granted access to the Marvin cluster hosted by the University of Bonn, as well as the Federal Ministry of Research, Technology and Space, the Ministry of Culture and Science of the State of North Rhine-Westphalia, the Ministry of Science, Research and Arts of the State of Baden-Wurttemberg, the Bavarian State Ministry of Science and the Arts and the Gauss Centre for Supercomputing e.V. (GCS) for funding this project by providing computing time on the Supercomputer JUPITER at Julich Supercomputing Centre (JSC) of Forschungszentrum Julich through the Gauss AI Compute Competition.

References

[1] 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. 2025. 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] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. 2019. 5

[11] Chin-Yang Lin, Chung-Ho Wu, Chang-Han Yeh, Shih-Han Yen, Cheng Sun, and Yu-Lun Liu. FrugalNeRF: Fast convergence for extreme few-shot novel view synthesis without learned priors. 2025. 8

[12] 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

[13] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. VideoStudio: Generating consistent-content and multi-scene videos. 2024. 2

[14] Raul Mur-Artal and Juan D. Tardos. ORB-SLAM2: An open-source slam system for monocular, stereo, and rgb-d cameras. 2017. 4

[15] 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

[16] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L. Schönberger. Global structure-from-motion revisited. 2024. 5

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. 2022. 2

[18] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. 2016. 8, 1

[19] 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

[20] 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

[21] 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

[22] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. VideoComposer: Compositional video synthesis with motion controllability. 2023. 1

[23] 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

[24] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. 2004. 5

[25] 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

[26] 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

[27] 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

[28] 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

[29] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. VideoGPT: Video generation using vq-vae and transformers. 2021. 5

[30] 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. 2024. 2

[31] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan

Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. 2025. 1, 2

[32] Zixuan Ye, Huijuan Huang, Xintao Wang, Pengfei Wan, Di Zhang, and Wenhan Luo. StyleMaster: Stylize your video with artistic generation and translation. 2024. 1

[33] Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. TrajectoryCrafter: Redirecting camera trajectory for monocular videos via diffusion models. 2025. 2

[34] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. 2023. 2

[35] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2VGen-XL: High-quality image-to-video synthesis via cascaded diffusion models. 2023. 2

[36] 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

[37] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. 2018. 4

[38] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. 2016. 2