FL4N: FEED-FORWARD LATENT MODEL FOR 4D NOVEL VIEW SYNTHESIS

Anonymous authors
Paper under double-blind review

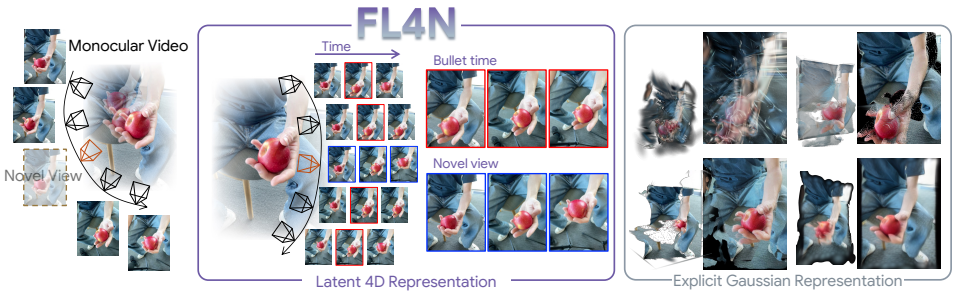

Figure 1: **Teaser.** FL4N models dynamics in a latent 4D space rather than reconstructing 3D/4D Gaussians. Gaussian methods suffer from temporal ghosting, floaters, and holes, while FL4N renders coherent frames.

ABSTRACT

4D Novel View Synthesis (NVS) from monocular videos aims to render dynamic scenes across viewpoints and timestamps, requiring jointly inferring 3D structure and temporal dynamics. Existing feed-forward methods rely on explicit 3D Gaussian representations, which are prone to temporal inconsistency and severe rendering artifacts. We present **FL4N**, the first feed-forward latent 4D NVS model. We bypass explicit 3D primitives by weaving a 4D scene prior into a unified latent representation across space and time. To overcome dynamic training data scarcity, we learn a temporal residual on top of a frozen, pretrained static latent backbone. A lightweight dynamic encoder learns to isolate true scene motion from camera parallax. Trained purely on image synthesis with binary mask guidance, FL4N achieves state-of-the-art rendering quality for 4D NVS in realtime. Our proposed latent representation successfully learns a strong 4D prior, generalizes across novel views and timestamps, and cleanly decouples view and time in the latent space. Codes will be publicly available.

1 INTRODUCTION

4D novel view synthesis (NVS) is a cornerstone of understanding and simulating our ever-changing physical world, enabling broad applications, *e. g.*, immersive graphics (Li et al., 2025), world models (Zhu et al., 2025), robotics (Kim et al., 2026b), autonomous driving (Song et al., 2025), and 4D-aware vision-language models (Cho et al., 2026). Unlike static NVS (Hong et al., 2025; Jin et al., 2025), 4D NVS requires capturing both persistent 3D structure and temporal dynamics. While slow optimization-based methods (Pumarola et al., 2021; Wang et al., 2025b) and multi-view methods (Balice et al., 2026; Xu et al., 2024b) are impractical in the wild, monocular feed-forward approaches offer a practical, real-time solution by inferring a 4D scene from a casual video.

However, monocular feed-forward 4D NVS is highly ill-posed, as it must jointly infer entangled 3D structure and object motion from monocular observations alone (Wang et al., 2025b). This requires the model to learn a strong 4D prior. To make this tractable, existing methods (Lin et al., 2026a; Xu et al., 2025; Yang et al., 2026) rely on the explicit 3D inductive bias of 3D Gaussians (Kerbl et al., 2023) with added 3D motion and dense geometry and motion supervision. However, this reliance on explicit 3D primitives can bottleneck 4D rendering quality. Inaccurate prediction of Gaussian

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

attributes across timestamps often results in rendering artifacts, *e. g.*, temporal ghosting, floaters, and holes (An et al., 2026; Kim et al., 2026c). Moreover, time-varying effects (*e. g.*, shadows and illumination changes) must be encoded in per-primitive attributes (Li et al., 2024), and regions with insufficient primitive coverage tend to be over-smoothed (Kim et al., 2026c; Li et al., 2024).

In recent static NVS literature, latent-based methods (Jin et al., 2025; Mitchel et al., 2025; Szymanowicz et al., 2026) avoid these limitations by learning view synthesis end-to-end with transformers, mapping input views directly to target views conditioned on a target camera, without an explicit 3D representation. With minimal 3D inductive biases and large-scale training, these latent models outperform 3D-Gaussian-based methods in visual fidelity (Jin et al., 2025; Kim et al., 2026a). However, directly extending the latent approach to dynamic scenes faces a fundamental challenge. Without explicit 3D priors, learning to disentangle 3D structure and temporal dynamics purely from data would require massive multi-view dynamic scene datasets (Bai et al., 2025; Van Hoorick et al., 2024), which are much scarcer than static datasets (Watson et al., 2025).

To resolve this challenge, we propose **FL4N**, a Feed-forward Latent 4D NVS model, which augments a frozen pretrained static latent NVS model with a trainable **dynamic encoder**. Our key insight is that a 4D latent model does not need to learn dynamic scene representation from scratch: pretrained static latent NVS models already possess a strong, generalizable 3D prior, allowing the dynamic encoder to exclusively focus on modeling temporal changes (*e. g.*, object motion or appearance variation). Consequently, our method can be trained with a small amount of dynamic data while fully inheriting the pretrained model's ability to synthesize coherent novel views.

Concretely, we attach the lightweight trainable dynamic encoder alongside the frozen static encoder, similar to ControlNet (Zhang et al., 2023). The dynamic encoder takes the same input views with a learnable per-view time token and cross-attends to frozen static encoder features to isolate true scene dynamics from camera-induced motion. A lightweight time module then modulates the residual latent for the queried timestamp, which is residually added to the static latent (He et al., 2016) and decoded by the frozen renderer at the target camera pose. Multi-view videos naturally disentangle viewpoint and temporal variation through same-time views and fixed views across time. Sampling unseen view–time pairs, FL4N learns this disentanglement through image synthesis with binary dynamic mask supervision, without dense geometry or motion supervision.

We evaluate our approach on synthetic and real datasets under a systematic protocol. To effectively demonstrate the performance of 4D novel view synthesis and the capability of view-time disentanglement, we separates the setting with novel view synthesis, cross-time view synthesis, and novel time synthesis. FL4N achieves state-of-the-art or competitive performance across datasets in the pose-free setting without test-time pose optimization, while optionally incorporating context camera poses further enables state-of-the-art performance across all three settings.

Our main contributions are summarized as follows:

- **First 4D latent NVS model.** We present to jointly model 3D structure and temporal variation in a latent space, rendering novel views at novel timestamps without explicit 3D representations or target-view/time observations.
- **A residual formulation for efficient dynamic adaptation.** We formulate 4D NVS as learning a temporal residual on top of a frozen pretrained static representation. This leverages the pretrained model's scene and viewpoint priors while concentrating learning on temporal variation through a lightweight branch.
- **State-of-the-art results.** We comprehensively benchmark 4D NVS across novel view, cross-time, and novel time synthesis on eight synthetic and real-world benchmarks, achieving state-of-the-art rendering quality on all three settings, with the fastest runtime.

2 Related work

Feed-forward Gaussian splatting for 4D novel view synthesis. Explicit feed-forward 4D methods typically represent scenes with 3D Gaussians extended with temporal dynamics, such as velocity (Xu et al., 2025; Yang et al., 2026) or deformation fields (Lin et al., 2026a;b). However, independently modeling motion per Gaussian often disrupts object rigidity and temporal consistency due to incoherent motion among nearby Gaussians (Kim et al., 2026c), yielding artifacts such as

temporal ghosting. Anchoring Gaussian prediction to each target timestamp (Liang et al., 2025) can sidestep the issue but is susceptible to view bias by primarily relying on nearby frames while underutilizing observation from distant frames. Query-based Gaussian representation (Kim et al., 2026c) alleviates this via dense motion supervision, yet its limited number of queries constrains rendering quality due to insufficient Gaussian coverage, particularly for large outdoor scenes. Multi-view approaches (Balice et al., 2026) can exploit observations from multiple time-synchronized static cameras, but the setup is far more restrictive than a monocular moving camera. We introduce a feed-forward latent-based method with minimal geometric inductive bias in a practical but more challenging monocular camera setup.

Latent-based novel view synthesis. Latent-based NVS models synthesize novel views with transformers conditioned on the target pose (Jiang et al., 2025; Jin et al., 2025; Kim et al., 2026a; Sajjadi et al., 2022; Szymanowicz et al., 2026) or latent pose variable (Mitchel et al., 2025), bypassing explicit 3D representations. Although RayDer (Prestel et al., 2026) trains on dynamic videos, it still targets static NVS while suppressing object motion via per-view nuisance token that is dropped at inference. Despite their strong generalization, these methods either assume static scenes or treat dynamic content as variation to be ignored rather than modeled. Toward dynamic latent representations, DyST (Seitzer et al., 2024) extends RUST (Sajjadi et al., 2023) with a dynamics estimator, yet keeps its scene latent time-invariant. Temporal variation is decoded via control codes extracted from target-pose and target-time frames, which limits explicit view and time control and risks appearance leakage from these frames. In contrast, our model constructs a time-conditioned scene latent that is rendered from arbitrary target cameras, enabling synthesis at novel views and novel timestamps.

Generative novel view synthesis. NVS becomes inherently ambiguous when target views extend into regions unseen in the input views. Deterministic methods maintain strong cross-view consistency on observed content but tend to blur unobserved regions (Prestel et al., 2026; Szymanowicz et al., 2026). Generative approaches (Bai et al., 2025; Chen et al., 2026; Ren et al., 2025; Van Hoorick et al., 2024; Yu et al., 2025) can extrapolate plausible content beyond the observed regions, yet their stochastic sampling often compromises cross-view consistency and scene faithfulness (Jin et al., 2025; Kim et al., 2026c). Although generative refinement can post-process rendered views from Gaussian representations (Kim et al., 2026c; Yang et al., 2026) or serve as a diffusion decoder for latent-based model (Szymanowicz et al., 2026), our work focuses on deterministic NVS and remains orthogonal to such generative refinement directions.

3 METHOD

We propose **FL4N**, a Feed-forward Latent-based 4D NVS model. Given an N-frame monocular video $\{\mathbf{I}_k\}_{k=1}^N$, our method synthesizes views from an arbitrary query view π_{tgt} at target timestamp t_{tgt} drawn from the observed time range. Figure 2 shows an overview of FL4N. We formulate 4D synthesis as *a residual adaptation* on top of a frozen pretrained static NVS model. Given the static encoder and the renderer from the pretrained model, we attach a lightweight dynamic encoder that focuses its capacity on isolating dynamic object motion.

In the following, we first review the latent static NVS baseline that forms a basis of our method (§ 3.1). We then introduce FL4N, describing the technical details on the dynamic encoder with target-time residual adaptation (§ 3.2). Finally, we detail the training strategy (§ 3.3) and conclude by analyzing the learned residual features (§ 3.4).

3.1 PRELIMINARIES: LATENT-BASED NOVEL VIEW SYNTHESIS

Latent-based static NVS methods (Jiang et al., 2025; Szymanowicz et al., 2026) encode multi-view images of a static scene into a latent scene representation and synthesize a novel view by decoding this latent conditioned on a target camera pose. Such a framework comprises a multi-view encoder $\mathcal{E}_{\text{st}}$ and a renderer $\mathcal{R}$. Camera parameters are optional, conditioning the encoder when available and otherwise letting the model operate pose-free.

Given N input images $\{\mathbf{I}_k\}_{k=1}^N$ and their optional corresponding cameras $\{\pi_k\}_{k=1}^N$, the static encoder produces the static scene latent $\mathbf{z}_{\text{st}}$:

$$\mathbf{z}_{\text{st}} = \mathcal{E}_{\text{st}}\big(\{\mathbf{x}_k,\ \pi_k\}_{k=1}^N\big) \in \mathbb{R}^{N\times P\times C}, \qquad \mathbf{x}_k = \mathcal{P}_{\text{st}}(\mathbf{I}_k), \tag{1}$$

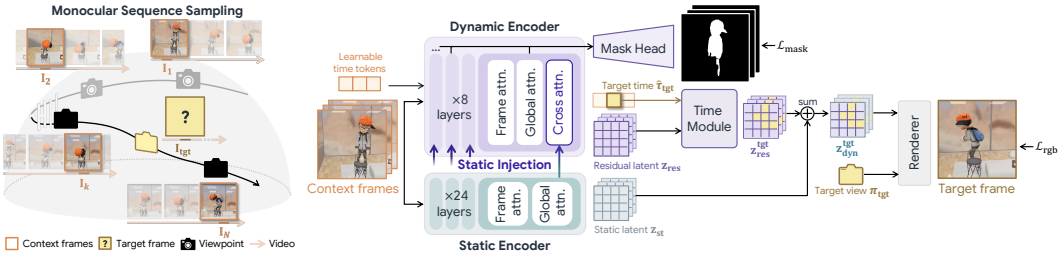

Figure 2: **FL4N Overview.** *(Left)* During training, context and target frames with varying views and timestamps are sampled from multi-view videos. *(Right)* A static encoder encodes static 3D structures (static latent) while a lightweight dynamic encoder extracts temporal residuals (residual latent) within the whole context frames. For 4D NVS, the residual latent is first adapted at the target time and added to the static latent resulting in a dynamic latent, which is then rendered into the novel view with target-view conditioned.

with $\mathcal{P}_{\text{st}}(\cdot)$ denotes the patchify layer, P the number of image patches, and C the feature dimension. The renderer decodes the latent into a novel-view image $\hat{\mathbf{I}} = \mathcal{R}(\mathbf{z}_{\text{st}}, \pi_{\text{tgt}})$ at a query camera π_{tgt}.

3.2 FL4N: Feed-forward Latent Model for 4D Novel View Synthesis

Dynamic encoder. To disentangle dynamic motion from static scene structure in monocular videos, we freeze the pretrained static encoder $\mathcal{E}_{\text{st}}$ and renderer $\mathcal{R}$, and introduce a lightweight dynamic encoder $\mathcal{E}_{\text{dyn}}$. The static encoder $\mathcal{E}_{\text{st}}$ already captures time-invariant content and camera-induced viewpoint changes from its static-scene training. During our dynamic-scene training, the frozen static encoder and renderer allow the dynamic encoder $\mathcal{E}_{\text{dyn}}$ to learn temporal variations as residuals to the pretrained static representation. Freezing the renderer $\mathcal{R}$ fixes latent-to-image mapping, forcing the residual feature to the pretrained latent space. This encourages them to complement the static representation rather than learn a new feature space from limited dynamic data.

To provide explicit temporal cues to the dynamic encoder $\mathcal{E}_{\text{dyn}}$, we introduce a learnable time token $\boldsymbol{\tau}_k \in \mathbb{R}^C$ for each context frame $\mathbf{I}_k$, which is concatenated with the patch tokens $\mathbf{x}_k = \mathcal{P}_{\text{dyn}}(\mathbf{I}_k)$ before being fed to the dynamic encoder $\mathcal{E}_{\text{dyn}}$. Through this encoding, each time token encodes a timestamp-related information from each frame.

In the dynamic encoder, identifying dynamic regions requires separating object motion from apparent camera-induced motion. We therefore condition the dynamic encoder on features from the static encoder through cross-attention. This allows it to factor out camera-induced changes captured in the static features and focus its representation on residual motion from dynamic objects. Each dynamic encoder block applies frame-wise self-attention $\text{FSA}(\cdot)$, global self-attention $\text{GSA}(\cdot)$, and cross-attention $\text{CA}(\cdot)$ sequentially,

$$\mathbf{F}^{\ell}_{\text{res}} \leftarrow \mathbf{F}^{\ell}_{\text{res}} + \text{FSA}^{\ell}(\mathbf{F}^{\ell}_{\text{res}}), \quad \mathbf{F}^{\ell}_{\text{res}} \leftarrow \mathbf{F}^{\ell}_{\text{res}} + \text{GSA}^{\ell}(\mathbf{F}^{\ell}_{\text{res}}), \quad \mathbf{F}^{\ell+1}_{\text{res}} = \mathbf{F}^{\ell}_{\text{res}} + \text{CA}^{\ell}\big(\mathbf{F}^{\ell}_{\text{res}}, \mathbf{F}_{\text{st}}\big), \quad (2)$$

where $\mathbf{F}^{\ell}_{\text{res}}$ denotes the residual features at the ℓ-th dynamic encoder block, $\mathbf{F}_{\text{st}}$ the features from the static encoder, and $\text{CA}(X, Y)$ computes queries from X and keys and values from Y.

Target-time residual adaptation. The residual latent and encoded time tokens $[\,\mathbf{z}_{\text{res}}; \hat{\boldsymbol{\tau}}_1, \ldots, \hat{\boldsymbol{\tau}}_N\,] = \mathcal{E}_{\text{dyn}}(\{\mathbf{I}_k\}_{k=1}^{N}) \in \mathbb{R}^{N\times(P+1)\times C}$ summarize temporal information across all N context frames. To extract a residual specific to a target timestamp, we condition the residual latent $\mathbf{z}_{\text{res}}$ with target-time token $\hat{\boldsymbol{\tau}}_{\text{tgt}}$. For an observed timestamp $t_{\text{tgt}} = t_k$, we directly use the corresponding token $\hat{\boldsymbol{\tau}}_k$. For an intermediate, novel timestamp $t_{\text{tgt}} \in [t_k, t_{k+1}]$, we linearly interpolate the two time tokens adjacent to the target timestamp (t_k, t_{k+1}), weighted by their temporal distance to the target.

$$\hat{\boldsymbol{\tau}}_{\text{tgt}} = \frac{(t_{k+1} - t_{\text{tgt}}) \cdot \hat{\boldsymbol{\tau}}_k + (t_{\text{tgt}} - t_k) \cdot \hat{\boldsymbol{\tau}}_{k+1}}{t_{k+1} - t_k}, \qquad t_{\text{tgt}} \in [t_k,\, t_{k+1}]. \quad (3)$$

The interpolated token then provides the temporal condition for rendering at the novel timestamp.

Given $\hat{\boldsymbol{\tau}}_{\text{tgt}}$, a lightweight time module $\mathcal{T}$ modulates the aggregated residuals $\mathbf{z}_{\text{res}}$ into the target-time residual $\mathbf{z}^{\text{tgt}}_{\text{res}}$. The time module $\mathcal{T}$ consists of two transformer blocks with adaptive layer normalization (AdaLN) conditioned on $\hat{\boldsymbol{\tau}}_{\text{tgt}}$, followed by *a residual gate* implemented as a zero-initialized

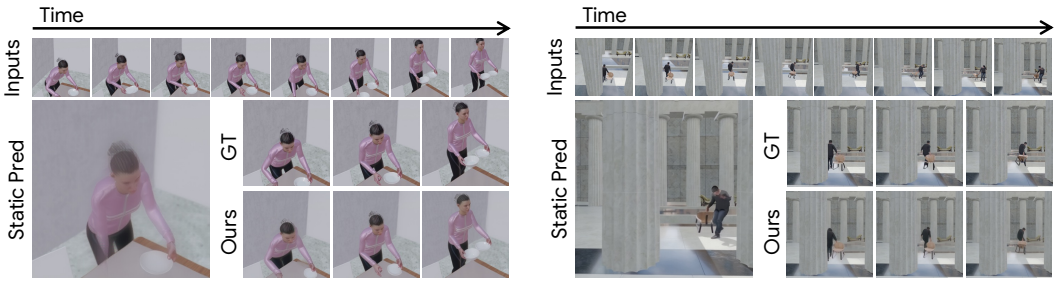

Figure 3: **Rendered frames from the static latent and our residual time correction.** In dynamic scenes, the static latent suppresses dynamic content (left) or copies it from the nearest context frame (right), whereas our residual latent places the moving content at the target time while preserving the static background.

linear projection layer. This gate initializes the residual contribution to zero (*i. e.*, preserving the pretrained static prior at the start of training) and progressively learns temporal corrections. The target residual $\mathbf{z}_{\text{res}}^{\text{tgt}}$ is then added to the static latent $\mathbf{z}_{\text{st}}$ to obtain the target-time dynamic latent $\mathbf{z}_{\text{dyn}}^{\text{tgt}}$:

$$\mathbf{z}_{\text{res}}^{\text{tgt}} = \mathcal{T}(\mathbf{z}_{\text{res}};\ \hat{\boldsymbol{\tau}}_{\text{tgt}}), \qquad \mathbf{z}_{\text{dyn}}^{\text{tgt}} = \mathbf{z}_{\text{st}} + \mathbf{z}_{\text{res}}^{\text{tgt}}. \tag{4}$$

The residual $\mathbf{z}_{\text{res}}$ is a time-invariant representation, processed only once for the N context frames by the dynamic encoder $\mathcal{E}_{\text{dyn}}$. For multiple target timestamps, we reuse $\mathbf{z}_{\text{res}}$ and apply the lightweight time module $\mathcal{T}$ with corresponding target-time tokens. The resulting target-specific latent is then fed to the frozen renderer $\mathcal{R}$ to render an image at the query camera π_{tgt}:

$$\hat{\mathbf{I}} = \mathcal{R}(\mathbf{z}_{\text{dyn}}^{\text{tgt}}, \pi_{\text{tgt}}). \tag{5}$$

3.3 Training

The rendered image $\hat{\mathbf{I}}$ at the target view and time is supervised by an image reconstruction loss:

$$\mathcal{L}_{\text{image}} = \mathcal{L}_2 + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_{\text{perc}}^{\text{dyn}}\mathcal{L}_{\text{perc}}^{\text{dyn}} + \lambda_{\text{grad}}^{\text{dyn}}\mathcal{L}_{\text{grad}}^{\text{dyn}}, \tag{6}$$

where the L2 loss $\mathcal{L}_2$ and the perceptual loss $\mathcal{L}_{\text{perc}}$ are computed over the full image, with an extra perceptual loss $\mathcal{L}_{\text{perc}}^{\text{dyn}}$ and the gradient loss $\mathcal{L}_{\text{grad}}^{\text{dyn}}$ on dynamic region. Since moving objects typically occupy a small fraction of the image, the full-image losses are dominated by static pixels and provide weak supervision for them. The dynamic-region losses provide stronger supervision for dynamic objects and yield sharper rendering of moving objects.

To help disentangle dynamic object motion from camera-induced motion without dense geometric or motion supervision, we indirectly supervise the dynamic encoder to locate dynamic regions. Specifically, we attach a DPT-based mask head (Ranftl et al., 2021) to the dynamic encoder to predict a binary dynamic-region mask $\hat{M}_k$, which is supervised with a pseudo label of dynamic mask M_k using a binary cross-entropy loss $\text{BCE}(\cdot)$,

$$\mathcal{L}_{\text{mask}} = \text{BCE}(\hat{M}_k, M_k). \tag{7}$$

The total training objective is

$$\mathcal{L} = \mathcal{L}_{\text{image}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}}. \tag{8}$$

Further details of training objective and pseudo-label generation are provided in Appendix A.

3.4 Analysis on residual latent

We conduct analyses to investigate what information is captured in the residual latent $\mathbf{z}_{\text{res}}$. Additional analysis on static and residual latent is provided in Appendix C.

Residual complements static rendering. The frozen static model can answer *where* to render, but not *when*. Having no target-time input, the static latent $\mathbf{z}_{\text{st}}$ compresses the video sequence into a single time-invariant prediction; moving content is thus discarded or frozen at a observed frames

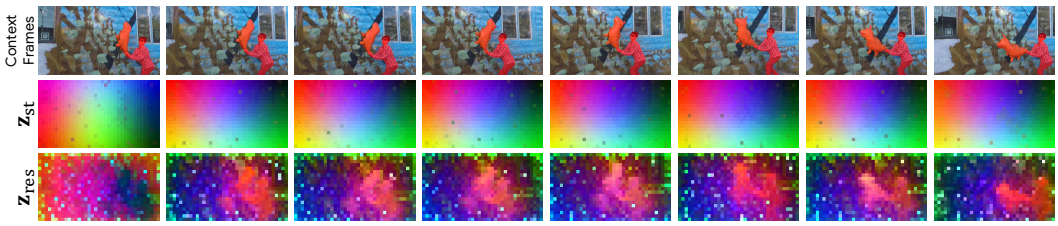

Figure 4: **PCA visualization of static and residual latent.** The static latent captures camera-ray-like structure, whereas the residual latent primarily attends to moving objects and exhibits temporal variation.

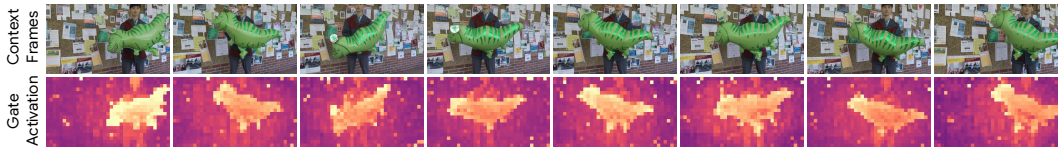

Figure 5: **Visualization of residual gate activation.** The residual gate magnitude is higher on dynamic objects while remaining suppressed in the background.

with the input pose entangled, as in Fig. 3. Our residual latent $\mathbf{z}_{\text{res}}$ supplies this missing temporal axis. Conditioned on the target time, it complements the static latent with time-dependent variation, placing the dynamic object at its target-time position while preserving the static background.

Residual feature encodes temporal dynamics. Figure 4 compares the static latent $\mathbf{z}_{\text{st}}$ and residual latent $\mathbf{z}_{\text{res}}$ using PCA visualization. While the static latent exhibits camera-ray-like patterns, the residual latent suppresses static regions and captures distinct motion across time, suggesting that the dynamic encoder effectively isolates temporal variation from the time-invariant scene representation.

This temporal specialization is further supported by the learned residual gate. Figure 5 visualizes gate activation using the channel-wise L2 norm of the gated residual. The gate remains largely inactive in static regions but activates strongly around moving objects, indicating that temporal modeling capacity is concentrated where motion occurs.

4 Experiments

Implementation details. Our model builds on a frozen static novel view synthesizer (Szymanowicz et al., 2026), with only the dynamic encoder, time module, and mask head trained. The dynamic encoder follows the architecture of the pretrained static encoder, with added cross-attention module for each layer but uses one-third of its depth, with 8 transformer layers compared to 24 in the static encoder, since temporal variation accounts for only a relatively small portion of the scene. Further architectural and training details are provided in Appendix A.

Training dataset. To effectively disentangle camera and object motion, we train our model using synchronized multi-view video datasets such as Kubric4D (Van Hoorick et al., 2024) and Camx-Time (Huang et al., 2026), captured by V cameras with fixed poses with T frames. Unlike costly, explicit 3D annotation (*e. g.*, dense depth or 3D motion) that is commonly used for 3D-Gaussian-based approaches, the multi-view datasets provide cost-effective indirect 4D supervision from 2D images. Furthermore, our residual formulation specifically makes this RGB-only supervision directly tractable, with requiring relatively small amount as well.

Given a multi-view video $\mathcal{I} = \{I_{v,t} \mid v \in \{1, \cdots, V\}, t \in \{1, \cdots, T\}\}$, we adopt an asynchronous sampling strategy to ensure generalization to monocular video inputs at test time as well as accurate conditioning on arbitrary viewpoints and timestamps. we sample N distinct view-time pairs $\{(v_i, t_i)\}_{i=1}^{N} \sim \mathcal{U}(\{1, \cdots, V\} \times \{1, \cdots, T\})$, then construct an asynchronous input set $\mathcal{I}_{\text{in}} = \{I_{v_i,t_i}\}_{i=1}^{N}$. We then sample the target frames from the remaining set such that $\mathcal{I}_{\text{tar}} \subseteq \mathcal{I} \setminus \mathcal{I}_{\text{in}}$. Holding the viewpoint fixed isolates temporal changes caused by dynamic objects, whereas fixing the timestamp provides observations from varying viewpoints, enabling supervision analogous to

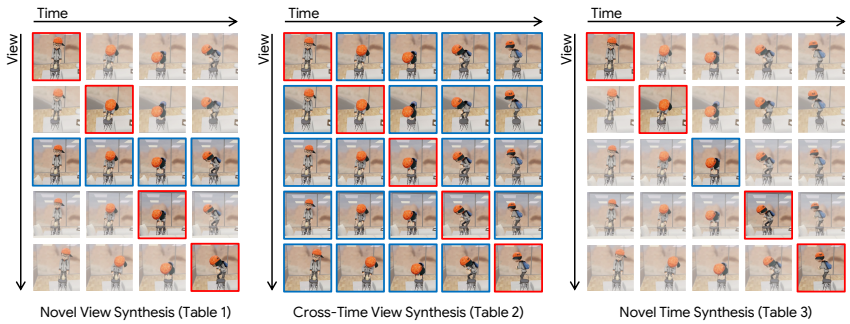

Figure 6: **Three evaluation settings** given synchronized multi-view video datasets. Red outlines indicate input context frames, while blue outlines indicate target frames for evaluation.

Table 1: **Quantitative results on Novel View Synthesis:** FL4N outperforms all Gaussian-based methods on all benchmark datasets, showing the strength of our latent representation bypassing explicit primitives. Context pose input further improves the metrics, achieving the state of the art. Best in **bold**, second best underlined.

Method	Pose	NVIDIA			N3DV			Kubric4D			CamxTime			Charge		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Feed-forward Gaussian Representation																
4DGT (Xu et al., 2025)	✓	16.127	0.3884	0.5972	17.813	0.6609	0.5131	14.148	0.5027	0.6096	17.125	0.6557	0.4893	18.017	0.7786	0.5653
MoVieS (Lin et al., 2026a)	✓	11.818	0.2546	0.6760	20.136	0.7045	0.4605	12.411	0.3550	0.6070	17.069	0.6223	0.4488	18.709	0.7473	0.5298
NeoVerse† (Yang et al., 2026)	✗	15.050	0.4135	0.4626	12.953	0.4806	0.5697	8.954	0.2573	0.6721	10.432	0.3110	0.5905	15.224	0.5211	0.5919
C4G† (Kim et al., 2026c)	✗	16.572	0.4182	0.6203	18.444	0.7050	0.5246	14.931	0.5656	0.6442	20.527	0.7056	0.4369	18.535	0.7797	0.5279
Feed-forward Latent-based 4D Representation																
FL4N (ours)	✗	17.945	0.4113	0.5229	18.581	0.6898	0.4776	20.916	0.6433	0.3767	21.268	0.7237	0.2743	20.924	0.8217	0.4952
FL4N (ours)	✓	**20.435**	**0.4835**	**0.3867**	**22.896**	**0.7716**	**0.3560**	**22.209**	**0.6737**	**0.3350**	**25.730**	**0.8270**	**0.1681**	**22.959**	**0.8324**	**0.4616**

Table 2: On the challenging **Cross-Time View Synthesis** task, FL4N substantially outperforms all Gaussian-based methods on all benchmark datasets, even without test-time pose alignment (†), showing the model successfully learns 4D priors and decouples view and time in the latent space. Best in **bold**, second best underlined.

Method	Pose	NVIDIA			N3DV			Kubric4D			CamxTime			Charge		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Feed-forward Gaussian Representation																
4DGT (Xu et al., 2025)	✓	15.937	0.3867	0.6052	16.097	0.6380	0.5483	15.680	0.5424	0.5786	17.024	0.6573	0.4939	17.939	0.7756	0.5617
MoVieS (Lin et al., 2026a)	✓	15.785	0.4614	0.5328	19.710	0.7400	0.4414	16.988	0.6368	0.4174	17.340	0.6470	0.4224	20.113	0.7771	0.5084
NeoVerse† (Yang et al., 2026)	✗	14.449	0.3819	0.4890	12.447	0.4729	0.5681	8.885	0.2256	0.6793	10.276	0.3091	0.5911	15.220	0.5380	0.5940
C4G† (Kim et al., 2026c)	✗	16.403	0.4096	0.6304	17.039	0.6829	0.5374	15.270	0.5572	0.6348	20.175	0.7009	0.4453	18.410	0.7640	0.5324
Feed-forward Latent-based 4D Representation																
FL4N (ours)	✗	21.552	0.5348	0.3739	24.550	0.8152	0.3353	23.547	0.7535	0.2558	24.546	0.8040	0.2016	24.222	**0.8500**	0.4413
FL4N (ours)	✓	**21.838**	**0.5578**	**0.3562**	**25.109**	**0.8284**	**0.3234**	**24.272**	**0.7850**	**0.2304**	**27.648**	**0.8732**	**0.1325**	**24.421**	0.8496	**0.4382**

static multi-view synthesis. This setup naturally trains ours to disentangle temporal dynamics from viewpoint changes.

Evaluation setting. We evaluate three settings, novel view synthesis (NVS), cross-time view synthesis (CTVS), and novel time synthesis (NTS), illustrated in Fig. 6. NVS evaluates held-out viewpoints at input timestamps, CTVS evaluates unseen viewpoints-timestamps combinations from a monocular video, and NTS evaluates intermediate timestamps from temporally strided inputs. NVS and CTVS use multi-view datasets from fixed cameras: NVIDIA (Yoon et al., 2020), N3DV (Li et al., 2022), Kubric4D, CamxTime, and Charge (Nazarczuk et al., 2026). NTS is evaluated on monocular videos captured by moving cameras: DyCheck (Gao et al., 2022), ADT (Pan et al., 2023), and TUM (Sturm et al., 2012). Kubric4D and CamxTime are evaluated on held-out test scenes from our training distribution, whereas the other datasets evaluate zero-shot generalization. Detailed evaluation protocol is provided in Appendix B.

4.1 Results

Quantitative results. Tables 1, 2, and 3 report the three protocols. Here, † denotes 2,000-step test-time pose alignment with target frames, and Pose (✓/✗) indicates context pose input. FL4N achieves state-of-the-art performance among all pose-free baselines, even without test-time alignment, except for SSIM on NVS of two real-world datasets. On real-world datasets, whose camera poses are noisy estimates, C4G benefits from test-time pose alignment while our model is forced to estimate slight inaccurate pose relation, resulting in a slight misalignment of pixels. An optionally incorporation of context camera poses further improves the accuracy by exploiting the explicit context-to-target

Table 3: On **Novel Time Synthesis**, FL4N achieves the state-of-the-art accuracy by a large margin over all methods, showing the strong generalization of our latent-based approach at novel interpolated timestamps.

Method	Pose	DyCheck			ADT			TUM		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Feed-forward Gaussian Representation										
4DGT (Xu et al., 2025)	✓	13.743	0.4079	0.6233	16.710	0.4999	0.5651	14.380	0.4468	0.5612
MoVieS (Lin et al., 2026a)	✓	11.755	0.3420	0.6428	15.432	0.4698	0.5828	15.683	0.5079	0.4426
NeoVerse† (Yang et al., 2026)	✗	16.756	0.4663	0.4566	16.179	0.4384	0.5296	16.120	0.5266	0.3981
C4G† (Kim et al., 2026c)	✗	15.755	0.4735	0.6111	18.900	0.5813	0.5833	17.508	0.5518	0.4951
Feed-forward Latent-based 4D Representation										
FL4N (ours)	✗	19.019	0.4977	0.4284	19.693	0.5948	0.4490	17.847	0.5586	0.3449
FL4N (ours)	✓	**20.527**	**0.5470**	**0.3757**	**24.115**	**0.6875**	**0.3083**	**21.162**	**0.6978**	**0.2601**

pose relation achieving state-of-the-art performance across all benchmarks on all protocols. These results demonstrate the substantial strength of our proposed latent representation: strong 4D prior, generalization to novel view and time, and decouplement of view and time, all successfully learned in the latent space. Additional rendering quality evaluation on dynamic region and discussion on test-time pose alignment are provided in Appendices D.1 and F.

Ablation. Table 4 ablates our key design choices to justify our formulation, with all variants trained and evaluated on Kubric4D for 10k steps. Note that, Since the ablation models are trained only on Kubric4D and evaluated in-domain, they achieve higher absolute scores than our general model in Table 1, which is trained on a mixture of datasets. These numbers should therefore be interpreted only relative to each other within this ablation. We also include metrics on dynamic region to see its gain exclusively. The residual branch denotes our newly introduced modules (*i. e.*, the dynamic encoder and the time module) on top of the static NVS model.

Table 4: **Ablation studies** on Kubric4D under the Novel View Synthesis protocol, with accuracy on full frames and dynamic region. Each design choice contributes to our final formulation.

Variant	Components				Full frame			Dynamic region		
	Residual branch	Static injection	Mask loss	Frozen renderer	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
(i)	✗	–	✗	✗	21.61	0.666	0.342	14.34	0.507	0.520
(ii)	✓	✓	✓	✗	21.80	0.674	0.332	14.55	0.523	0.497
(iii)	✓	✓	✗	✓	22.31	0.691	0.293	14.51	0.537	0.469
(iv)	✓	✗	✓	✓	22.51	0.683	0.289	14.42	0.539	0.467
FL4N	✓	✓	✓	✓	**22.69**	**0.696**	**0.288**	**14.71**	**0.549**	**0.465**

Variant (i) is a baseline that fully fine-tunes the static NVS model on the dynamic dataset, showing huge regression from FL4N. Freezing the render is crucial (variant (ii)); unfreezing it disrupts dynamic branch training by destabilizing the established latent-to-image mapping in the renderer. Dynamic mask loss (variant (iii)) modestly improves the accuracy by guiding dynamic region during training. Without it, the model tends to collapse as training progresses. Static feature injection to the residual encoder (variant (iv)) further improves the overall accuracy, helping the residual branch to effectively isolate temporal variation from injected intermediate features from the static encoder.

Runtime. Table 5 reports runtime at 224 × 224 resolution on one RTX 5880 Ada. FL4N achieves the fastest runtime. Ours is 1.08× faster than NeoVerse and 565× faster than that with pose alignment, with ours achieving substantially better accuracy.

Table 5: **Runtime.** † denotes 2,000 step test-time pose alignment.

Method	sec / view ↓	FPS ↑
4DGT (Xu et al., 2025)	0.040	25.00
MoVieS (Lin et al., 2026a)	0.037	27.02
NeoVerse (Yang et al., 2026)	0.027	37.03
NeoVerse† (Yang et al., 2026)	14.127	0.07
C4G (Kim et al., 2026c)	0.036	27.77
C4G† (Kim et al., 2026c)	23.885	0.04
FL4N (ours)	**0.025**	**40.00**

Qualitative results. Figure 7 qualitatively compares our results with Gaussian-based methods. Gaussian-based methods often exhibit holes, object inconsistencies, and blurred details when rendering at novel viewpoints and timestamps. In contrast, our method produces visually consistent results. This again confirms the superiority of our latent-based method over explicit Gaussian primitives: 4D prior and view-time disentanglement can be learned successfully in the latent space, and novel view are seamlessly rendered. Appendix D.3 includes comprehensive qualitative comparisons on all benchmark datasets.

4.2 Latent Probing on Downstream Tasks

To assess whether the dynamic encoder captures temporal changes while preserving scene geometry encoded by static feature, we probe the representations via zero-shot temporal matching and video-depth probing. In tables, "All" covers the entire frame, while "Dyn" is restricted to dynamic objects. Full matching and depth results, along with point-tracking results, are provided in Appendix E.



Figure 7: **Qualitative comparison on 4D Novel View Synthesis.** Gaussian-based methods often exhibit holes, object inconsistencies, and blurred details. Our method produces more coherent and visually consistent results, demonstrating the superiority of our latent-based representation over explicit geometry primitives.

Zero-shot matching. While the static prior establishes dense correspondences across views, exploiting epipolar geometry, moving content violates the multi-view consistency, degrading correspondence quality in dynamic regions. On the other hand, dynamic prior improves correspondence in such regions. In Table 6, we evaluate zero-shot matching of static ($\mathbf{z}_{\text{st}}$) and dynamic latent ($\mathbf{z}_{\text{dyn}}$) on DAVIS, using the protocol of Nam et al. (2026). The dynamic latent consistently demonstrates better accuracy than the static latent; the gap slightly widens on dynamic region.

Table 6: **Zero-shot matching on DAVIS.** Dynamic latent outperform static latent by encoding temporal information.

Region	Feat.	δ_{avg}	$\delta 1$	$\delta 4$	$\delta 16$
All	$\mathbf{z}_{\text{st}}$	27.33	5.65	25.37	53.40
	$\mathbf{z}_{\text{dyn}}$	**29.04** (+1.71)	**5.76** (+0.11)	**26.50** (+1.13)	**57.64** (+4.24)
Dyn	$\mathbf{z}_{\text{st}}$	23.45	3.10	21.14	49.03
	$\mathbf{z}_{\text{dyn}}$	**25.29** (+1.84)	**3.12** (+0.02)	**22.37** (+1.23)	**53.68** (+4.65)

Video depth estimation. Although the model receives no explicit geometric supervision, its latent representations can encode scene geometry information. We probe this by attaching DPT depth predictors on the static and dynamic latent and training them Kubric (Greff et al., 2022). Table 7 reports the video depth accuracy. Dynamic latent consistently outperforms static latent, across the full frame and dynamic-object regions. Notably, the advantage becomes more pronounced as the temporal stride between context frames increases, corresponding to larger object motion. This suggests that dynamic latent largely preserves geometry-aware information encoded by the static latent while incorporating additional cues from temporal dynamics.

Table 7: **Probing video depth on Kubric.** Dynamic latent yields consistent gains, with larger gains at greater temporal strides.

Region	Feat.	$\Delta 4$		$\Delta 8$	
		AbsRel↓	$\delta < 1.25$ ↑	AbsRel↓	$\delta < 1.25$ ↑
All	$\mathbf{z}_{\text{st}}$	0.0944	0.8908	0.0868	0.9081
	$\mathbf{z}_{\text{dyn}}$	**0.0863** (−0.0081)	**0.9035** (+0.0127)	**0.0751** (−0.0117)	**0.9326** (+0.0244)
Dyn	$\mathbf{z}_{\text{st}}$	0.1178	0.8699	0.1149	0.8790
	$\mathbf{z}_{\text{dyn}}$	**0.1050** (−0.0128)	**0.8941** (+0.0242)	**0.0996** (−0.0153)	**0.9099** (+0.0309)

5 Conclusion

4D NVS is a fundamental task for modeling and understanding the world around us, with broad applications. Yet existing approaches typically rely on explicit 4D primitives to model dynamic scenes, often requiring dense geometry and motion supervision but suffer from temporal inconsistency and holes. In this work, we introduced FL4N, a feed-forward latent model that renders dynamic scenes from a monocular video at novel views and timestamps. By learning a temporal residual on top of a frozen static latent model, FL4N acquires a 4D representation from limited dynamic data, supervised only by images and binary dynamic masks. Across eight benchmarks and three evaluation protocols, FL4N outperforms prior feed-forward Gaussian-based methods and generalizes to real-world videos, while its latent space successfully disentangles view from time and preserves scene geometry. We hope our work opens the door to further research in latent 4D understanding.

AI USE STATEMENT

In this work, we used generative AI tools (LLMs) to aid and polish the writing to improve readability and clarity. We have not used generative AI tools for generating synthetic datasets, developing theoretical models, formulating or proving mathematical claims, proposing hypotheses, designing experiments, or interpreting results. Additionally, we used generative AI tools to create or edit software code. We have reviewed all AI-assisted work. We treated all AI outputs only as references and did not adopt any of them as is. Based on these references, the authors selectively revised and incorporated only the necessary parts. Every AI-assisted decision was discussed and made by at least two authors. We take full responsibility for the final content of this work produced with the aid of generative AI.

REFERENCES

Honggyu An, Jaewoo Jung, Mungyeom Kim, Chaehyun Kim, Minkyeong Jeon, Jisang Han, Kazumi Fukuda, Takuya Narihira, Hyunah Ko, Junsu Kim, Sunghwan Hong, Yuki Mitsufuji, and Seungryong Kim. Learning compact 3D representations from feed-forward novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 363–373, June 2026.

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. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14834–14844. IEEE, 2025.

Matteo Balice, Yanik Kunzi, Chenyangguang Zhang, Matteo Matteucci, Marc Pollefeys, and Sungwhan Hong. No Pose, No Problem in 4D: Feed-forward dynamic gaussians from unposed multi-view videos. *arXiv preprint arXiv:2605.22190*, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts, 2025.

Kaihua Chen, Tarasha Khurana, and Deva Ramanan. Reconstruct, inpaint, test-time finetune: Dynamic novel-view synthesis from monocular videos. *Advances in Neural Information Processing Systems*, 38:47723–47753, 2026.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Easi3R: Estimating disentangled motion from DUSt3R without training. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9158–9168. IEEE, 2025.

Seokju Cho, Abhishek Badki, Hang Su, Jindong Jiang, Ziyao Zeng, Seungryong Kim, Sifei Liu, and Orazio Gallo. 4DP-QA: Scalable qa for 4D perception in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23869–23879, June 2026.

Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. *Advances in Neural Information Processing Systems*, 35: 33768–33780, 2022.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3749–3761, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jisang Han, Jiaolong Yang, Chong Luo, and Seungryong Kim. PF3plat: Pose-free feed-forward 3D gaussian splatting for novel view synthesis. In *Forty-second International Conference on Machine Learning*, 2025.

Yu Hu, Chong Cheng, Sicheng Yu, Xiaoyang Guo, and Hao Wang. VGGT4D: Mining motion cues in visual geometry transformers for 4D scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 414–424, 2026.

Zhening Huang, Hyeonho Jeong, Xuelin Chen, Yulia Gryaditskaya, Tuanfeng Y. Wang, Joan Lasenby, and Chun-Hao Huang. SpaceTimePilot: Generative rendering of dynamic scenes across space and time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11218–11228, 2026.

Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, et al. RayZer: A self-supervised large view synthesis model. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4918–4929, 2025.

Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. LVSM: A large view synthesis model with minimal 3D inductive bias. In *The Thirteenth International Conference on Learning Representations*, 2025.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labeling real videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10. IEEE, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Evan Kim, Hyunwoo Ryu, Thomas W Mitchel, and Vincent Sitzmann. Scaling view synthesis transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28893–28902, 2026a.

Jisoo Kim, Jungbin Cho, Sanghyeok Chu, Ananya Bal, Jinhyung Kim, Gunhee Lee, Sihaeng Lee, Seung Hwan Kim, Bohyung Han, Hyunmin Lee, et al. Pri4R: Learning world dynamics for vision-language-action models with privileged 4D representation. *arXiv preprint arXiv:2603.01549*, 2026b.

Mungyeom Kim, Minkyeong Jeon, Honggyu An, Jaewoo Jung, Hyuna Ko, Jisang Han, Hyeonseo Yu, Donghwan Shin, Sunghwan Hong, Takuya Narihira, et al. Learning global motion with compact gaussians for feed-forward 4D reconstruction. *arXiv preprint arXiv:2605.31595*, 2026c.

Renjie Li, Panwang Pan, Bangbang Yang, Dejia Xu, Shijie Zhou, zhang xuanyang, Zeming Li, Achuta Kadambi, Zhangyang Wang, Zhengzhong Tu, and Zhiwen Fan. 4K4DGen: Panoramic 4D generation at 4K resolution. In *International Conference on Learning Representations*, pp. 100692–100713, 2025.

Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3D video synthesis from multi-view video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5511–5521. IEEE, 2022.

Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. Spacetime gaussian feature splatting for real-time dynamic view synthesis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8508–8520. IEEE, 2024.

Hanxue Liang, Jiawei Ren, Ashkan Mirzaei, Antonio Torralba, Ziwei Liu, Igor Gilitschenski, Sanja Fidler, Cengiz Oztireli, Huan Ling, Zan Gojcic, and Jiahui Huang. Feed-forward bullet-time reconstruction of dynamic scenes from monocular videos. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Chenguo Lin, Yuchen Lin, Panwang Pan, Yifan Yu, Tao Hu, Honglei Yan, Katerina Fragkiadaki, and Yadong Mu. MoVieS: Motion-aware 4D dynamic view synthesis in one second. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 295–306, 2026a.

Chieh Lin, Zhaoyang Lv, Songyin Wu, Zhen Xu, Thu Nguyen-Phuoc, Hung-Yu Tseng, Julian Straub, Numair Khan, Lei Xiao, Ming-Hsuan Yang, et al. DGS-LRM: Real-time deformable 3D gaussian reconstruction from monocular videos. *Advances in Neural Information Processing Systems*, 38:46915–46936, 2026b.

Thomas W. Mitchel, Hyunwoo Ryu, and Vincent Sitzmann. True self-supervised novel view synthesis is transferable. *arXiv preprint arXiv:2510.13063*, 2025.

Jisu Nam, Soowon Son, Dahyun Chung, Jiyoung Kim, Siyoon Jin, Junhwa Hur, and Seungryong Kim. Emergent temporal correspondences from video diffusion transformers. *Advances in Neural Information Processing Systems*, 38:123683–123727, 2026.

Michal Nazarczuk, Thomas Tanay, Arthur Moreau, Zhensong Zhang, and Eduardo Pérez-Pellitero. Charge: A comprehensive novel view synthesis benchmark and dataset to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15323–15333, 2026.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3D machine perception. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20076–20086. IEEE, 2023.

Ulrich Prestel, Stefan Andreas Baumann, Nick Stracke, and Björn Ommer. RayDer: Scalable self-supervised novel view synthesis from real-world video. In *European Conference on Computer Vision*, pp. 558–578, 2026.

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10313–10322, 2021.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12159–12168, 2021.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132, 2025.

Mehdi SM. Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6219–6228, 2022.

Mehdi SM Sajjadi, Aravindh Mahendran, Thomas Kipf, Etienne Pot, Daniel Duckworth, Mario Lučić, and Klaus Greff. RUST: Latent neural scene representations from unposed imagery. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17297–17306. IEEE, 2023.

Maximilian Seitzer, Sjoerd van Steenkiste, Thomas Kipf, Klaus Greff, and Mehdi SM. Sajjadi. DyST: Towards dynamic neural scene representations on real-world videos. In *International Conference on Learning Representations*, volume 2024, pp. 37052–37065, 2024.

Soowon Son, Honggyu An, Jisu Nam, Hyunah Ko, Chaehyun Kim, Dahyun Chung, Siyoon Jin, Jung Yi, Junhwa Hur, and Seungryong Kim. Probing and leveraging video diffusion transformer features for robust point tracking. *arXiv preprint arXiv:2512.20606*, 2025.

Rui Song, Chenwei Liang, Yan Xia, Walter Zimmer, Hu Cao, Holger Caesar, Andreas Festag, and Alois Knoll. CoDa-4DGS: Dynamic gaussian splatting with context and deformation awareness for autonomous driving. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28031–28041. IEEE, 2025.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 573–580. IEEE, 2012.

Stanislaw Szymanowicz, Minghao Chen, Jianyuan Wang, Christian Rupprecht, and Andrea Vedaldi. LagerNVS: Latent geometry for fully neural real-time novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15443–15453, 2026.

Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative camera dolly: Extreme monocular dynamic novel view synthesis. In *European Conference on Computer Vision*, pp. 313–331. Springer, 2024.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025a.

Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4D reconstruction from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9660–9672, 2025b.

Yihan Wang and Jia Deng. WAFT: Warping-alone field transforms for optical flow. In *International Conference on Learning Representations*, volume 2026, pp. 157389–157402, 2026.

Daniel Watson, Saurabh Saxena, Lala Li, Andrea Tagliasacchi, and David Fleet. Controlling space and time with diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 95989–96012, 2025.

Kai Xu, Tze Ho Elden Tse, Jizong Peng, and Angela Yao. DAS3R: Dynamics-aware gaussian splatting for static scene reconstruction. *arXiv preprint arXiv:2412.19584*, 2024a.

Zhen Xu, Sida Peng, Haotong Lin, Guangzhao He, Jiaming Sun, Yujun Shen, Hujun Bao, and Xiaowei Zhou. 4K4D: Real-time 4D view synthesis at 4K resolution. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20029–20040. IEEE, 2024b.

Zhen Xu, Zhengqin Li, Zhao Dong, Xiaowei Zhou, Richard Newcombe, and Zhaoyang Lv. 4DGT: Learning a 4D gaussian transformer using real-world monocular videos. In *Advances in Neural Information Processing Systems*, pp. 42606–42632, 2025.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. NeoVerse: Enhancing 4D world model with in-the-wild monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40340–40351, June 2026.

Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3D gaussian splats from sparse unposed images. In *International Conference on Learning Representations*, volume 2025, pp. 54009–54033, 2025.

Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5335–5344. IEEE, 2020.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. TrajectoryCrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 100–111, 2025.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, trevor darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A simple approach for estimating geometry in the presence of motion. In *International Conference on Learning Representations*, pp. 82863–82886, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. IEEE, 2023.

Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8535–8546. IEEE, 2025.

APPENDIX CONTENTS

A Implementation details **16**

A.1 Architecture details 16

A.2 Training details 16

B Evaluation protocol **16**

B.1 Evaluation datasets 16

B.2 Evaluation protocols 17

B.3 Model specific configurations 17

C Latent analysis **18**

D Additional results **18**

D.1 Quantitative comparison on dynamic regions 18

D.2 Additional analysis on quantitative results 19

D.3 Additional qualitative results 19

D.4 Dynamic mask 20

D.5 Context frame number ablation 20

E Probings **20**

E.1 Full probing results 20

E.2 Point tracking 21

F Test-time pose alignment for latent models **22**

G Limitations **22**

A Implementation details

A.1 Architecture details

Static encoder and renderer. The static prior used is a pretrained feed-forward novel view synthesizer whose encoder is a 24-layer transformer alternating frame-wise and global attention, followed by a 12-layer transformer renderer. Both use a hidden size of 768 with 12 heads on 8×8 patches, and remain frozen during our training.

Dynamic encoder. The dynamic encoder follows the same architecture as the static encoder but uses one-third of its depth, with 8 layers. Each layer additionally cross-attends to the last layer of static features, allowing the residual branch to access camera-pose-aware representations throughout encoding. This provides a spatial anchor for learning motion residuals, complementing the static features rather than relearning scene structure. A DPT decoder (Ranftl et al., 2021) attached to the dynamic encoder predicts the binary dynamic mask.

Temporal conditioning. We use a single learnable time token $\boldsymbol{\tau} \in \mathbb{R}^C$ with $C = 1024$, and attach a copy of it to each input view by concatenating it with that view's tokens. Each copy becomes view-specific ($\hat{\boldsymbol{\tau}}_k$) through attention with its own view, allowing the same token to handle an arbitrary number and ordering of context frames. The resulting time representation is processed by a time module consisting of two transformer blocks with adaptive layer normalization conditioned on the target-time token, followed by a zero-initialized linear residual gate.

A.2 Training details

Training setup. We train with AdamW at a learning rate of 3.5×10^{-5}, weight decay 0.05, and 3k warmup steps, while keeping the static encoder and renderer frozen for whole training. FL4N is trained on four NVIDIA RTX 5880 Ada Generation GPUs with a batch size of 8 for total of 32k steps. Each training sample contains 8 conditioning views and 12 target views at a resolution of 336×512.

Training datasets. We use Kubric4D (Van Hoorick et al., 2024) and CamxTime (Huang et al., 2026), which provide synchronized multi-view videos captured by static cameras. To disentangle camera and object motion, we train the model to reconstruct target frames at view–time pairs not included in the input set.

Pose dropout. To support pose-free conditioning at inference, we randomly zero out the conditioning camera tokens with probability 0.4 during training. At inference time, FL4N supports fully unposed setting by zeroing all conditioning camera tokens.

Loss. We use an ℓ_2 photometric loss $\mathcal{L}_2$, LPIPS losses $\mathcal{L}_{\text{perc}}$ and $\mathcal{L}^{\text{dyn}}_{\text{perc}}$, and an ℓ_1 image-gradient loss $\mathcal{L}^{\text{dyn}}_{\text{grad}}$. The dynamic-region losses are computed using the dynamic mask M_k to select moving pixels, on 128×128 crops centered on the moving content. For the image loss term, the dynamic and static regions are averaged over their respective areas and weighted equally. Dynamic mask loss is only comptued for context frames. We set $\lambda_{\text{perc}} = 0.5$, $\lambda^{\text{dyn}}_{\text{perc}} = 2.0$, $\lambda^{\text{dyn}}_{\text{grad}} = 1.0$, and $\lambda_{\text{mask}} = 1.0$.

Pseudo dynamic mask generation. We generate pseudo dynamic masks by thresholding the magnitude of optical flow estimated by an off-the-shelf flow estimator (Wang & Deng, 2026). Since the flow is computed between frames captured by the same static camera, camera-induced motion and parallax are absent, making the remaining flow primarily reflect object motion. This leverages the static-camera multi-view configuration of our training data to identify dynamic regions with binary pseudo labels. The mask generation can also be easily replaced with off-the-shelf segmentation models, such as SAM3 (Carion et al., 2025), without changing the training framework.

B Evaluation protocol

B.1 Evaluation datasets

Our benchmark comprises five multi-view datasets for NVS and CTVS: NVIDIA, N3DV, Kubric4D, CamxTime, and Charge, and three datasets for NTS: ADT, DyCheck, and TUM. Kubric4D and

CamxTime use held-out scenes, while the remaining datasets are unseen during training and evaluate cross-dataset generalization.

For static camera multi-view datasets, we simulate moving camera monocular video by selecting N diagonal frames, *i. e.* camera i observed at time i as input.

Kubric4D. We use 100 test scenes from Kubric4D, which contains 16 static cameras, with the first four providing top-down views and the remaining 12 cameras arranged around the scene on a horizontal ring. Each camera captures 60 frames. We exclude the first four cameras and sample cameras consecutively from the horizontal ring. We use a frame interval of 8.

CamxTime. CamxTime contains four separate camera trajectories, each comprising 120 densely spaced cameras with 120 frames per camera. We hold out the last 10% of scenes for testing and use `Trajectory_01` for evaluation, with camera and frame intervals both set to 8.

Charge. we use the `Sparse_Test` cameras. Charge stores cameras in a non-spatial order, so sampling by file index leaves large gaps between consecutive conditioning views. We therefore reorder the cameras of each scene along a continuous open path through the rig and uniformly sample along this path to approximate a smooth monocular camera trajectory. we use the following camera order:

`C05 C01 C04 C06 C02 C08 C00 C07 C03 C09`,

which follows the shortest open path through the rig where `Ci` denotes the i-th test camera. CTVS uses the first 8 cameras of this order, and NVS holds out `C02`. We use a frame interval of 8.

B.2 Evaluation protocols

NVS (Novel View Synthesis). We take 9 frames and hold out the 5th as a target. The remaining 8 are used as the input, and the held-out frame's viewpoint is the target at all 8 conditioning times. For N3DV, `cam0` is used for a target camera.

CTVS (Cross-time View Synthesis). Given N diagonal frames as input, the targets are the N^2-N off-diagonal cells of the camera–time grid, evaluating synthesis at unseen viewpoint–timestamp combinations. NVIDIA uses all available cameras (N=12), and N3DV uses all cameras except `cam0` (N=17–20 depending on the scene). Kubric4D, CamxTime, and Charge each use N=8.

NTS (Novel Time Synthesis). Following C4G (Kim et al., 2026c), we use the following scenes from each dataset:

- **DyCheck**
  - `apple`, `block`, `paper-windmill`, `spin`, `teddy`
- **ADT**
  - `Apartment_release_multiuser_cook_seq141_M1292`
  - `Apartment_release_multiskeleton_party_seq114_M1292`
  - `Apartment_release_meal_skeleton_seq135_M1292`
  - `Apartment_release_work_skeleton_seq137_M1292`
- **TUM**
  - `fr2_desk_with_person`
  - `fr3_sitting_{halfsphere, rpy, static, xyz}`
  - `fr3_walking_{halfsphere, rpy, static, xyz}`

All three datasets use a single moving camera for both context and target views. We follow the C4G (Kim et al., 2026c) evaluation protocol and sample 16 frames with stride 6 and evaluate on the 15 interleaved frames.

B.3 Model specific configurations

Every model runs at the native input resolution. For the pose-free baselines, NeoVerse (Yang et al., 2026) and C4G (Kim et al., 2026c), we apply 2,000 steps of test-time pose alignment (Ye et al.,

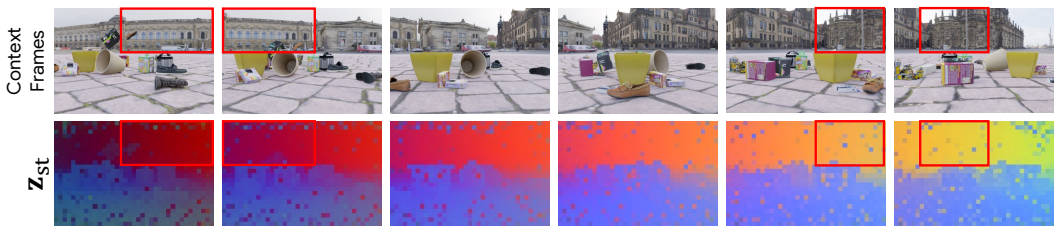

Figure A: **Visualization of static latent.** The static latent encodes camera pose, with similar colors for cameras pointing in the same direction. Red boxes highlight overlapping regions across context frames, which exhibit similar color gradients.

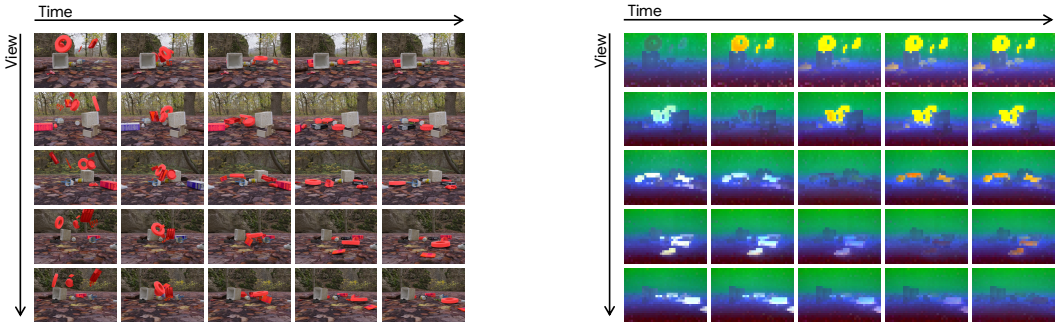

Figure B: **Visualization of residual latent.** *(Left)* Ground-truth RGB frame overlaid with the dynamic mask. *(Right)* Corresponding residual latent grid. Dynamic objects exhibit more pronounced residual latents as the target pose and time diverge from the input, indicating a larger discrepancy from the context frames.

2025). In contrast, FL4N does not perform pose alignment, despite being evaluated in the pose-free setting. For real-world datasets, FL4N uses camera poses predicted by VGGT (Wang et al., 2025a), using the estimated intrinsics for NVIDIA and the estimated intrinsics and extrinsics for N3DV.

C Latent Analysis

Figure A shows a PCA visualization of the static latent $\mathbf{z}_{\text{st}}$ under camera motion. The static latent primarily captures camera pose, with similar features corresponding to similar viewing directions.

Figure B visualizes the time-conditioned residual features along with their corresponding GT dynamic masks. For each view i and target time j, the residual feature is spatially anchored to the diagonal frame at view i and time i. Dynamic objects become more pronounced in the residual as their position at time j increasingly differs from that at the anchor time i, while showing weaker responses near the diagonal where the temporal displacement is small. This trend indicates that the residual latent encodes timestamp-dependent differences, with larger feature variations corresponding to larger object motion.

D Additional Results

D.1 Quantitative Comparison on Dynamic Regions

On the two datasets with ground-truth motion masks, Kubric4D and Charge (Nazarczuk et al., 2026), we additionally report metrics computed only within dynamic regions. Since moving content typically occupies a small fraction of the frame, dynamic-region metrics provide a more focused evaluation of reconstruction quality on moving subjects. Table A reports the dynamic-region metrics for NVS and CTVS, where our model consistently outperforms the baselines.

Table A: **4D NVS result for Dynamic region.** † denotes 2,000 step test-time pose alignment, the best result in each column is in **bold**, and the second best is underlined. Metrics are restricted to the moving pixels and are computed over the same target subset, and their full-frame counterparts for the same cells are reported in Tables 1 and 2.

Method	Pose	NVS						CTVS					
		Kubric4D			Charge			Kubric4D			Charge		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Feed-forward 4D Gaussian Representation													
4DGT (Xu et al., 2025)	✓	12.089	0.2406	0.0305	17.703	0.6516	0.1127	12.405	0.2557	0.0303	17.708	0.6588	0.1157
MoVieS (Lin et al., 2026a)	✓	11.401	0.1741	0.0318	17.969	0.6079	0.1130	11.797	0.2034	0.0307	18.068	0.6119	0.1147
NeoVerse† (Yang et al., 2026)	✗	10.847	0.1527	0.0381	16.292	0.5161	0.1242	10.254	0.1372	0.0403	16.152	0.4966	0.1307
C4G† (Kim et al., 2026c)	✗	13.031	0.3082	0.0322	18.552	0.6617	0.1143	12.901	0.3052	0.0333	18.385	0.6566	0.1166
Feed-forward Latent-based 4D Representation													
FL4N (ours)	✗	14.774	0.3479	0.0240	19.906	0.6708	0.1063	**14.482**	0.3351	0.0246	19.748	**0.6774**	0.1019
FL4N (ours)	✓	**14.800**	**0.3485**	**0.0236**	**19.993**	**0.6711**	**0.1029**	14.451	**0.3358**	**0.0244**	**19.935**	0.6772	**0.1009**



Figure C: **Further qualitative comparison.** Gaussian-based methods exhibit holes, object inconsistencies, and blurred details. Our method produces more coherent and visually consistent results, demonstrating the superiority of our latent-based representation over explicit geometry primitives.

D.2 Additional analysis on quantitative results

In the main table, NeoVerse performs notably poorly on Kubric4D. This is because NeoVerse tends to provide weaker information across the temporal axis and produces many holes, leaving regions unseen at a given timestamp largely unmodeled. This issue is particularly severe on Kubric4D, where the large camera gaps span nearly the entire viewing range, making regions observed at different timestamps, such as the background, especially difficult to reconstruct.

D.3 Additional qualitative results

Figure C shows additional qualitative comparisons for 4D NVS. Figures F to K provide comparisons across all evaluated methods and scenes. Figures L to N further compare cross-time view synthesis results using view-time grids. Finally, Figs. O to Q show additional view-time grid results of FL4N alongside the corresponding ground-truth views. Across these results, FL4N exhibits stronger temporal consistency and more effective disentanglement of viewpoint and time, producing stable renderings with fewer temporal inconsistencies and visual artifacts.

Table B: **Quantitative results on dynamic regions.** On the dynamic object segmentation task, Our method substantially outperforms other methods for 3D or 4D geometry estimation.

Method	DAVIS-2016				DAVIS-2017				DAVIS-all			
	JM↑	JR↑	FM↑	FR↑	JM↑	JR↑	FM↑	FR↑	JM↑	JR↑	FM↑	FR↑
Easi3R$_{\text{dust3r}}$ (Chen et al., 2025)	50.10	55.77	43.40	37.25	46.86	50.54	39.06	30.05	44.10	50.85	35.16	27.24
Easi3R$_{\text{monst3r}}$ (Chen et al., 2025)	54.93	68.00	45.29	47.30	54.75	66.16	44.09	42.36	51.64	63.06	40.98	38.49
MonST3R (Zhang et al., 2025)	40.42	40.39	49.54	52.12	38.07	36.05	48.24	49.01	36.98	34.52	47.03	46.72
DAS3R (Xu et al., 2024a)	41.13	38.67	44.50	36.94	44.51	43.95	46.71	44.96	43.33	38.93	45.24	38.78
VGGT4D (Hu et al., 2026)	62.12	76.80	56.04	67.49	56.45	65.62	51.09	56.85	50.75	55.59	47.04	46.43
FL4N (ours)	**74.93**	**93.78**	**74.16**	**90.64**	**75.44**	**92.32**	**74.37**	**90.38**	**68.76**	**80.11**	**69.87**	**83.99**



Figure D: **Qualitative comparison of dynamic masks.** Our method show more accurate dynamic object masks with clearer object boundary than VGGT4D (Hu et al., 2026).

D.4 Dynamic mask

In Table B, we report the dynamic mask estimation accuracy with IoU and Boundary F-measure as JM and FM, respectively, and their corresponding recall metrics as JR and FR, following the evaluation protocol of Hu et al. (2026). The reported metrics for the other methods are adopted from Hu et al. (2026). Figure D shows that our model predicts dynamic masks with cleaner boundaries and greater robustness, demonstrating its ability to generalize dynamic object localization to unseen real-world datasets.

D.5 Context frame number ablation

Our model is trained with only 8 context frames, yet Figure E shows that the rendering quality remains on a flat plateau as the number of context frames increases to 32+, more than four times the training configuration. This demonstrates that the dynamic encoder generalizes to substantially longer temporal contexts than trained. The result highlights that our model can effectively exploit temporal context far beyond the number of frames seen during training.

We use CamxTime test scenes, which provide enough number of context frames while maintaining a sufficiently dense set of conditioning candidates. Context frames are sampled along a moving camera trajectory from the camera-time grid within a fixed temporal window. Thus, all settings cover the same camera and temporal extent, while increasing the number of context frames increases the sampling density. We evaluate 12 target frames across four novel camera viewpoints, $21, 45, 69, 93$, and three timestamps, $8, 56, 104$.

E Probings

E.1 Full probing results

Tables C and D provide the full probing results, while the main paper reports a subset of matching thresholds and depth across temporal strides.

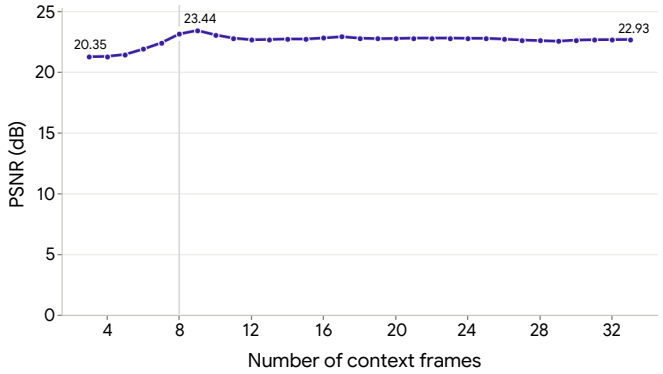

Figure E: **Generalization to the number of context frames**, reporting PSNR on CamxTime as the number of conditioning frames varies at inference. The vertical line marks the 8 frames the model is trained with. The rendering quality remains on a flat plateau as the number of context frames increases to 32+. The dynamic encoder generalizes to substantially longer temporal contexts than trained and effectively exploits temporal context far beyond the number of frames seen during training.

Table C: **Zero-shot matching on DAVIS.** Performance on all threshold are reported. Dynamic latent demonstrates consistently better zero-shot matching accuracy than static latent on all thresholds.

Region	Feat.	δ_{avg}	$\delta 1$	$\delta 2$	$\delta 4$	$\delta 8$	$\delta 16$
All	$\mathbf{z}_{\text{st}}$	27.33	5.65	12.43	25.37	39.78	53.40
	$\mathbf{z}_{\text{dyn}}$	**29.04**	**5.76**	**12.50**	**26.50**	**42.81**	**57.64**
		(+1.71)	(+0.11)	(+0.07)	(+1.13)	(+3.03)	(+4.24)
Dyn	$\mathbf{z}_{\text{st}}$	23.45	3.10	9.12	21.14	34.84	49.03
	$\mathbf{z}_{\text{dyn}}$	**25.29**	**3.12**	**9.20**	**22.37**	**38.05**	**53.68**
		(+1.84)	(+0.02)	(+0.08)	(+1.23)	(+3.21)	(+4.65)

Table D: **Video depth probed on Kubric.** Performance on all strides are reported. The gain is larger at larger temporal gap.

Region	Feat.	$\Delta 1$		$\Delta 2$		$\Delta 4$		$\Delta 8$	
		AbsRel↓	$\delta < 1.25$ ↑	AbsRel↓	$\delta < 1.25$ ↑	AbsRel↓	$\delta < 1.25$ ↑	AbsRel↓	$\delta < 1.25$ ↑
All	$\mathbf{z}_{\text{st}}$	0.1129	0.8550	0.1023	0.8729	0.0944	0.8908	0.0868	0.9081
	$\mathbf{z}_{\text{dyn}}$	**0.1037**	**0.8722**	**0.0932**	**0.8937**	**0.0863**	**0.9035**	**0.0751**	**0.9326**
		(−0.0092)	(+0.0171)	(−0.0092)	(+0.0208)	(−0.0081)	(+0.0127)	(−0.0117)	(+0.0244)
Dyn	$\mathbf{z}_{\text{st}}$	0.1180	0.8597	0.1194	0.8614	0.1178	0.8699	0.1149	0.8790
	$\mathbf{z}_{\text{dyn}}$	**0.1127**	**0.8747**	**0.1078**	**0.8824**	**0.1050**	**0.8941**	**0.0996**	**0.9099**
		(−0.0052)	(+0.0150)	(−0.0116)	(+0.0209)	(−0.0128)	(+0.0242)	(−0.0153)	(+0.0309)

Table E: **Point tracking probed on DAVIS** Tracking performance across different temporal strides are reported.

Region	Feat.	$\Delta 1$			$\Delta 2$			$\Delta 4$			$\Delta 8$		
		AJ↑	$< \delta^x_{\text{avg}}$ ↑	OA↑	AJ↑	$< \delta^x_{\text{avg}}$ ↑	OA↑	AJ↑	$< \delta^x_{\text{avg}}$ ↑	OA↑	AJ↑	$< \delta^x_{\text{avg}}$ ↑	OA↑
All	$\mathbf{z}_{\text{st}}$	37.63	52.35	70.40	36.15	51.65	69.20	33.10	49.00	66.42	28.92	45.82	63.42
	$\mathbf{z}_{\text{dyn}}$	**40.45**	**54.84**	**73.25**	**39.80**	**54.76**	**72.61**	**36.59**	**51.73**	**69.79**	**32.29**	**47.60**	**67.06**
		(+2.82)	(+2.49)	(+2.85)	(+3.64)	(+3.11)	(+3.41)	(+3.49)	(+2.73)	(+3.37)	(+3.38)	(+1.78)	(+3.64)
Dyn	$\mathbf{z}_{\text{st}}$	32.26	47.85	66.12	30.83	47.39	64.80	28.57	45.69	62.65	24.84	42.88	59.72
	$\mathbf{z}_{\text{dyn}}$	**34.82**	**49.86**	**69.23**	**34.09**	**49.53**	**68.65**	**32.01**	**47.70**	**66.64**	**28.75**	**44.81**	**64.38**
		(+2.56)	(+2.02)	(+3.11)	(+3.27)	(+2.14)	(+3.85)	(+3.43)	(+2.00)	(+3.99)	(+3.91)	(+1.93)	(+4.65)

E.2 Point tracking

Correspondence between frames does not necessarily capture how scene points evolve over time. We therefore evaluate point tracking by attaching a CoTracker (Karaev et al., 2025) head to each representation. All variants are trained on Kubric (Greff et al., 2022) for the same number of epochs and evaluated following the protocol of (Son et al., 2025). Table E shows that dynamic feature

Table F: **Test-time pose-alignment results for pose-free methods.** Alignment optimizes each target pose against the ground-truth target image for 2,000 steps, and the green row reports its effect. Best in **bold**.

Method	Alignment	ADT			TUM		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeoVerse (Yang et al., 2026)	✗	12.217	0.2982	0.6391	14.398	0.4159	0.4587
	✓	16.179	0.4384	0.5296	16.120	0.5266	0.3981
		(+3.962)	(+0.1402)	(−0.1095)	(+1.721)	(+0.1107)	(−0.0606)
C4G (Kim et al., 2026c)	✗	12.963	0.4487	0.6870	12.378	0.3935	0.5984
	✓	18.900	0.5813	0.5833	17.508	0.5518	0.4951
		(+5.937)	(+0.1326)	(−0.1038)	(+5.130)	(+0.1583)	(−0.1033)
FL4N (ours)	✗	19.693	0.5948	0.4490	17.847	0.5586	0.3449
	✓	**23.288**	**0.6648**	**0.3683**	**20.392**	**0.6643**	**0.2860**
		(+3.594)	(+0.0700)	(−0.0807)	(+2.545)	(+0.1057)	(−0.0589)

substantially improves tracking of moving points over the static feature, with the performance gap increasing as the temporal stride grows. Consistent with the depth probing results, this widening gap indicates that the residual captures temporal information that is particularly useful under larger object displacement.

F Test-time pose alignment for latent models

Evaluating models that do not receive conditioning poses can introduce scale misalignment with the ground truth, resulting in poor rendering metrics even when the predicted scene is geometrically accurate. Since the input images do not determine an absolute scale, a pose-free model may reconstruct the scene at an arbitrary scale. Pose-free methods therefore perform test-time pose alignment before evaluation and report rendering metrics using the aligned poses (Kim et al., 2026c; Ye et al., 2025). This alignment only brings the prediction into the coordinate frame and scale of the ground truth without modifying the underlying scene content.

The pose-free Gaussian baselines in our quantitative comparisons are evaluated with 2,000 step pose-alignment steps, whereas FL4N is reported without pose alignment in the main tables. However, FL4N is also pose-free and therefore suffers from the same scale ambiguity that motivates this evaluation procedure. We thus additionally evaluate FL4N with the same 2,000 pose alignment to provide a consistent comparison.

For Gaussian models, Gaussians are predicted once from the conditioning images and kept fixed throughout the alignment process. Differentiable rasterization allows the rendering loss to be back-propagated to the camera pose, so each alignment step re-rasterizes the same frozen Gaussians under an updated pose, without modifying the network or scene representation.

The analogous procedure can be done to our latent renderer. The scene latent is predicted solely from the conditioning images and is independent of the rendering process. We therefore compute the latent once and keep it fixed throughout the alignment process, with the renderer serving the role of rasterization by re-rendering the cached latent under an updated target pose at each step. For both Gaussian and latent models, we optimize a rotation and translation offset for each target pose (Ye et al., 2025), initialized at zero such that the initial rendering exactly matches the feed-forward prediction. We use the same objective as the Gaussian baselines, with 2,000 Adam steps using pixel and perceptual losses. Table F reports the effect of this alignment under the NTS protocol on ADT and TUM. Alignment is worth 5.94 dB to C4G on ADT and 5.13 dB on TUM, against 3.59 dB and 2.55 dB for our model, so a large part of the published margin of C4G on these cells comes from fitting the pose to the image it is then scored on. How much alignment recovers therefore measures how much error the supplied cameras carry rather than the quality of the reconstruction, and our model without any alignment remains ahead of both aligned baselines on every metric.

G Limitations

Our approach inherits the limitations of the pretrained static latent NVS backbone, particularly when rendering from extrapolated viewpoints that are unseen in the input views. Latent NVS models can-

not synthesize unseen content (Jin et al., 2025), resulting in blurry or collapsed renderings under extrapolation, as also observed in (Szymanowicz et al., 2026). In contrast, explicit Gaussian representations construct scene content in 3D space, allowing the observed portions of the scene to remain usable even when extrapolation leaves holes in unseen regions.

The residual formulation can compensate for certain errors in the static prediction, but its capacity is naturally limited when the underlying scene representation or camera pose is substantially incorrect. These limitations can be alleviated by using stronger static NVS models with broader scene coverage and more robust viewpoint understanding.



Figure F: **More qualitative comparisons on NVIDIA across all scenes.** Our method shows much better rendering quality than all other methods based on 3D Gaussian.

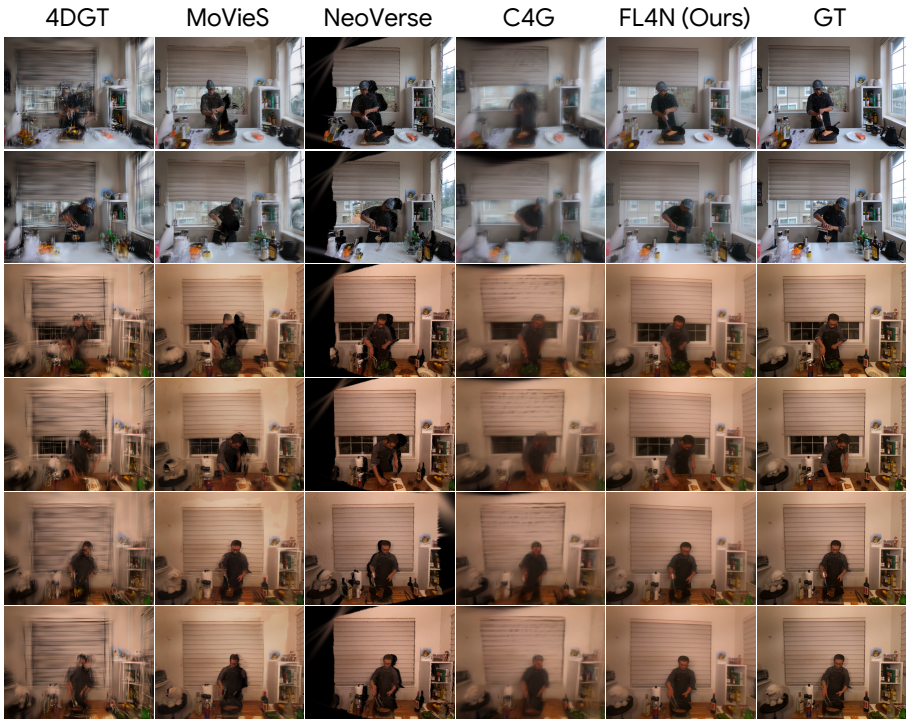

Figure G: **More qualitative comparisons on N3DV across all scenes.** Our method shows much better rendering quality than all other methods based on 3D Gaussian.



Figure H: **More qualitative comparisons on TUM across all scenes.**

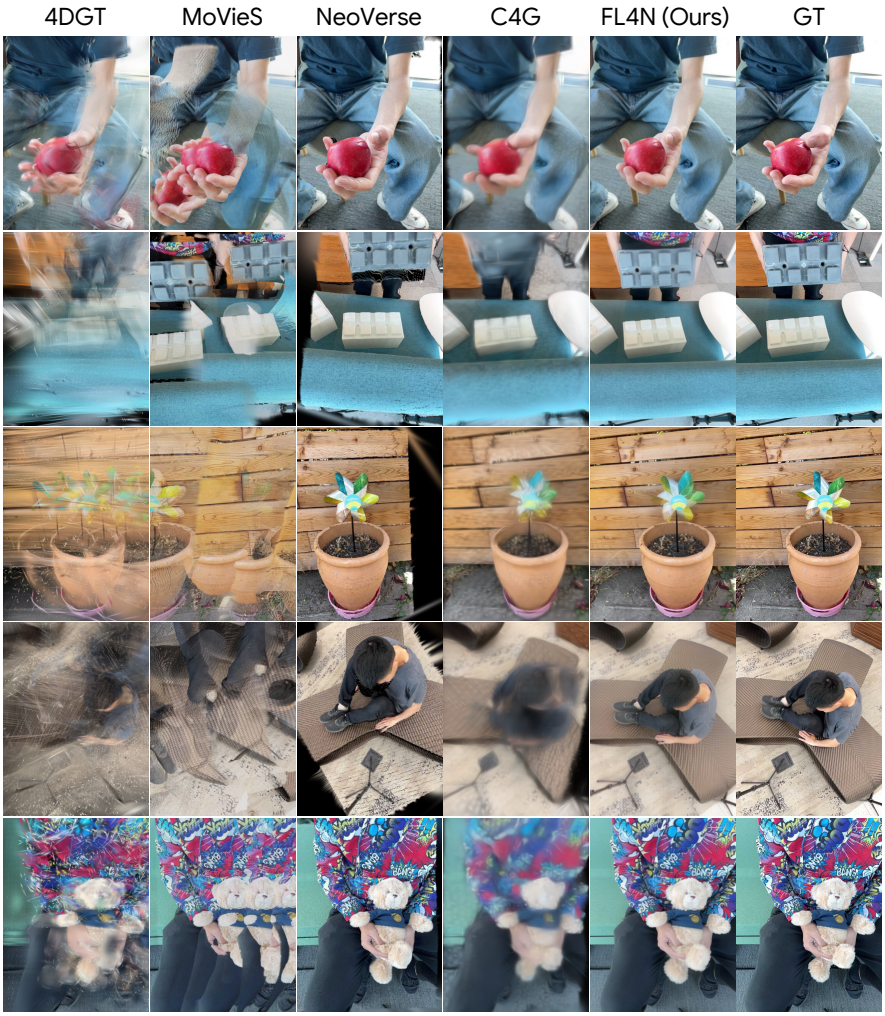

Figure I: **More qualitative comparisons on DyCheck across all scenes.**

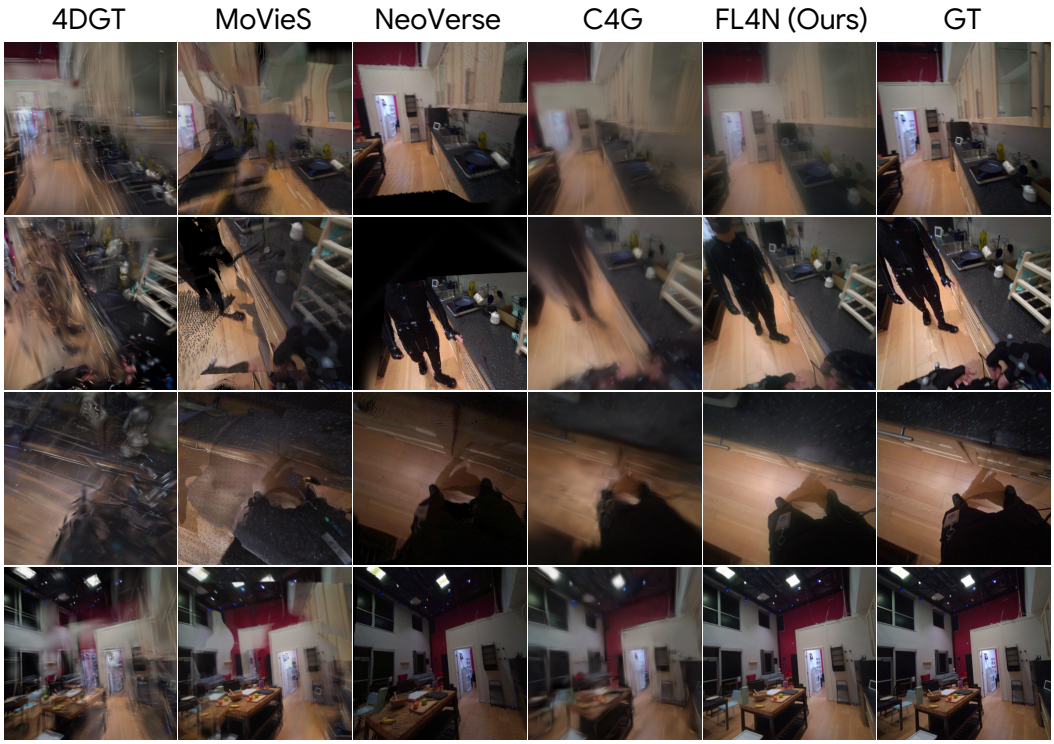

Figure J: **More qualitative comparisons on ADT across all scenes.** Our method shows much better rendering quality than all other methods based on 3D Gaussian. Gaussian-based approaches show severe artifacts such as holes caused by misaligned/mispositioned Gaussians.

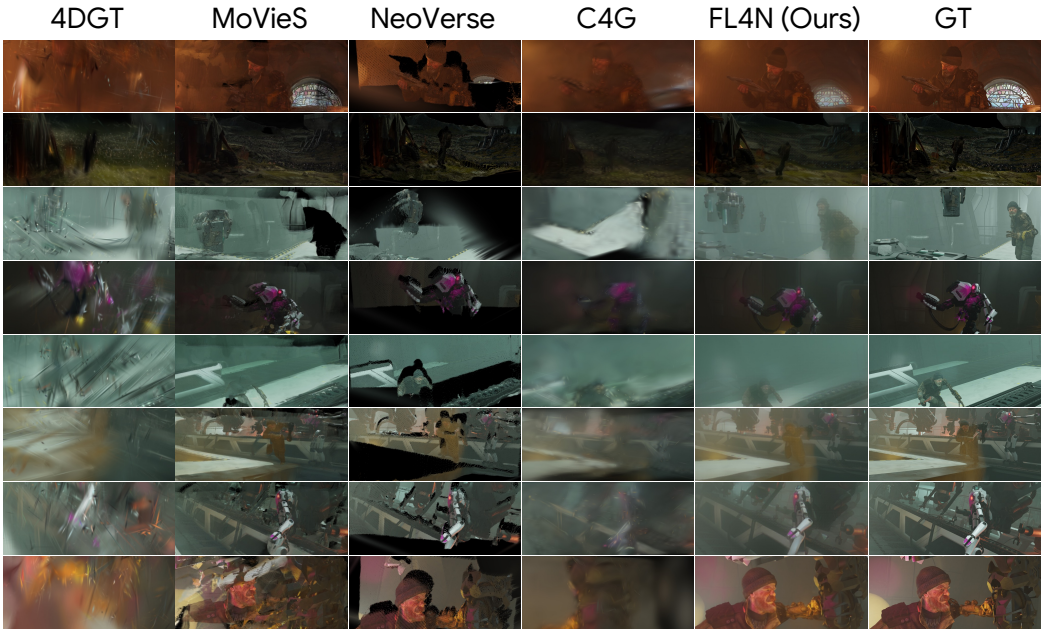

Figure K: **More qualitative comparisons on Charge across all scenes.** Our method shows much better rendering quality than all other methods based on 3D Gaussian. Gaussian-based approaches show severe artifacts such as holes caused by misaligned/mispositioned Gaussians.

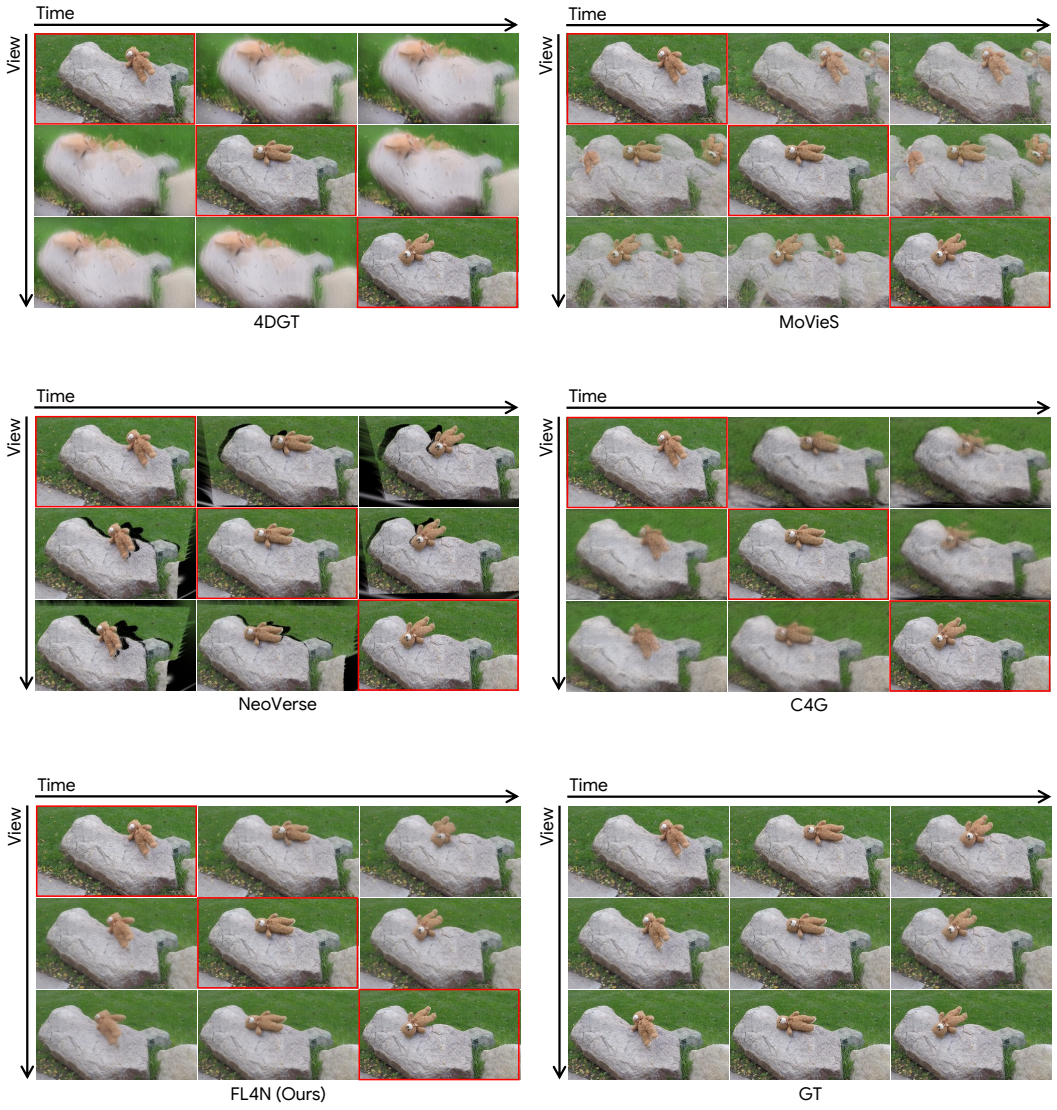

Figure L: **Qualitative comparisons across view-time grid on NVIDIA.** Red outlines indicate context frames. Compared with other methods, our methods show substantially better rendering quality with clear view-time disentanglement.

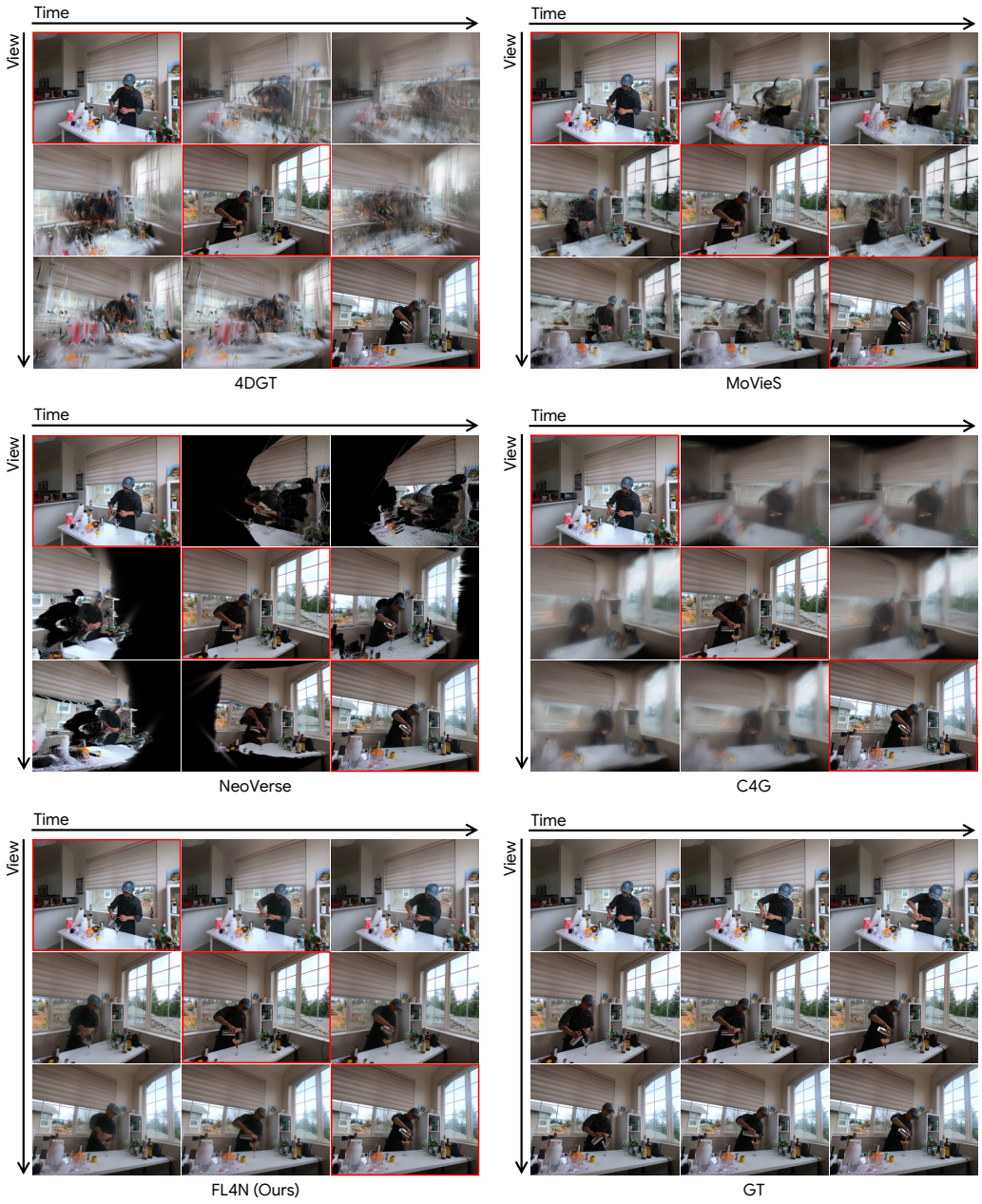

Figure M: **Qualitative comparisons across view-time grid on N3DV.** Red outlines indicate context frames. Compared with other methods, our methods show substantially better rendering quality with clear view-time disentanglement.

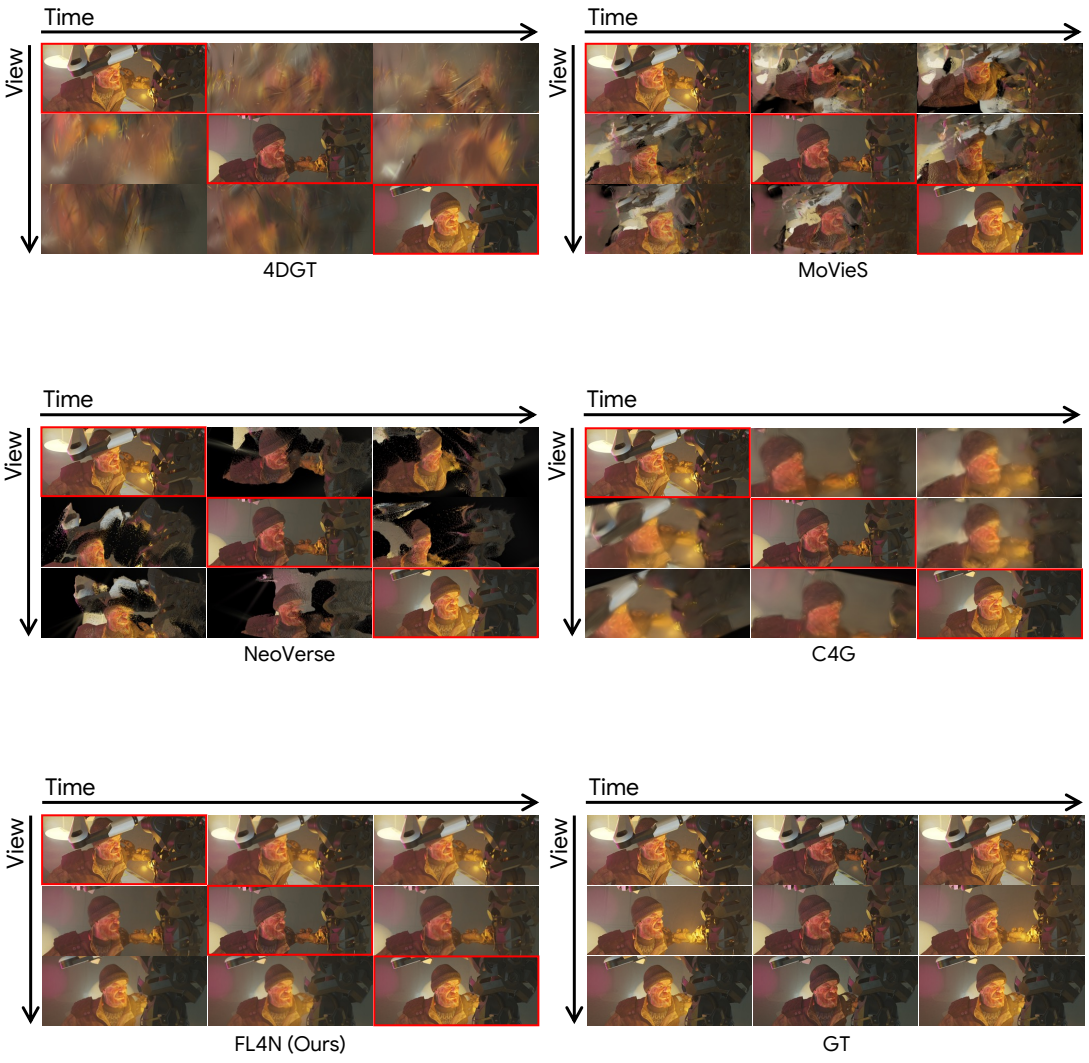

Figure N: **Qualitative comparisons across view-time grid on Charge.** Red outlines indicate context frames. Compared with other methods, our methods show substantially better rendering quality with clear view-time disentanglement.

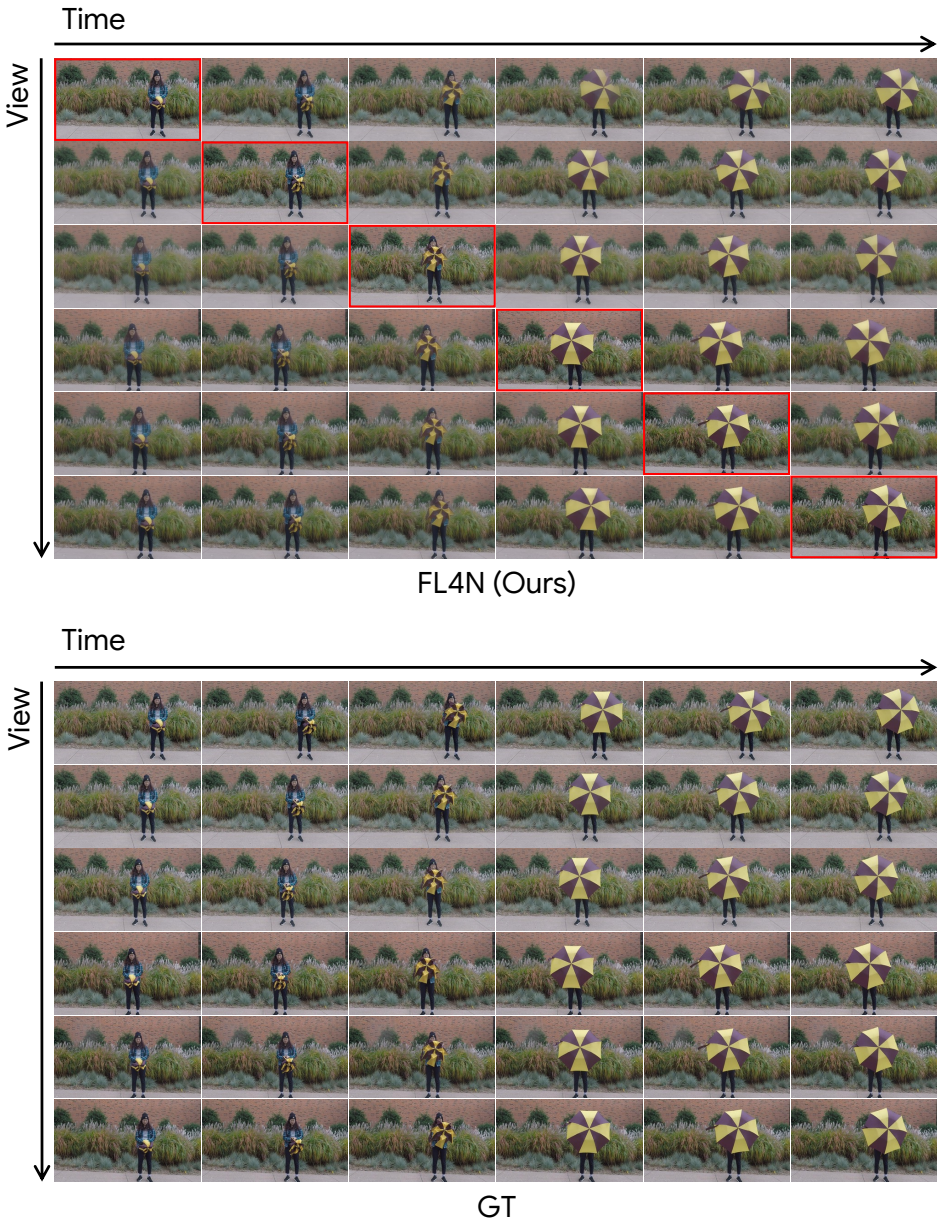

Figure O: **Qualitative results across view-time grid on NVIDIA.** Red outlines indicate context frames. Our method successfully synthesizes images at novel view and time, using target conditioning on predicted latent, without relying on explicit geometric primitives.

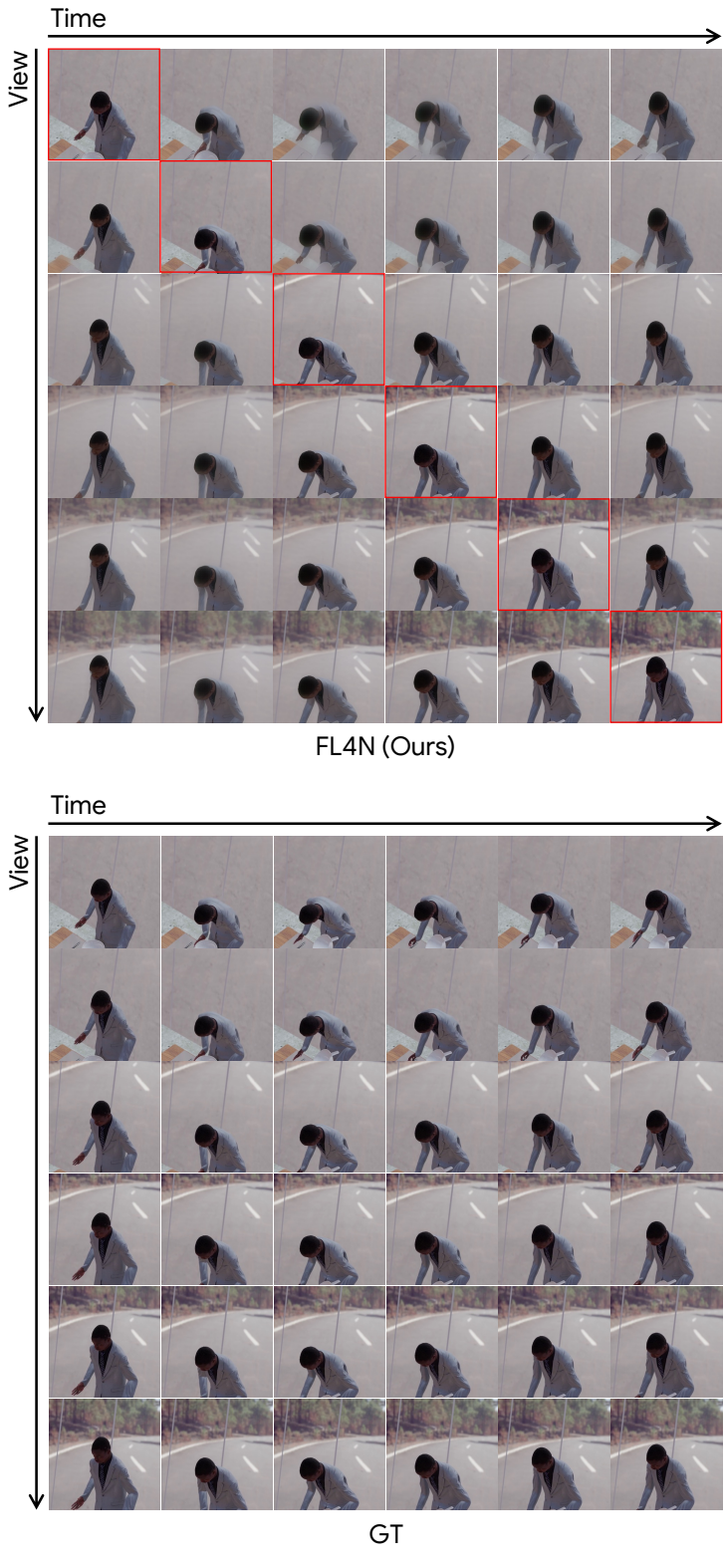

Figure P: **Qualitative results across view-time grid on CamxTime.** Red outlines indicate context frames. Our method successfully synthesizes images at novel view and time, using target conditioning on predicted latent, without relying on explicit geometric primitives.

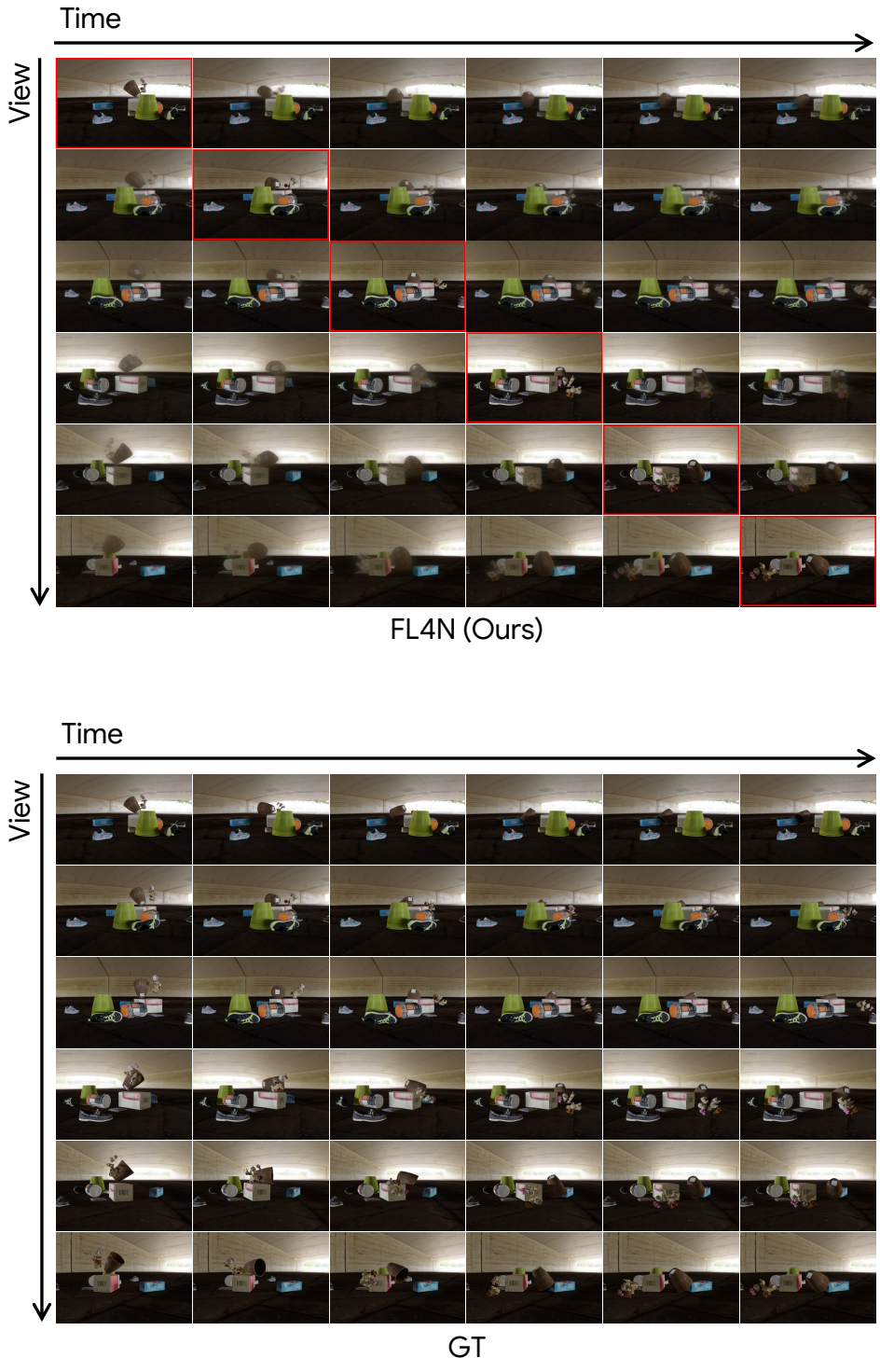

Figure Q: **Qualitative results across view-time grid on Kubric4D.** Red outlines indicate context frames. Our method successfully synthesizes images at novel view and time, using target conditioning on predicted latent, without relying on explicit geometric primitives.