

FillGS: Filling Observation Gaps in 4D Gaussian Splatting via Viewpoint-Time Selection and Generative Refinement

Takashi Otonari[✉] and Toshihiko Yamasaki[✉]

The University of Tokyo, Japan
{otonari,yamasaki}@cvm.t.u-tokyo.ac.jp

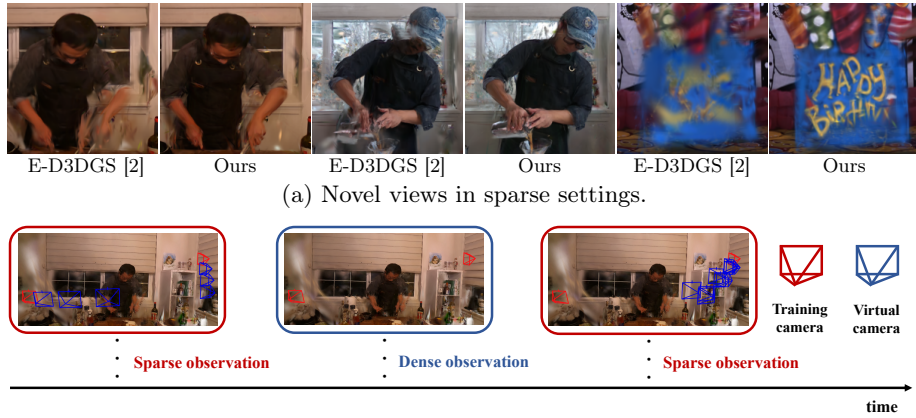
Abstract. 4D Gaussian Splatting (4DGS) can render dynamic scenes photorealistically. However, with limited viewpoint coverage, some spatiotemporal regions remain sparsely observed, leading to artifacts, particularly in scenes with large motion. Existing approaches leveraging generative models rely on heuristic virtual-viewpoint selection before refining rendered views. As a result, they cannot actively explore such sparsely observed regions. To address this issue, we propose a pipeline that actively selects spatiotemporal virtual viewpoints to improve 4DGS reconstruction. Our method selects virtual viewpoints for generative enhancement based on the rendering sensitivity and motion-aware observation density of 4D Gaussians, prioritizing views that alleviate observation sparsity. In the refined images, we filter out regions that conflict with captured observations or are likely to contain generative artifacts and then fine-tune 4DGS using only the reliable regions. We evaluate our method on multi-view video benchmarks using new train/test splits designed to induce observation gaps. Results show consistent improvements over prior viewpoint selection strategies and fine-tuning methods in both qualitative and quantitative evaluations, while reducing artifacts.¹

Keywords: 4D Gaussian Splatting · Novel view synthesis · Active view selection · Diffusion model

1 Introduction

Recent advances in neural rendering, such as Neural Radiance Fields (NeRF) [35], have enabled high-fidelity novel view synthesis from multi-view images. Moreover, 3D Gaussian Splatting (3DGS) [14] and its dynamic extension, 4D Gaussian Splatting (4DGS) [56], use explicit representations to achieve both high rendering quality and fast rendering speed, making these methods promising foundations for real-time rendering of dynamic scenes. These techniques support a wide range of applications, including augmented and virtual reality, gaming, and autonomous driving.

¹ Our project page is available at <https://otonari726.github.io/fillgs/>.



(b) Our virtual-viewpoint selection based on motion-aware observation sparsity.

Fig. 1: 4D Gaussian Splatting (4DGS) suffers from limited spatiotemporal observations. In this paper, we address the challenge of insufficient spatiotemporal observations by actively selecting viewpoints to identify spatiotemporal gaps and then refining the rendered images with a generative model. Our method enables high-fidelity rendering from novel spatiotemporal viewpoints.

However, 4DGS is highly sensitive to the distribution of observed viewpoints, and thus often exhibits artifacts in sparsely observed regions, as shown in Fig. 1(a). When each time step is captured by only a few cameras, fast-moving objects are seen at a given location only briefly; as a result, the corresponding 4D Gaussians are weakly constrained. In real-world captures, where viewpoint diversity is often limited, this effect is further amplified, leading to degraded reconstruction quality.

Recent studies have aimed to improve 4D novel view synthesis by leveraging generative models to refine rendered images, thereby compensating for sparse observations [15, 17]. However, the effectiveness of such enhanced images critically depends on the choice of virtual viewpoints used for generation; existing methods do not actively select them. Instead, they typically rely on heuristic strategies that give little consideration to the downstream effects of the generated views, such as interpolating between training viewpoints or uniformly sampling from the viewpoint space. In static scenes, selecting virtual viewpoints to maximize spatial coverage has been shown to be effective [16]. However, extending this idea to dynamic scenes is fundamentally challenging because sparsity evolves over time. As motion continuously changes which regions are sparsely observed, observation sparsity is inherently spatiotemporal rather than purely spatial. Consequently, virtual viewpoints chosen solely to maximize spatial coverage fail to target sparsely observed spatiotemporal regions, especially with large motion. Accordingly, a key open problem in dynamic-scene settings is how to select spatiotemporal viewpoints for refinement by generative models.

In this paper, we propose FillGS, which actively selects spatiotemporal virtual viewpoints for the generative enhancement of 4DGS-rendered views, as shown in Fig. 1(b). FillGS improves rendering quality under sparse-observation settings without relying on hand-crafted heuristics, such as interpolation or uniform sampling, that fail to account for the effects of generative enhancement. We define a motion-aware observation density to identify sparsely observed regions, and we prioritize spatiotemporal regions where the 4D Gaussians are under-constrained due to sparse observations. Furthermore, we do not treat generated images as equivalent to real observations. We exclude (i) regions that are already sufficiently observed and (ii) regions likely to contain generative artifacts that are inconsistent with real observations. We then fine-tune 4DGS using only reliable pixels. As a result, FillGS enables stable improvement in sparsely observed regions without degrading already well-reconstructed regions.

To validate the proposed method, we create new training and evaluation splits from existing multi-view video datasets, thereby enabling evaluation in sparse-observation settings. Across multiple datasets under sparse-view settings, our method consistently outperforms existing viewpoint selection strategies and fine-tuning methods both qualitatively and quantitatively.

2 Related Work

2.1 4D Novel View Synthesis

Dynamic novel view synthesis (NVS) is considerably more challenging than static NVS because variations in space and time occur simultaneously. Among dynamic NeRF methods, many frameworks have been proposed that represent space and time as continuous functions and model scene motion via deformation fields, flow fields, or mappings to a canonical space [4, 25, 26, 39–41, 48, 59]. Although these approaches achieve high-quality rendering, they incur significant optimization overhead. In addition, limited viewpoint coverage exacerbates artifacts in sparsely observed regions.

In recent years, 3D Gaussian Splatting (3DGS) [14] has emerged as a promising approach that achieves both high visual quality and fast rendering through an explicit point-based scene representation. Building on this line of work, 4D Gaussian Splatting (4DGS), which extends 3DGS to dynamic scenes, has advanced rapidly [2, 5, 13, 19, 24, 28, 33, 56, 62, 63]. Furthermore, subsequent studies have addressed challenging observation settings, such as monocular or casually captured videos [8, 15, 20, 27, 31, 46, 49, 50].

In NVS for dynamic scenes under sparse-view settings, artifacts caused by insufficient observations and inconsistencies in spatiotemporal geometric reconstruction become more pronounced. Consequently, many studies have proposed various priors and regularization techniques to mitigate these issues [1, 17, 22, 31, 45]. In research closely related to ours, UA-4DGS [15] and GS-GS [17] augment training data by using diffusion models to refine rendered images from unobserved viewpoints, compensating for sparse observations in monocular or

sparse-view settings. However, the virtual viewpoints used to refine these rendered images are chosen via random sampling or interpolation between training views. As a result, these methods do not perform active virtual-viewpoint selection targeting spatiotemporal regions that are likely to improve reconstruction quality.

2.2 View Selection for Applying Generative Priors in Novel View Synthesis

Recent works have improved 3DGS under sparse observations by leveraging diffusion models as generative priors. In these approaches, a diffusion model refines rendered images, and the refined outputs are used as pseudo-observations to further fine-tune the reconstruction model. However, little attention has been paid to which viewpoints are effective for refining images. Existing approaches rely on heuristics such as interpolating between training views [32, 36, 64], optimizing for target viewpoints [3, 57, 58], manually selecting virtual viewpoints for generation [54], expanding the field of view at training viewpoints [10], and interpolating time in dynamic scenes [61]. These designs largely ignore the viewpoint-dependent effectiveness of generated images for improving reconstruction.

In contrast, ExploreGS [16] improves static reconstruction by selecting virtual viewpoints to maximize spatial coverage under sparse observations. However, dynamic scenes require spatiotemporal exploration, where observations are temporally sparse. Consequently, spatial-coverage-based selection may overlook viewpoints that are informative for motion reconstruction, thereby reducing motion estimation accuracy and degrading 4D reconstruction quality.

2.3 Active View Selection for Data Acquisition

To achieve high-fidelity NVS from a small number of input images, active view selection (AVS), which aims to identify informative camera views or view subsets, has been widely studied for both NeRF and 3DGS. Existing frameworks select viewpoints expected to yield the greatest improvement using criteria such as reconstruction uncertainty, information gain, scene coverage, or Fisher information [7, 11, 18, 38, 43, 44, 47, 51, 53, 55, 60]. Moreover, in robotics and online mapping, recent works have integrated active reconstruction with path planning based on NeRF or 3DGS, enabling efficient viewpoint acquisition under sensing and motion constraints [12, 23, 30, 34, 65]. However, many AVS methods assume that additional real observations can be obtained by capturing new images. In contrast, FillGS actively selects virtual viewpoints to refine the corresponding rendered views using a generative model over a broad spatiotemporal domain, including past time steps that are no longer physically observable.

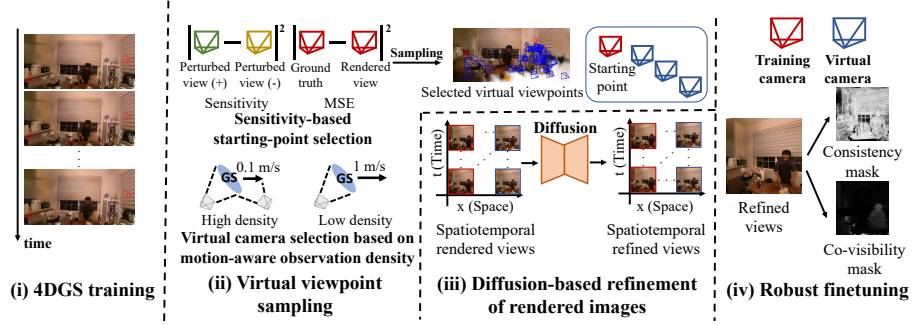


Fig. 2: Overview of our FillGS pipeline. (i) We first train a 4DGS model. (ii) We then actively select virtual viewpoints. To do so, we first identify training views that are locally unstable due to sparse observations (Sec. 3.2), and then generate candidate virtual viewpoints using both local and global strategies. Each candidate is scored using a motion-aware observation-density metric, and we prioritize candidates with low observation density (Sec. 3.3). (iii) The rendered images from the selected virtual viewpoints are refined using a diffusion model (Sec. 3.4). (iv) Finally, we assign weights to the refined images based on their conflicts with the training images and generation errors, and use these weights to fine-tune the 4DGS model (Sec. 3.5).

3 Method

3.1 Overview

When 4DGS is trained from a limited set of viewpoints, sparsely observed spatiotemporal regions become under-constrained, often leading to rendering artifacts. We address this by using images refined by a video diffusion model as auxiliary supervision. Our pipeline actively selects spatiotemporal virtual viewpoints that cover under-observed regions, refines the rendered images at these viewpoints, and robustly fine-tunes 4DGS while suppressing generative artifacts and conflicts with real observations. Fig. 2 summarizes the overall pipeline.

Uniformly sampling unobserved viewpoints does not necessarily improve the conditioning of 4DGS optimization. Instead, virtual viewpoints should be prioritized according to how well they observe regions with insufficient observation support. We therefore perform viewpoint selection in two stages. First, we select an observed spatiotemporal viewpoint where the trained 4DGS exhibits high local sensitivity, which serves as a starting point for exploration (Sec. 3.2). Second, we select virtual viewpoints that observe many 4D Gaussians with low motion-aware observation density (Sec. 3.3).

Since the goal of 4D NVS is to remain consistent with real observations, generated images are used only as auxiliary supervision for under-constrained regions. During fine-tuning, we down-weight losses from generated images in regions that are either unreliable or already well supported by observed views, using generation-error detection and motion-aware co-visibility weighting (Sec. 3.5). This allows our method to improve sparsely observed regions without degrading well-reconstructed areas.

3.2 Sensitivity-Based Starting-Viewpoint Selection Strategy

We select observed spatiotemporal viewpoints as starting points for exploration by identifying viewpoints where the trained 4DGS exhibits local instability. These viewpoints provide useful starting points for subsequent exploration.

Why sensitivity indicates good starting points. We aim to select observed viewpoints whose local neighborhoods are likely to be unstable as starting points for exploration. However, because ground-truth images are unavailable for nearby unobserved views, we cannot directly evaluate reconstruction errors in these neighborhoods. We therefore approximate error-prone neighborhoods by measuring the sensitivity of the rendered output to small perturbations, using local variation of $\hat{I}_\theta$ as a surrogate uncertainty signal. Such weakly constrained regions often arise from insufficiently observed 4D Gaussians, where small changes in viewpoint or time can alter contributing Gaussians and visibility, causing abrupt rendering changes. Since under-constrained regions may also yield large errors even at observed views, we additionally use the squared error at observed viewpoints as an auxiliary indicator. We prioritize observed viewpoints with high error and large perturbation-induced fluctuations as starting points. We provide an analysis of this score on synthetic data in the supplementary materials.

Score definition. For an observed viewpoint v , we compute an exploration priority score

$$\mathcal{E}(v) = \mathcal{L}(v) + \mathcal{V}(v), \quad (1)$$

and prioritize viewpoints with higher $\mathcal{E}(v)$ as exploration starting points. Here, $\mathcal{L}(v)$ measures the normalized reconstruction error of the observed image, and $\mathcal{V}(v)$ measures the normalized local variation of the rendered image under small perturbations. Specifically, we define

$$\mathcal{L}(v) = \frac{1}{\bar{\mathcal{L}}} \left\| \hat{I}_\theta(v) - I(v) \right\|^2, \quad (2)$$

where $\bar{\mathcal{L}}$ is a normalization constant chosen such that the dataset-level mean of $\mathcal{L}(v)$ is 1. This normalization ensures that $\mathcal{L}(v)$ is properly scaled and directly comparable to $\mathcal{V}(v)$. For the local variation term, we define

$$\mathcal{V}(v) = \sum_{\delta \in \mathcal{D}} w_\delta \left\| \hat{I}_\theta(v \oplus \delta) - \hat{I}_\theta(v \ominus \delta) \right\|^2, \quad (3)$$

where $\mathcal{D}$ denotes a set of small perturbations (*e.g.*, three translations, three rotations, and one temporal shift), and $v \oplus \delta$ ($v \ominus \delta$) represents the viewpoint obtained by applying perturbation δ to viewpoint v in the positive (negative) direction. The weights w_δ are normalized across perturbations so that the dataset-level average contribution of each δ in the score is $1/|\mathcal{D}|$. As a result, the total over all $\delta \in \mathcal{D}$ averages to 1, matching the scale of $\mathcal{L}(v)$.

Connection to uncertainty estimation. The proposed sensitivity term $\mathcal{V}(v)$ can be interpreted as a finite-difference approximation to a local Lipschitz constant of the rendering function with respect to viewpoint perturbations. Regions

that are weakly constrained by observations tend to exhibit higher sensitivity to small pose or temporal changes, as minor perturbations may alter Gaussian visibilities and the blending weights. Under such conditions, the rendering function is locally unstable, and this local instability tends to correlate with epistemic uncertainty in the reconstruction. Therefore, the sensitivity score serves as a practical proxy for identifying regions with high uncertainty in the reconstruction without requiring explicit probabilistic modeling.

3.3 Viewpoint Scoring Based on Motion-Aware Observation Density

In this paper, we employ two strategies to generate candidate virtual viewpoints. The first is a local strategy that searches within the neighborhood of the observed viewpoints. The second is a global strategy that samples candidate goal viewpoints directly. We then compute a motion-aware observation density based on the velocity of each 4D Gaussian and the number of observations. We assign higher scores to candidates that observe Gaussians with lower observation densities and select the highest-scoring candidate.

Candidate view creation. We generate candidate virtual viewpoints to refine rendered images using (i) a local strategy that perturbs the current pose and (ii) a global strategy that samples a goal viewpoint and interpolates a trajectory from the start to that viewpoint, thereby covering regions that are difficult to reach through local exploration.

In the local approach, we generate candidate virtual viewpoints in the neighborhood of the current viewpoint by applying six translational and four rotational perturbations. In addition, we select two training viewpoints near the current viewpoint and include an interpolated viewpoint between the current viewpoint and each selected training viewpoint. In total, we obtain 12 candidate viewpoints. By incorporating interpolation with training viewpoints, the candidate set not only explores the local neighborhood but also fills gaps toward the training viewpoints, leveraging the benefits of prior interpolation-based viewpoint selection methods.

In the global approach, we create a bounding box computed from the selected starting viewpoint and its neighboring observed viewpoint at that timestamp. We then expand this bounding box by the maximum displacement reachable by the local strategy, which we calculate from the number of steps and the local search trajectory length, and uniformly sample 12 candidate viewpoints within the expanded box. For each sampled viewpoint, we compute the viewing direction by interpolating between the two neighboring observed viewpoints associated with the starting viewpoint’s timestamp. By selecting candidates in regions that are difficult to reach by the local strategy, we improve global coverage. Once a global viewpoint is selected, we re-select, at the same timestamp, the training camera closest to it and construct a camera trajectory by interpolating between the selected global viewpoint and this nearest training camera. We adopt this interpolated path as the virtual-camera trajectory.

Viewpoint scoring. We assign scores to the viewpoint candidates generated by the local and global strategies, and select the highest-scoring candidate as a virtual viewpoint to be used as input to a generative model. First, for each 4D Gaussian g_i , we construct an observation-count table $C_{g_i}(t)$ over time t , which indicates how many times the 4D Gaussian is observed at each time t . In general, a Gaussian with a large $C_{g_i}(t)$ can be regarded as having a high observation density, whereas a Gaussian with a small $C_{g_i}(t)$ can be regarded as having a low observation density. However, while the background tends to be observed consistently across all times, dynamic objects may become invisible at certain time steps or undergo substantial appearance changes, and thus at a given spatial location they are often observable only at specific times. For this reason, simple observation counts alone cannot adequately reflect the degree of motion-aware observation sparsity.

Therefore, we define a motion-aware observation-deficiency score that captures how sparsely each Gaussian is observed while accounting for its motion. Specifically, let the magnitude of the velocity of Gaussian g_i at time t be $s_{g_i}(t)$, and define the observation deficiency $D_{g_i}(t)$, which corresponds to the inverse of the observation density, as

$$D_{g_i}(t) = \frac{s_{g_i}(t)}{1 + C_{g_i}(t)}. \quad (4)$$

According to Eq. (4), a Gaussian with few observations and high speed receives a high deficiency score, indicating that it is insufficiently constrained by the captured views. In contrast, a Gaussian with many observations and low speed is assigned a low deficiency score, indicating that it is already sufficiently constrained by the captured views. For each candidate viewpoint, we render $D_{g_i}(t)$ in place of the color of each Gaussian, and take the mean pixel value of the resulting image as the candidate’s score. That is, a candidate that observes more Gaussians with higher observation-deficiency scores receives a higher score, enabling the selection of highly informative viewpoints.

Relation to information gain. From an active-view-selection perspective, $D_{g_i}(t)$ can be interpreted as a proxy for the information gained by observing Gaussian g_i at time t . Gaussians exhibiting large motion while being infrequently observed tend to be less constrained, so additional observations are expected to reduce parameter uncertainty. By prioritizing viewpoints that cover such Gaussians, our scoring strategy aims to alleviate reconstruction ambiguity.

3.4 Video Diffusion Model

We leverage a video diffusion model to refine images rendered from virtual viewpoints, which often exhibit artifacts when observations are sparse. The model builds on 3DGS-Enhancer [32] and extends it to handle dynamic scenes. Specifically, we use spatiotemporal patches as the model’s input and output. Each patch is formed by concatenating 3 consecutive frames in time and 4 views along the spatial dimension. To construct a spatiotemporal patch, we start with an observed image, append rendered images from the selected virtual viewpoints, and

extend the sequence to adjacent time steps in both temporal directions. As additional conditioning inputs, we provide the camera parameters and timestamps corresponding to each observed and rendered frame. Conditioned on these inputs, the model predicts a refined image sequence with completed content and improved spatiotemporal consistency.

Training follows the procedure of 3DGS-Enhancer. We initialize the model from weights pretrained on the DL3DV dataset [29] and then further fine-tune it on a dataset of dynamic scenes. Details of the dataset and the diffusion training procedure are provided in the supplementary materials.

3.5 Robust 4D Model Fine-tuning Using Generated Views

We propose a robust fine-tuning strategy that preserves observation consistency in 4DGS. Specifically, we modulate the loss using (i) matching-based consistency masks and (ii) motion-aware co-visibility masks, so that fine-tuning concentrates on non-conflicting, under-constrained regions.

Consistency mask driven by matching confidence. To assess the reliability of the generated images, we use RoMa v2 [6] to compute geometric consistency between the observed and generated images, producing a per-pixel confidence map. For each generated image, we match it to its two adjacent observed views and take the maximum of the two confidence scores for each pixel. This assigns higher weight to regions that are observed from at least one neighboring view. This confidence map can be interpreted as the likelihood that a generated pixel is consistent with the observed views. We fine-tune the 4DGS model using the consistency mask to suppress the influence of incorrectly generated regions.

Co-visibility mask with motion-aware observation-density weighting. When fine-tuning a model using generated images, it is important not to overly modify regions that are already supported by existing observations. Existing approaches for incorporating generated images into dynamic scenes either render per-view uncertainty maps from static per-Gaussian uncertainty values [15] or apply the same photometric loss weights as those used for real images [17]. These weighting strategies do not explicitly account for dynamic components in the scene.

To address this issue, we leverage the motion-aware observation density of 4D Gaussians to adaptively assign weights to generated images. Specifically, we assign small weights to regions with high observation density that are well-constrained by existing observations, thereby suppressing updates, while assigning large weights to sparsely observed regions, thereby strengthening the constraint imposed by generated images. These weights are obtained by normalizing the observation deficiency score $D_{g_i}(t)$, which is also used for virtual-viewpoint selection. Consequently, we achieve stable model fine-tuning that exploits the benefits of generated images without degrading observation-consistent reconstructions.



Fig. 3: Our sparse-view camera split. We train on two cameras and evaluate on spatially separated views under both interpolation and extrapolation splits. Red cameras indicate training cameras, and blue cameras indicate testing cameras.

4 Experiments

4.1 Experimental Settings

Implementation details. We adopt E-D3DGS [2] as our baseline, initializing it according to the original training configuration. The model is then fine-tuned for 30,000 iterations; at 6,000-iteration intervals, we generate 20 virtual trajectories and incorporate them into the fine-tuning process. All experiments are conducted using a single NVIDIA A100 GPU with either 40 GB or 80 GB of memory. Additional details are provided in the supplementary materials.

Evaluation metrics. We evaluate rendering quality using standard image-level metrics, including PSNR, SSIM [52], and LPIPS [66]. To better capture perceptual and semantic consistency, we additionally report (i) FID [9], which measures the distance between the feature distributions of rendered and ground-truth images, and (ii) the cosine similarity between DINOv2 [37] features extracted from rendered and ground-truth images.

Baselines. To fairly evaluate the proposed pipeline for virtual-viewpoint selection and model fine-tuning, we control for all factors other than the viewpoint selection strategy, fine-tuning method, and the use of generated images. Concretely, we adopt E-D3DGS [2] as a common backbone for all methods. When generated images are introduced during training, we fix both the generative model and the number of generated images. For the baseline methods, the weighting scheme for incorporating generated images follows ExploreGS [16] and is computed using G-IoU and LPIPS. Under this setup, we compare the following viewpoint selection strategies proposed in prior work: (i) **E-D3DGS** [2], which trains without any generated images and uses only the original training images; (ii) **Interpolation-based selection** [17], which restricts virtual viewpoints to interpolations of training views; (iii) **FisherRF-based selection** [11] using Fisher information scores; (iv) **Coverage-based (3D) selection** [16], which follows ExploreGS for virtual-viewpoint selection by maintaining a single coverage map shared across all time steps and prioritizing views that observe more 4D Gaussians from novel directions; and (v) **Coverage-based (4D) selection** [16], which extends ExploreGS viewpoint selection to 4D by maintaining a separate coverage map for each time step, thereby prioritizing views that observe more 4D Gaussians from novel directions without considering coverage at other time steps and maximizing per-time-step spatial coverage.

	Interpolation					Extrapolation				
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	DINOv2 $\uparrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	DINOv2 $\uparrow$
E-D3DGS [2]	20.42	0.775	0.293	129.33	0.871	14.68	0.723	0.363	160.01	0.832
Interpolation-based selection [17]	20.75	0.777	0.293	120.08	0.896	14.75	0.716	0.370	157.52	0.826
FisherRF-based selection [11]	21.01	0.774	0.303	135.36	0.853	15.59	0.716	0.373	180.27	0.800
Coverage-based (3D) selection [16]	19.91	0.754	0.317	148.68	0.878	15.30	0.700	0.388	175.53	0.809
Coverage-based (4D) selection [16]	19.09	0.737	0.325	140.19	0.866	15.51	0.681	0.407	182.99	0.792
Ours	21.73	0.792	0.256	83.81	0.920	16.24	0.740	0.312	137.32	0.855

Table 1: Quantitative comparison on Neural 3D Video Dataset [21].

	Interpolation					Extrapolation				
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	DINOv2 $\uparrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	DINOv2 $\uparrow$
E-D3DGS [2]	19.63	0.606	0.378	137.86	0.736	16.10	0.690	0.341	131.48	0.785
Interpolation-based selection [17]	19.90	0.612	0.376	115.24	0.768	16.88	0.701	0.338	123.77	0.786
FisherRF-based selection [11]	20.08	0.614	0.421	152.56	0.686	18.32	0.646	0.444	163.51	0.636
Coverage-based (3D) selection [16]	20.12	0.622	0.370	106.63	0.770	17.61	0.710	0.332	110.54	0.805
Coverage-based (4D) selection [16]	20.03	0.619	0.373	110.07	0.769	17.68	0.717	0.329	106.47	0.811
Ours	22.55	0.748	0.250	73.44	0.866	19.88	0.720	0.315	106.45	0.821

Table 2: Quantitative comparison on Technicolor Dataset [42].

Datasets. Most dynamic novel view synthesis datasets are captured with synchronized multi-camera rigs, and benchmarks typically sample test views between densely spaced training cameras, which mainly evaluates interpolation. This is inadequate for sparse-view robustness, where test viewpoints lie far from the training trajectory. We therefore introduce new benchmark splits from 6 scenes in Neural 3D Video [21] and 5 scenes in Technicolor [42]. We train on only two cameras and evaluate on two test sets: interpolation views between the training cameras and extrapolation views outside their trajectory (Fig. 3). This setting better measures generalization to sparsely observed regions.

4.2 Results

Quantitative comparison. Quantitative results on the Neural 3D Video [21] and Technicolor [42] datasets are reported in Table 1 and Table 2, respectively. Our method improves performance across all evaluation metrics and remains robust under both interpolation and extrapolation. While prior methods often focus on either interpolation between training views or extrapolation beyond them to expand viewpoint coverage, our approach performs well in both settings. Additional experiments using other datasets, generative models, and backbones are provided in the supplementary materials.

Qualitative comparison. Fig. 4 and Fig. 5 show qualitative comparisons on the Neural 3D Video [21] and Technicolor [42] datasets. In all rows of both figures, the proposed method effectively restores fine details in dynamic regions, such as thin structures and motion boundaries (Fig. 4 top and second rows; Fig. 5 top and second rows), while simultaneously mitigating background artifacts (Fig. 4 bottom row; Fig. 5 bottom row). In contrast, prior approaches often exhibit oversmoothing in fast-moving areas and residual artifacts in static background regions. These results indicate that our model enables more accurate reconstruction of geometry and appearance under challenging motion and artifact-prone

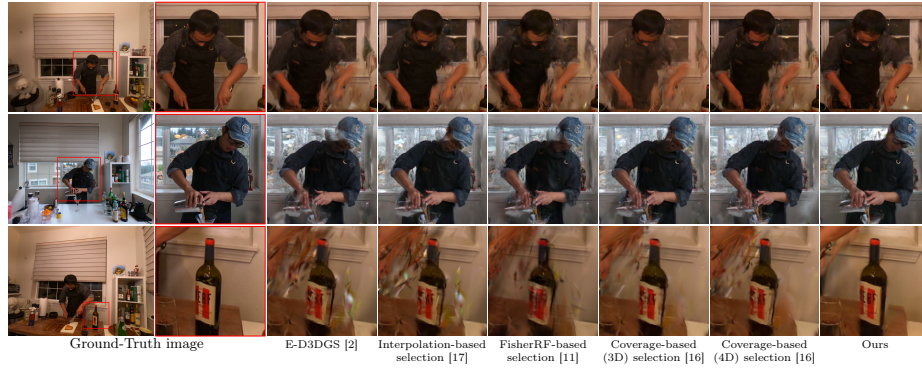


Fig. 4: Qualitative comparison on Neural 3D Video Dataset [21].

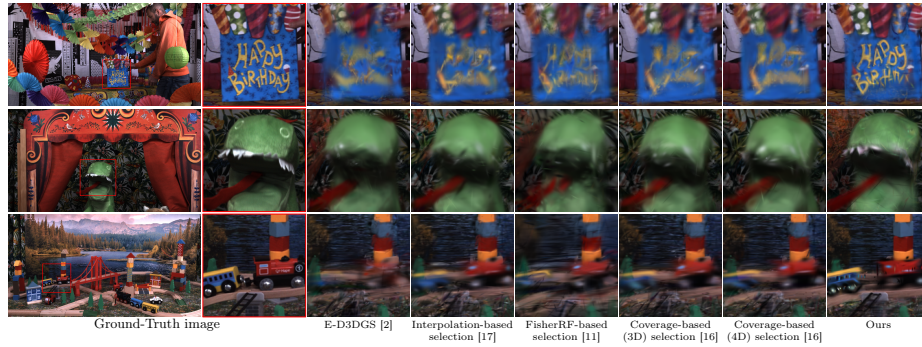


Fig. 5: Qualitative comparison on Technicolor Dataset [42].

conditions. Overall, the proposed method consistently achieves higher visual fidelity than existing methods. Additional visualizations of the failure cases are provided in the supplementary materials.

4.3 Ablation Study

Viewpoint selection. To assess the effectiveness of our virtual-viewpoint selection, we compare it with several baseline strategies. The results on the Neural 3D Video dataset [21] are shown in Table 3. Random selection chooses virtual viewpoints uniformly at random as in UA-4DGS [15]. Negative selection inverts the score used in our method. The interpolation-based selection samples virtual viewpoints between training views. FisherRF-based selection, following FisherRF [11], selects viewpoints to maximize the Fisher information score. The coverage-based selection, following ExploreGS [16], selects viewpoints to maximize spatial coverage. Our method consistently outperforms these baselines. Moreover, the performance drop with negative selection suggests that our score is a meaningful indicator. Restricting virtual-viewpoint candidate creation to local strategies limits global exploration, reduces viewpoint diversity, and conse-

	Interpolation					Extrapolation				
	PSNR↑	SSIM↑	LPIPS↓	FID↓	DINOv2↑	PSNR↑	SSIM↑	LPIPS↓	FID↓	DINOv2↑
Ours	21.73	0.792	0.256	83.81	0.920	16.24	0.740	0.312	137.32	0.855
<i>Our starting-viewpoint selection</i>										
w/o starting-viewpoint selection	21.06	0.773	0.286	126.78	0.880	15.75	0.705	0.393	144.98	0.822
w/o sensitivity-based selection	21.24	0.763	0.279	100.77	0.922	16.01	0.714	0.379	146.25	0.826
w/o error-based selection	21.52	0.766	0.275	100.14	0.919	16.22	0.711	0.377	142.81	0.830
<i>Our candidate-viewpoint selection</i>										
w/o candidate-viewpoint selection	21.13	0.778	0.282	113.72	0.902	15.07	0.650	0.391	147.03	0.825
w/o observation-count-based selection	20.55	0.768	0.294	135.20	0.874	15.66	0.703	0.346	150.63	0.833
w/o speed-based selection	21.06	0.761	0.283	106.35	0.914	15.79	0.703	0.350	146.81	0.823
w/o global candidate view creation	21.39	0.778	0.283	116.88	0.894	15.85	0.701	0.353	150.18	0.830
<i>Other selection strategy</i>										
Random selection [15]	20.59	0.762	0.295	130.43	0.869	15.07	0.701	0.384	172.37	0.811
Negative selection	20.31	0.765	0.294	129.30	0.866	15.06	0.700	0.379	175.95	0.808
Interpolation-based selection [17]	21.15	0.776	0.283	120.52	0.888	15.30	0.703	0.386	173.37	0.817
FisherRF-based selection [11]	20.27	0.761	0.302	139.41	0.855	15.21	0.709	0.363	157.92	0.823
Coverage-based (3D) selection [16]	20.94	0.770	0.290	133.28	0.870	15.51	0.705	0.384	171.75	0.820
Coverage-based (4D) selection [16]	20.62	0.771	0.280	131.06	0.895	15.69	0.700	0.388	167.36	0.819

Table 3: An ablation study of our viewpoint selection on Neural 3D Video Dataset [21].

	Interpolation					Extrapolation				
	PSNR↑	SSIM↑	LPIPS↓	FID↓	DINOv2↑	PSNR↑	SSIM↑	LPIPS↓	FID↓	DINOv2↑
Ours	21.73	0.792	0.256	83.81	0.920	16.24	0.740	0.312	137.32	0.855
<i>Our fine-tuning method</i>										
w/o weight	20.05	0.716	0.334	128.57	0.891	14.70	0.639	0.409	167.40	0.829
w/o consistency mask	21.26	0.778	0.283	107.96	0.897	15.81	0.705	0.355	149.50	0.833
w/o co-visibility mask	21.96	0.780	0.270	98.43	0.912	15.93	0.700	0.359	153.95	0.839
<i>Other weighting method</i>										
3DGS-Enhancer [32]	21.41	0.763	0.281	100.11	0.910	15.48	0.694	0.374	156.29	0.811
ExploreGS [16]	20.97	0.770	0.287	109.74	0.895	15.14	0.697	0.353	151.38	0.820
UA-4DGS [15]	19.96	0.745	0.322	135.11	0.847	14.69	0.689	0.419	201.89	0.772

Table 4: An ablation study of our fine-tuning methods on Neural 3D Video Dataset [21].

quently degrades accuracy. In addition, both starting-viewpoint and candidate-viewpoint selection are important. Removing either component reduces the likelihood of sampling informative viewpoints, leading to lower accuracy. Furthermore, the error and sensitivity scores used for starting-viewpoint selection, as well as the observation count and speed used for candidate-viewpoint selection, are all important; removing any of them degrades performance. Additional experiments on the cost–accuracy trade-off of viewpoint selection are provided in the supplementary materials.

In Fig. 6, we visualize the spatial distribution of virtual viewpoints generated by each viewpoint selection strategy by collapsing the temporal dimension. With interpolation-based selection between the training viewpoints, the sampled virtual viewpoints cluster in limited regions. In contrast, the coverage-based selection is biased toward extrapolation, rather than filling the gaps between the training viewpoints. By contrast, our proposed method samples virtual viewpoints that effectively cover the gaps between training viewpoints while avoiding excessive extrapolation. In the extrapolation setting, the global strategy samples extrapolative viewpoints beyond the locally reachable region, mainly improv-

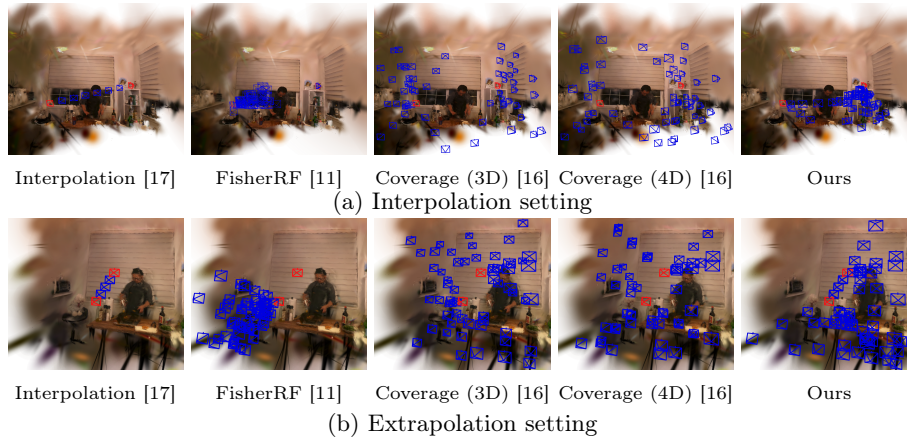


Fig. 6: Visualization of virtual trajectory sampling. Red cameras indicate training cameras, and blue cameras indicate sampled virtual cameras.

ing extrapolation performance, while local candidates stabilize nearby geometry. Further analysis of starting-viewpoint selection using synthetic data, including the temporal axis, is provided in the supplementary materials.

Fine-tuning. To verify the effectiveness of our fine-tuning strategy, we compare it with existing fine-tuning approaches, as shown in Table 4. 3DGS-Enhancer [32] assigns weights based on the distance from the training viewpoints and Gaussian scales. ExploreGS [16] uses LPIPS [66] between the rendered and refined images, as well as G-IoU derived from Gaussian co-visibility. UA-4DGS [15] weights samples by uncertainty computed from the rendering contribution weights on the training images. Our method consistently outperforms existing fine-tuning approaches across all metrics. In addition, while both the consistency mask and the co-visibility mask improve performance, the consistency mask, which suppresses generation errors, has a greater impact. These results suggest that a well-designed weighting scheme that effectively filters out generation errors is key to improving performance.

5 Conclusion

We present a framework for active spatiotemporal virtual-viewpoint selection in 4D Gaussian Splatting. Our framework integrates (i) an active spatiotemporal virtual-viewpoint selection strategy to refine rendered images using rendering sensitivity and motion-aware observation density, and (ii) a fine-tuning method that mitigates generation errors and conflicts with real observations. We evaluate our method on multi-view video datasets using new train/test splits and show that it consistently outperforms existing viewpoint selection and fine-tuning approaches, reducing artifacts and improving rendering quality both quantitatively and qualitatively.

Limitations. The representational capacity of the underlying 4DGS model is limited. Consequently, even with high-quality generated images, 4DGS may struggle to capture complex motion, rendering some regions unrecoverable. Importantly, this limitation is independent of our pipeline for integrating generated images, and may be mitigated by introducing additional constraints.

Moreover, our method is limited by the diffusion model’s generation quality. Sampling viewpoints that the model cannot reliably generate is computationally wasteful. Therefore, the sampling range should be carefully constrained. As diffusion models improve, this limitation may be alleviated.

Acknowledgements

This work was partially supported by JST ACT-X (JPMJAX25C5), JST SPRING (Grant No. JPMJSP2108), and JST ASPIRE (Grant No. JPMJAP2303).

References

1. Azzarelli, A., Anantrasirichai, N., Bull, D.R.: Splatography: Sparse multi-view dynamic gaussian splatting for film-making challenges. In: 3DV (2026)
2. Bae, J., Kim, S., Yun, Y., Lee, H., Bang, G., Uh, Y.: Per-gaussian embedding-based deformation for deformable 3d gaussian splatting. In: ECCV (2024)
3. Bose, S., Dutta, A., Nag, S., Zhang, J., Li, J., Karydis, K., Chowdhury, A.K.R.: Uncertainty-aware diffusion-guided refinement of 3d scenes. In: ICCV (2025)
4. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: CVPR (2023)
5. Duan, Y., Wei, F., Dai, Q., He, Y., Chen, W., Chen, B.: 4d-rotor gaussian splatting: Towards efficient novel-view synthesis for dynamic scenes. In: SIGGRAPH (2024)
6. Edstedt, J., Nordström, D., Zhang, Y., Bökman, G., Astermark, J., Larsson, V., Heyden, A., Kahl, F., Wadenbäck, M., Felsberg, M.: Roma v2: Harder better faster denser feature matching. arXiv preprint arXiv:2511.15706 (2025)
7. Ewen, P., Chen, H., Isaacson, S., Wilson, J., Skinner, K.A., Vasudevan, R.: These magic moments: Differentiable uncertainty quantification of radiance field models. arXiv preprint arXiv:2503.14665 (2025)
8. Guo, F., Hsu, C.C., Ding, S., Zhang, C.: Uncertainty matters in dynamic gaussian splatting for monocular 4d reconstruction. In: ICLR (2026)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
10. Huang, Y.C., Chien, H.J., Lin, C.Y., Chen, Y.H., Liu, Y.L.: Gamo: Geometry-aware multi-view diffusion outpainting for sparse-view 3d reconstruction. arXiv preprint arXiv:2512.25073 (2025)
11. Jiang, W., Lei, B., Daniilidis, K.: Fisherrf: Active view selection and mapping with radiance fields using fisher information. In: ECCV (2024)
12. Jin, L., Chen, X., Rückin, J., Popović, M.: Neu-nbv: Next best view planning using uncertainty estimation in image-based neural rendering. In: IROS (2023)
13. Katsumata, K., Vo, D.M., Nakayama, H.: A compact dynamic 3d gaussian representation for real-time dynamic view synthesis. In: ECCV (2024)

14. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. *ACM TOG* **42**(4) (2023)
15. Kim, M., Lim, J., Han, B.: 4d gaussian splatting in the wild with uncertainty-aware regularization. In: *NeurIPS* (2024)
16. Kim, M., Jeon, S., Cho, I., Yoo, M., Kim, S.J.: Exploregs: Explorable 3d scene reconstruction with virtual camera samplings and diffusion priors. In: *ICCV* (2025)
17. Kong, H., Yang, X., Wang, X.: Generative sparse-view gaussian splatting. In: *CVPR* (2025)
18. Kopanas, G., Drettakis, G.: Improving nerf quality by progressive camera placement for free-viewpoint navigation. In: *VMV* (2023)
19. Kratimenos, A., Lei, J., Daniilidis, K.: Dynmf: Neural motion factorization for real-time dynamic view synthesis with 3d gaussian splatting. In: *ECCV* (2024)
20. Lei, J., Weng, Y., Harley, A., Guibas, L., Daniilidis, K.: Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In: *CVPR* (2025)
21. Li, T., Slavcheva, M., Zollhoefer, M., Green, S., Lassner, C., Kim, C., Schmidt, T., Lovegrove, S., Goesele, M., Newcombe, R., Lv, Z.: Neural 3d video synthesis from multi-view video. In: *CVPR* (2022)
22. Li, Y., Cao, J., Ruan, P., Saxena, D., Zhu, S., Cao, Y.: Geometry-consistent 4d gaussian splatting for sparse-input dynamic view synthesis. *arXiv preprint arXiv:2511.23044* (2025)
23. Li, Y., Kuang, Z., Li, T., Hao, Q., Yan, Z., Zhou, G., Zhang, S.: Activesplat: High-fidelity scene reconstruction through active gaussian splatting. *IEEE RA-L* **10**(8), 8099–8106 (2025)
24. Li, Z., Chen, Z., Li, Z., Xu, Y.: Spacetime gaussian feature splatting for real-time dynamic view synthesis. In: *CVPR* (2024)
25. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: *CVPR* (2021)
26. Li, Z., Wang, Q., Cole, F., Tucker, R., Snavely, N.: Dynibar: Neural dynamic image-based rendering. In: *CVPR* (2023)
27. Liang, Y., Xu, T., Kikuchi, Y.: Himor: Monocular deformable gaussian reconstruction with hierarchical motion representation. In: *CVPR* (2025)
28. Lin, Y., Dai, Z., Zhu, S., Yao, Y.: Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In: *CVPR* (2024)
29. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., Li, X., Sun, X., Ashok, R., Mukherjee, A., Kang, H., Kong, X., Hua, G., Zhang, T., Benes, B., Bera, A.: D13dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In: *CVPR* (2024)
30. Liu, G., Jiang, W., Lei, B., Pandey, V., Daniilidis, K., Motee, N.: Beyond uncertainty: Risk-aware active view acquisition for safe robot navigation and 3d scene understanding with fisherrf. *arXiv preprint arXiv:2403.11396* (2024)
31. Liu, Q., Liu, Y., Wang, J., Lyv, X., Wang, P., Wang, W., Hou, J.: Modgs: Dynamic gaussian splatting from casually-captured monocular videos with depth priors. In: *ICLR* (2025)
32. Liu, X., Zhou, C., Huang, S.: 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. In: *NeurIPS* (2024)
33. Lu, Z., Guo, X., Hui, L., Chen, T., Yang, M., Tang, X., Zhu, F., Dai, Y.: 3d geometry-aware deformable gaussian splatting for dynamic view synthesis. In: *CVPR* (2024)
34. Marza, P., Matignon, L., Simonin, O., Batra, D., Wolf, C., Chaplot, D.S.: Autonerf: Training implicit scene representations with autonomous agents. In: *IROS* (2024)

35. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
36. Nazarczuk, M., Catley-Chandar, S., Tanay, T., Zhang, Z., Slabaugh, G., Pérez-Pellitero, E.: Vidar: Video diffusion-aware 4d reconstruction from monocular inputs. In: NeurIPS (2025)
37. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. TMLR (2024)
38. Pan, X., Lai, Z., Song, S., Huang, G.: Activenerf: Learning where to see with uncertainty estimation. In: ECCV (2022)
39. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: ICCV (2021)
40. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. ACM TOG **40**(6) (2021)
41. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: CVPR (2021)
42. Sabater, N., Boisson, G., Vandame, B., Kerbiriou, P., Babon, F., Hog, M., Gendrot, R., Langlois, T., Bureller, O., Schubert, A., Allié, V.: Dataset and pipeline for multi-view light-field video. In: CVPRW (2017)
43. Safadoust, S., Tosi, F., Güney, F., Poggi, M.: Warprf: Multi-view consistency for training-free uncertainty quantification and applications in radiance fields. In: WACV (2026)
44. Shen, J., Ruiz, A., Agudo, A., Moreno-Noguer, F.: Stochastic neural radiance fields: Quantifying uncertainty in implicit 3d representations. In: 3DV (2021)
45. Shi, C., Yang, C., Hu, X., Chen, M., Pan, W., Yang, Y., Ding, J., Yu, Z., Yu, J.: Sparse4dgs: 4d gaussian splatting for sparse-frame dynamic scene reconstruction. In: AAAI (2026)
46. Stearns, C., Harley, A., Uy, M., Dubost, F., Tombari, F., Wetzstein, G., Guibas, L.: Dynamic gaussian marbles for novel view synthesis of casual monocular videos. In: SIGGRAPH Asia (2024)
47. Sünderhauf, N., Abou-Chakra, J., Miller, D.: Density-aware nerf ensembles: Quantifying predictive uncertainty in neural radiance fields. In: ICRA (2023)
48. Wang, C., Eckart, B., Lucey, S., Gallo, O.: Neural trajectory fields for dynamic novel view synthesis. arXiv preprint arXiv:2105.05994 (2021)
49. Wang, Q., Ye, V., Gao, H., Zeng, W., Austin, J., Li, Z., Kanazawa, A.: Shape of motion: 4d reconstruction from a single video. In: ICCV (2025)
50. Wang, S., Yang, X., Shen, Q., Jiang, Z., Wang, X.: Gflow: Recovering 4d world from monocular video. In: AAAI (2025)
51. Wang, Z., Zhou, H., Blaschko, M.B., Tuytelaars, T., Wu, M.: Diversity-driven view subset selection for indoor novel view synthesis. TMLR (2025)
52. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP **13**(4), 600–612 (2004)
53. Wang, Z., Bhalgat, Y., Li, R., Prisacariu, V.A.: Active view selector: Fast and accurate active view selection with cross reference image quality assessment. arXiv preprint arXiv:2506.19844 (2025)
54. Wei, J., Leutenegger, S., Schaefer, S.: Gsfix3d: Diffusion-guided repair of novel views in gaussian splatting. In: 3DV (2026)

55. Wilson, J., Almeida, M., Mahajan, S., Labrie, M., Ghaffari, M., Ghasemalizadeh, O., Sun, M., Kuo, C.H., Sen, A.: Pop-gs: Next best view in 3d-gaussian splatting with p-optimality. In: CVPR (2025)
56. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: CVPR (2024)
57. Wu, J.Z., Zhang, Y., Turki, H., Ren, X., Gao, J., Shou, M.Z., Fidler, S., Gojcic, Z., Ling, H.: Difx3d+: Improving 3d reconstructions with single-step diffusion models. In: CVPR (2025)
58. Wu, S., Xu, C., Huang, B., Geiger, A., Chen, A.: Genfusion: Closing the loop between reconstruction and generation via videos. In: CVPR (2025)
59. Xian, W., Huang, J.B., Kopf, J., Kim, C.: Space-time neural irradiance fields for free-viewpoint video. In: CVPR (2021)
60. Xiao, W., Cruz, R.S., Ahméd-Aristizabal, D., Salvado, O., Fookes, C., Lebrat, L.: Nerf director: Revisiting view selection in neural volume rendering. In: CVPR (2024)
61. Xiao, Y., Lin, Z., Lu, C., Zhai, D., Jiang, K., Zhao, W., Zhang, W., Jiang, J., Wang, H., Liu, X.: Vdegaussian: Video diffusion enhanced 4d gaussian splatting for dynamic urban scenes modeling. arXiv preprint arXiv:2508.02129 (2025)
62. Yang, Z., Yang, H., Pan, Z., Zhang, L.: Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In: ICLR (2024)
63. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In: CVPR (2024)
64. Yin, X., Zhang, Q., Chang, J., Feng, Y., Fan, Q., Yang, X., Pun, C.M., Zhang, H., Cun, X.: Gsfixer: Improving 3d gaussian splatting with reference-guided video diffusion priors. In: ICML (2026)
65. Zhan, H., Zheng, J., Xu, Y., Reid, I., Rezatofighi, H.: Activermap: Radiance field for active mapping and planning. arXiv preprint arXiv:2211.12656 (2022)
66. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)