Dynamic-Static Decomposition for Novel View Synthesis of Dynamic Scenes with Spiking Neurons

Supplementary Material

8. Appendix

This supplementary material provides:

1) Section 8.1 introduces the details of the Side View Setting as well as the dataset specifications.

2) Section 8.2 provides the details of the static mask M_{fine} introduced in Section 4.1 of the main text, as well as the implementation of the Variant of Spiking Neuron examined in the ablation study in Section 5.3 of the main text.

3) Section 8.3 presents comprehensive quantitative and qualitative comparisons for various datasets and evaluation settings.

4) Section 8.4 presents our discussion of the limitations in Section 6 of the main text.

8.1. Side View Setting & Datasets

In this section, we first introduce the side-view settings, followed by detailed descriptions of the datasets used in our experiments.

8.1.1. Details of the Side View Setting

We adopt a different train-test split for the additional side view setting in the N3DV [22] and MeetRoom [21] datasets, mainly because the standard test views of them only include center viewpoints and do not include side viewpoints. It is worth noting that the VRU [50] dataset already contains side view observations within its test views; therefore, we adhere to its standard train–test split.

Importance of the Side View Setting. Existing evaluations on datasets such as N3DV [22] and MeetRoom [21] are typically conducted under the Center View setting, where the training and testing viewpoints are relatively close. Although convenient, this setting is insufficient for assessing the quality of dynamic-scene reconstruction: when the model overfits to the training views, such overfitting cannot be easily revealed from center-view test images.

In real application scenarios, however, users often observe the scene from viewpoints that differ significantly from the training views. For example, in immersive AR/VR scenarios, users frequently observe the reconstructed dynamic scene from diverse viewpoints, where maintaining consistent motion, dynamic occlusion changes, and spatiotemporal structure is crucial for a realistic experience.

Motivated by these practical requirements, we introduce a Side-View setting, in which the test viewpoints are deliberately placed far from the training viewpoints. This large viewpoint shift makes dynamic-scene overfitting much easier to detect and provides a more reliable evaluation of a model's generalization ability and reconstruction quality.

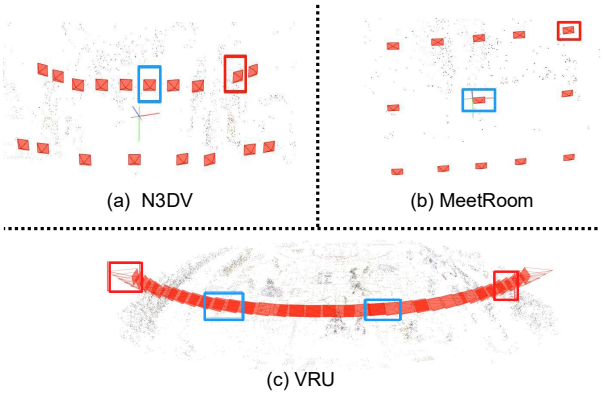

Figure 9. Visualization of center view and side view of **(a)** N3DV Dataset, **(b)** MeetRoom Dataset and **(c)** VRU Dataset. The blue bounding boxes indicate center views, while the red bounding boxes correspond to side views.

Center View vs. Side View. We distinguish center views and side views based on camera location and poses with respect to the scene. Fig. 9 showcases the center views and side views locations of the N3DV [22], MeetRoom [21] and VRU [50] datasets. In the N3DV Dataset and MeetRoom Datasets, we both choose ***Camera 00*** as the center view and ***Camera 02*** as the side view. In the VRU Dataset, we choose ***Camera 10, 20*** as the center views and ***Camera 00, 30*** as the side views.

8.1.2. Details of the Datasets

N3DV Dataset. The N3DV dataset [22] contains dynamic scenes with relatively small motions captured by 18–21 static cameras at 2704×2028 resolution and 30 FPS. Following prior work [17, 22, 55], videos are downsampled by a factor of two, with 1 camera used for testing and the rest for training. **We additionally evaluate a side-view setting using an alternative train–test split**.

MeetRoom Dataset. The MeetRoom dataset [21] consists of dynamic indoor scenes recorded by 13 cameras at 1280×720 resolution and 30 FPS. Following [21, 46], we use 1 camera for testing and 12 cameras for training. **We additionally evaluate a side-view setting using an alternative train–test split**.

VRU dataset. The VRU dataset [50] contains two dynamic basketball court scenes featuring large-scale motion, captured at 1920 × 1080 resolution and 25 FPS with 34 cameras. Following [50], we use 30 cameras for training and 4 for testing. **Since in the VRU [50] dataset the test views are uniformly sampled across all viewpoints, including both side and center views, additional evaluation under the side view setting is unnecessary**.

8.2. Methodological Details

This section presents the details of the static mask M_{fine} and the implementation of the Variant of Spiking Neuron examined in the ablation study.

8.2.1. Details of Static Mask M_{fine} in Sec. 4.1

As mentioned in Sec. 4.1, with the help of the static masks M_{coarse}, $M_{diffuse}$ and M_{temp}, we can generate a fine-grained static mask M_{fine}.

We first combine M_{coarse} and $M_{diffuse}$ to generate a intermediate static mask:

$$M_{inter} = M_{coarse} | M_{diffuse}. \tag{17}$$

However, this intermediate mask might lose some high-frequency details which might observe large residuals and be misclassified into dynamic pixels. So we use a temporal mask M_{temp} [50] to help exclude these static pixels from true dynamic regions, which rely on per-view pixel intensity differences to distinguish dynamic and static pixels:

$$S(x) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(C(x,t) - \frac{1}{T}\sum_{t=1}^{T}C(x,t))^2}, \tag{18}$$

where $C(x,t)$ represents the intensity of pixel x in the t-th frame and pixels with nearly constant constant color intensity are marked as static pixels:

$$M_{temp}(x) = \begin{cases} 1 & S(x) \leq \gamma \\ 0 & S(x) > \gamma \end{cases}. \tag{19}$$

Since a static pixel should observe small residuals across viewpoints and timestamps or obtain nearly constant pixel intensity across timestamps in a single viewpoint, we can generate a fine-grained dynamic mask and reverse it to get a fine-grained static mask M_{fine}, as shown in Fig. 14:

$$M_{fine} = M_{inter} | M_{temp}. \tag{20}$$

Then M_{fine} is further used for 4D Mask Field $\mathcal{F}$ optimization.

8.2.2. Details of the Variant of Spiking Neuron (SN) in Sec. 5.3

Here, we describe the differences between the two Variants used in the ablation study and provide their implementation details.

Variant A. Swift4d [50] utilize continuous dynamic-static attribute $d^c \in \mathbb{R}$ to render a dynamic map by intersected Gaussian set N and then normalize it with *sigmoid* function $\sigma(\cdot)$:

$$\hat{M} = \sigma(\sum_{i \in N} d_i^c \alpha_i T_i), \tag{21}$$

where $T_i = \prod_{j=1}^{i-1}(1-\alpha_i)$ and d_i^c denotes d^c attribute of the i-th Gaussian $\mathcal{G}_i$.

Variant B. Different from Variant A, Variant B (*sigmoid+th*) implementation first normalize d^c and then render the dynamic map, which coule be formulated as belows:

$$\hat{M} = \sum_{i \in N} \sigma(d_i^c) \alpha_i T_i. \tag{22}$$

Variant B is more similar with our optimization process since it does the normalization first.

By minimizing the binary cross-entropy loss $\mathcal{L}$ (e.g., mask loss described in Eq. 16) between the mask prior M and the dynamic map $\hat{M}$, they both optimize the dynamic value d_i^c for each Gaussian point and then use a threshold to generate dynamic-static labels.

8.3. More Experimental Results

This section presents comprehensive quantitative and qualitative results for all datasets under all evaluation settings.

8.3.1. Overall Experimental Results

We report overall results on the standard setting and side view setting of the N3DV, MeetRoom and VRU Datasets, as shown in Tab. 4. Our method outperforms existing methods [17, 46, 49, 50] on novel view synthesis results.

We detail more per-scene quantitative and qualitative results on the standard setting in Tab. 8 and side view setting in Tab. 9, together with Fig. 15, 16 (N3DV), Fig. 12, 13 (MeetRoom) and Fig. 17, 18 (VRU). We provide per-scene decomposition results in Fig. 16 (N3DV), Fig. 13 (MeetRoom) and Fig. 18 (VRU).

We evaluate FPS and training time to validate our effectiveness on improving training and rendering efficiency, as shown in Tab. 5.

8.3.2. Experimental Results under the standard setting

Except for the provided results on the side view setting in Sec. 5.2, we provide detailed novel view synthesis and dynamic-static decomposition results on the standard setting of the N3DV, MeetRoom datasets, as shown in Fig. 10, 11.

Our method present better fine-grained motions, specifically in the boundary regions. In the small-motion N3DV

Table 4. Quantitative comparison with 3DGS-based methods in terms of PSNR and LPIPS on the N3DV, MeetRoom, and VRU datasets. For N3DV and MeetRoom, we report results under both the Standard (Center View) and Side View Settings. For VRU, the Standard Setting already includes both center and side views, so only the standard results are provided. The best and the second best results are denoted by red and blue.

	N3DV						MeetRoom						VRU*	
	Standard (Center View)		Side View		Avg		Standard (Center View)		Side View		Avg		Standard (Center+Side)	
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
4DGS [49]	31.12	0.0583	26.19	0.0753	28.66	0.0668	30.57	0.0382	26.27	0.0638	28.42	0.0510	28.61	0.2320
3DGStream [46]*	31.67	-	25.55	0.0740	28.60	-	31.31	-	25.68	0.0916	28.50	-	27.55	0.2116
Ex4DGS [17]	32.11	0.0479	26.10	0.0667	29.11	0.0573	30.40	0.0460	25.10	0.0764	27.75	0.0612	24.23	0.2838
Swift4D [50]	32.23	0.0434	25.80	0.0619	29.01	0.0520	30.64	0.0421	25.16	0.0676	27.90	0.0549	29.05	0.1777
Ours	32.11	0.0461	26.30	0.0615	29.21	0.0538	31.57	0.0376	26.64	0.0626	29.11	0.0501	29.43	0.1702

* Due to lack of reported LPIPS results in original paper and pretrained model, we do not report the LPIPS metric of [46].

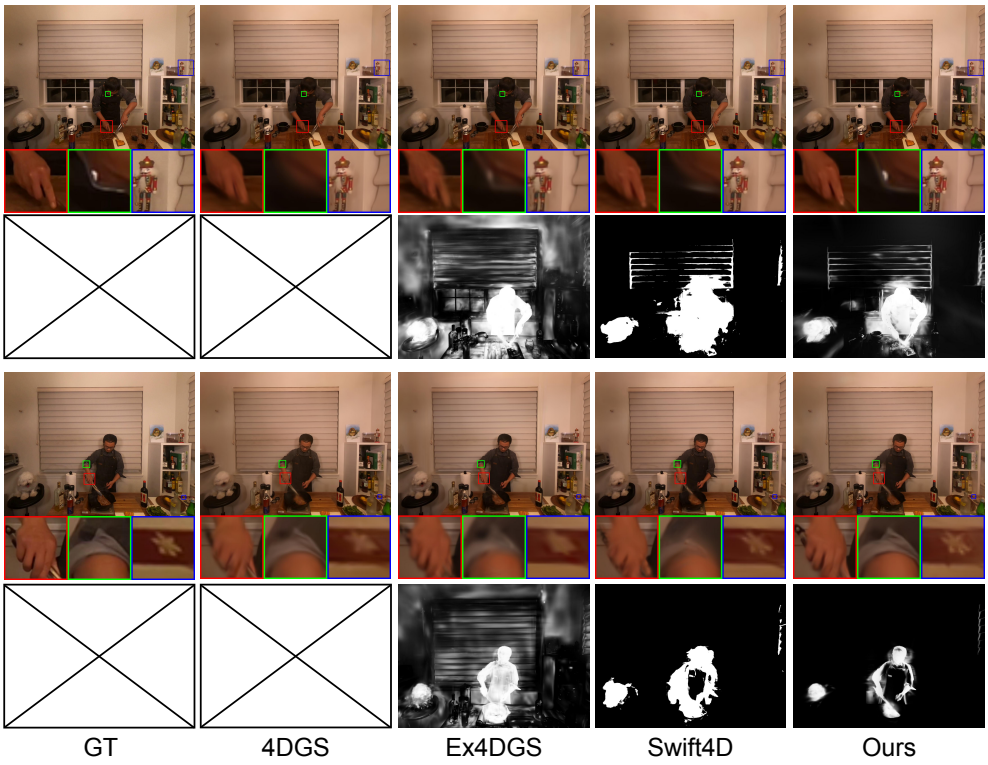

Figure 10. Visualization of detailed novel view synthesis and decomposition results on the *cut beef* and *sear steak* scenes in the N3DV dataset under the standard setting (center view). Please zoom in to observe details.

dataset, our method captures finer details of moving hands and swaying window blinds in the *cut beef* scene, as well as hands and sleeves in the *sear steak* scene (see Fig. 10). In the MeetRoom dataset, our method renders clearer facial features of moving figures. Beyond dynamic region enhancement, our method also retains high-fidelity static regions, which can be observed in the intricate Nutcracker and flower pattern in Fig. 10 as well as the clear edge of the chair in Fig. 11).

Our dynamic-static decomposition significantly reduces redundant dynamic Gaussians in static regions while effectively preserving dynamic Gaussians in dynamic regions, as shown in Figs. 16, 13 and 18. Taking Fig. 16 as an example, compared with [17, 50], our method eliminates most Gaussians on the static walls and only marks those necessary at the current timestamp as dynamic, resulting in a clear boundary for the person.

8.3.3. More Ablation Studies

Ablation on β. We analyze the curves, the gradients, and their impact on the final results under different β values in Fig. 19, 22. Larger β produces sharper and more accurate surrogate gradients but increase the risk of gradient explosion, affecting performance, while smaller β results in in-

Figure 11. Visualization of detailed novel view synthesis and decomposition results on the *discussion* scene in the MeetRoom dataset under the standard setting (center view) . Please zoom in to observe details.

Figure 12. Visualization of novel view synthesis on the MeetRoom dataset under the center view and side view. Please zoom in to observe details.

Table 5. Efficiency comparison of FPS and training time on the N3DV datasets under the side view setting.

	FPS↑	Training Time↓
4DGS [49]	30	50 mins
3DGStream [46]	215	60 mins
Ex4DGS [17]	129	60 mins
Swift4D [50]	125	25 mins
Ours	154	23 mins

accurate approximation. We set $\beta = 2$ in all experiments.

Ablation on box filter size. We compare filter sizes for box filter $\mathcal{B}$ in Fig. 23 and find that a 3×3 filter achieves the best performance. Fig. 21 shows that a larger box filter reduces large pseudo dynamic primitives (discussed in Sec. 8.4) but degrades performance due to excessive smoothing.

Ablation on τ_r and $\tau_{\circledast}$. The ablation studies on τ_r and $\tau_{\circledast}$ are shown in Fig. 20. The best-performing setting is $\tau_r = PERCENTILE(r, 0.7)$, $\tau_{\circledast} = 0.5$, as adopted in the paper.

8.3.4. More Metrics

Comparison on dynamic regions is shown in Tab. 6. Our method achieves superior dynamic part reconstruction across most scenes, validating its effectiveness. The best and the second best results are denoted by red and blue. The dynamic region masks (shown in Fig. 25) are manually annotated using Track Anything [53] since these datasets do not provide such annotations.

The statistics of dynamic Gaussians. Tab. 7 reports the statistics of dynamic Gaussians across three datasets, demonstrating the efficiency of our method. On VRU dataset, our method uses more dynamic Gaussians because many dynamic crowd instances (shown in Fig. 24) are labeled as dynamic, whereas they are misclassified as static by other methods. Full results among all datasets are shown in Tab. 7.

8.4. Discussion of Limitations

Blurry Inputs. The 4D Mask Field $\mathcal{F}$ depends on large photometric residuals of dynamic objects and is insensitive to static regions. When the input frames include severe motion blur which would reduce photometric differences, the 4D Mask Field $\mathcal{F}$ may misclassify blurred dynamic regions as static and thus provide incorrect mask priors and may

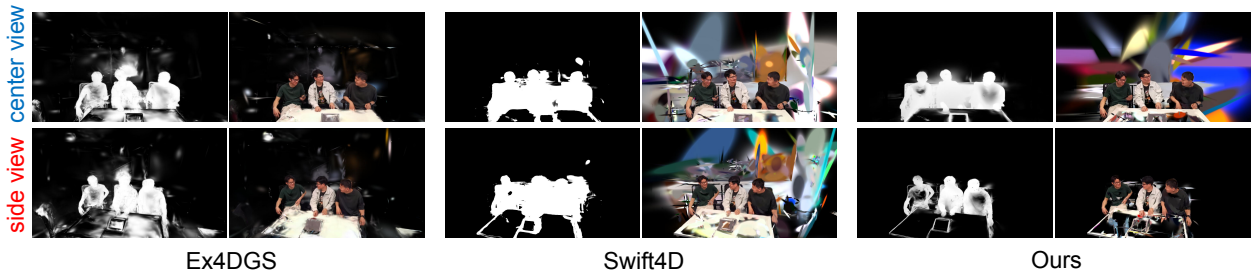

Figure 13. Visualization of dynamic-static decomposition results in the MeetRoom dataset under the center view and side view. Please zoom in to observe details.

Table 6. Quantitative Comparison of Rendering Quality on Masked Dynamic Regions. Our method achieves superior dynamic part reconstruction across most scenes, validating its effectiveness. The best and the second best results are denoted by red and blue.

Method	N3DV				MeetRoom				VRU		Avg	
	Standard (Center View)		Side View		Standard (Center View)		Side View					
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
4DGS [49]	40.35	0.0114	39.09	0.0119	38.36	0.0089	34.68	0.0090	43.57	0.0048	39.21	0.0092
Ex4DGS [17]	41.21	0.0095	39.65	0.0121	38.73	0.0078	34.63	0.0083	38.65	0.0073	38.57	0.0090
Swift4D [50]	41.56	0.0087	39.51	0.0097	37.94	0.0075	34.08	0.0079	44.86	0.0031	39.59	0.0074
Ours	41.61	0.0086	39.82	0.0096	38.82	0.0068	34.78	0.0073	45.93	0.0031	40.19	0.0071

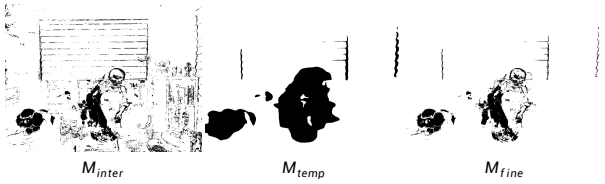

Figure 14. Visualization of M_{inter}, M_{temp} and M_{fine} on the *cook spinach* scene of the N3DV Dataset.

lead to suboptimal visualization on few frames.

Large Pseudo Dynamic Gaussians. Indoor lighting conditions (close light sources in N3DV or specular highlights in VRU) cause color variations across views on static regions, forcing models to fit illumination with large dynamic Gaussians (e.g., the *cut beef* scene in Fig. 16). Occlusion ambiguity also misclassifies some static Gaussians as dynamic. Both factors result in **very few** pseudo dynamic Gaussians with extremely **small rendering weights**, which have negligible impact on rendering quality and only slightly influence dynamic-static separation visualization.

Color discrepancies. Severe cross-view color discrepancies may lead to suboptimal performance of our method. Fig. 26 shows the color distributions for each view in N3DV scenes. Four scenes (Type A) present obvious cross-view color inconsistencies. As our method depends on accurate photometric information, these discrepancies result in slightly degraded performance in these scenes.

Long Sequence Inconsistencies. As shown in Fig. 27, minor cross-chunk inconsistencies may arise in static regions with specular highlights when merging chunk-wise reconstruction results for a long video. Such highlights vary frame-by-frame due to shadows from moving objects and are smoothed as static within each chunk, causing slight inconsistencies during merging.

Table 7. Quantitative Comparison of the GS counts (#G), dynamic GS counts (#Gd) and Dynamic Gaussian Ratios(#Gd/#G) across various datasets. The best and the second best results are denoted by red and blue.

Method	N3DV						MeetRoom						VRU			Avg		
	Standard (Center View)			Side View			Standard (Center View)			Side View								
	#G	#Gd	#Gd/#G	#G	#Gd	#Gd/#G	#G	#Gd	#Gd/#G	#G	#Gd	#Gd/#G	#G	#Gd	#Gd/#G	#G	#Gd	#Gd/#G
ex4dgs	268884	51322	19.09%	278593	72242	25.93%	177464	92665	52.22%	142702	58201	40.78%	368968	93221	25.27%	247322	73530	32.66%
swift4d	307261	38697	12.59%	304756	39567	12.98%	124099	67768	54.61%	127898	68321	53.42%	568285	89208	15.70%	286460	60712	29.86%
ours	298651	27398	9.17%	289340	24382	8.43%	125896	54832	43.55%	128436	54374	42.34%	602481	133455	22.15%	288961	58888	25.13%

Table 8. Quantitative comparison with 3DGS-based methods of PSNR, LPIPS, and FPS on the N3DV, MeetRoom and VRU datasets under the **standard** train-test split, where **Beef**, **Martini**, **Spinach**, **Salmon**, **Steak**, **Sear** respectively stands for *cut roasted beef*, *coffee martini*, *cook spinach*, *flame salmon*, *flame steak*, *sear steak* scenes.

	Metrics	N3DV							MeetRoom	VRU*		
		Beef	**Martini**	**Spinach**	**Salmon**	**Steak**	**Sear**	**Avg**	**Discussion**	**dg**	**gz**	**Avg**
4DGS [49]	PSNR↑	32.66	27.34	32.46	29.00	32.75	32.49	31.12	30.57	27.85	29.36	28.61
	LPIPS↓	0.0530	0.0830	0.0520	0.0810	0.0400	0.0410	0.0583	0.0382	0.2264	0.2376	0.2320
3DGStream [46]*	PSNR↑	32.21	27.75	33.31	28.42	34.30	33.01	31.67	31.31	26.75	28.35	27.55
	LPIPS↓	-	-	-	-	-	-	-	-	0.1977	0.2256	0.2116
Ex4DGS [17]	PSNR↑	33.73	28.79	33.23	29.29	33.91	33.69	32.11	30.40	23.53	24.94	24.23
	LPIPS↓	0.0404	0.0700	0.0420	0.0660	0.0340	0.0350	0.0479	0.0460	0.3040	0.2635	0.2838
Swift4D [50]	PSNR↑	33.72	29.13	33.05	29.75	33.67	33.98	32.23	30.64	27.86	30.25	29.05
	LPIPS↓	0.0395	0.0613	0.0415	0.0551	0.0319	0.0313	0.0434	0.0421	0.1592	0.1958	0.1777
Ours	PSNR↑	33.53	29.50	33.02	29.52	33.58	33.53	32.11	31.57	28.29	30.56	29.43
	LPIPS↓	0.0399	0.0639	0.0426	0.0593	0.0364	0.0345	0.0461	0.0376	0.1529	0.1875	0.1702

* Due to lack of reported LPIPS results in original paper and pretrained model, we do not report the LPIPS metric of [46].

Table 9. Quantitative comparison with 3DGS-based methods of PSNR, LPIPS, and FPS on the N3DV, MeetRoom and VRU datasets under the **side-view** train-test split, where **Beef**, **Martini**, **Spinach**, **Salmon**, **Steak**, **Sear** respectively stands for *cut roasted beef*, *coffee martini*, *cook spinach*, *flame salmon*, *flame steak*, *sear steak* scenes. The best and the second best results are denoted by red and blue.

	Metrics	N3DV							MeetRoom
		Beef	**Martini**	**Spinach**	**Salmon**	**Steak**	**Sear**	**Avg**	**Discussion**
4DGS [49]	PSNR↑	26.31	26.46	26.05	26.73	26.07	25.54	26.19	26.27
	LPIPS↓	0.0783	0.0898	0.0690	0.0790	0.0695	0.0661	0.0753	0.0638
3DGStream [46]	PSNR↑	22.79	25.70	26.87	26.22	26.14	25.58	25.55	25.68
	LPIPS↓	0.0981	0.0821	0.0692	0.0790	0.0580	0.0577	0.0740	0.0916
Ex4DGS [17]	PSNR↑	26.26	26.26	26.54	26.24	26.22	25.10	26.10	25.10
	LPIPS↓	0.0665	0.0798	0.0644	0.0681	0.0576	0.0641	0.0667	0.0764
Swift4D [50]	PSNR↑	25.23	26.22	25.79	26.53	25.48	25.54	25.80	25.16
	LPIPS↓	0.0699	0.0615	0.0660	0.0534	0.0577	0.0632	0.0619	0.0676
Ours	PSNR↑	26.39	26.24	26.09	26.78	26.27	26.05	26.30	26.64
	LPIPS↓	0.0646	0.0663	0.0646	0.0539	0.0586	0.0612	0.0615	0.0626

Figure 15. Visualization of per-scene novel view synthesis in the N3DV dataset under the center view and side view. Please zoom in to observe details.

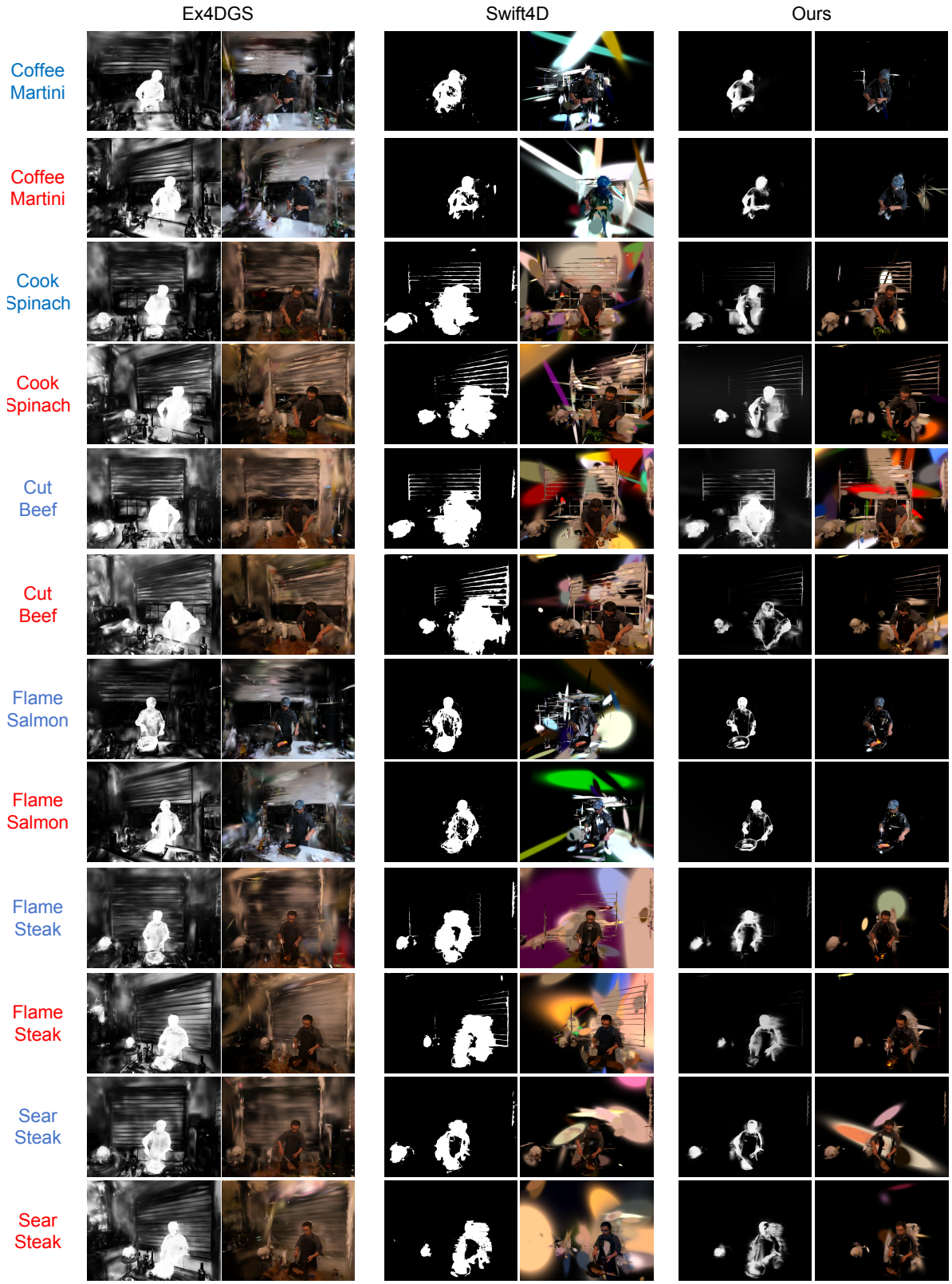

Figure 16. Visualization of per-scene dynamic-static decomposition results in the N3DV dataset under the center view and side view. Please zoom in to observe details.

Figure 17. Visualization of per-scene novel view synthesis in the VRU dataset under the center view and side view. Please zoom in to observe details.

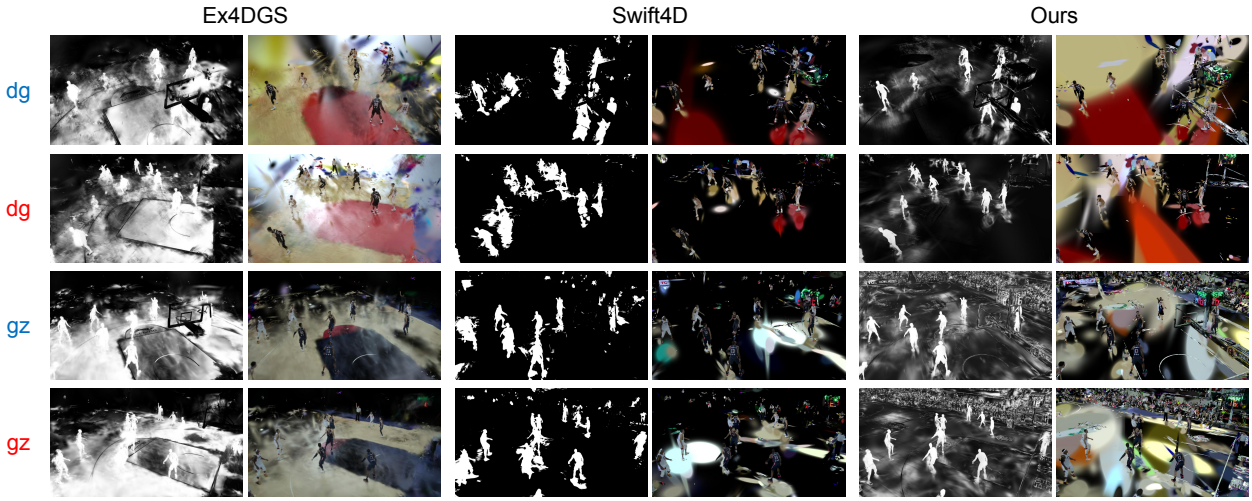

Figure 18. Visualization of per-scene dynamic-static decomposition results in the VRU dataset under the center view and side view. Please zoom in to observe details.

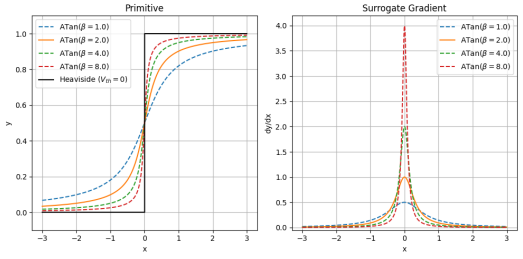

Figure 19. Effect of β on the shape of ATan primitive function and surrogate gradient.

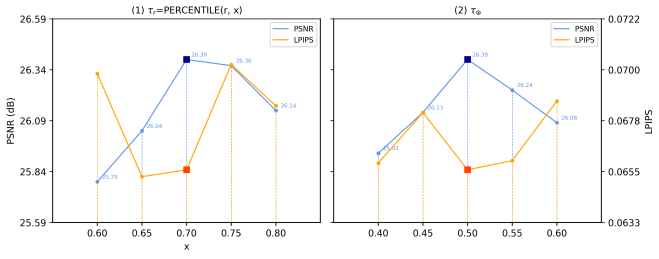

Figure 20. Ablation study on τ_r (left panel) and $\tau_{\circledast}$ (right panel) on the N3DV/cut beef scene under the side view setting.

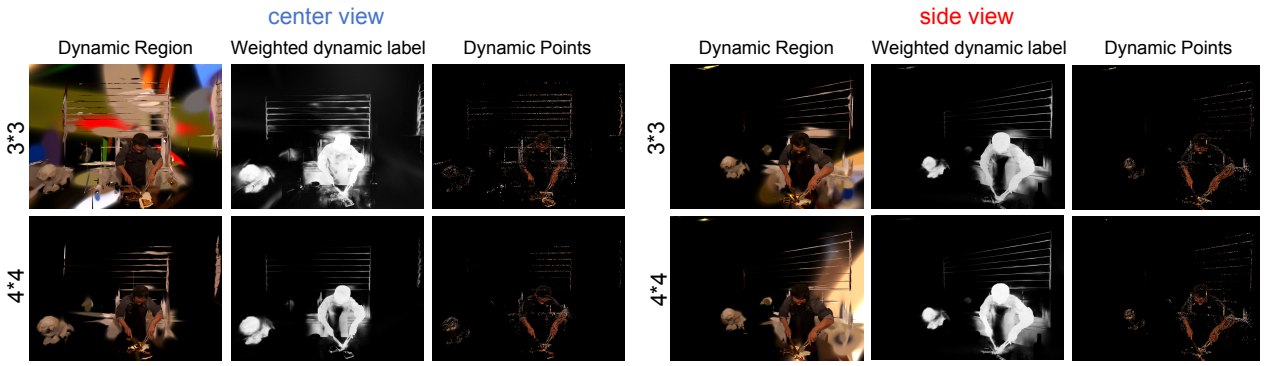

Figure 21. Visualization of ablation study on box filter size on the N3DV/cut beef scene. Larger box filter sizes reduce background pseudo-dynamic GS but over-smooth subtle motions, ultimately degrading the overall rendering quality.

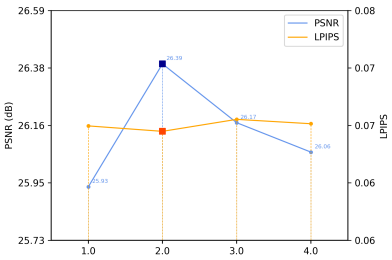

Figure 22. Ablation of β on the N3DV/*cut beef* scene under the side view setting, with optimal performance at β=2.0 (as adopted in this paper).

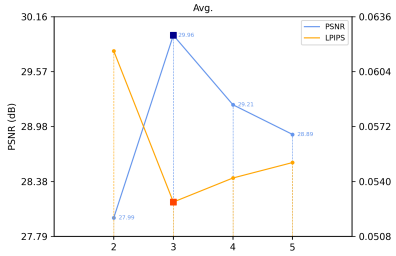

Figure 23. Ablation study on box filter size on the N3DV/cut beef scene (averaged over center and side views), with optimal performance at a box filter size of 3 (as adopted in this work).

Figure 24. VRU dataset contains numerous slightly moving audience, which is challenging to capture.

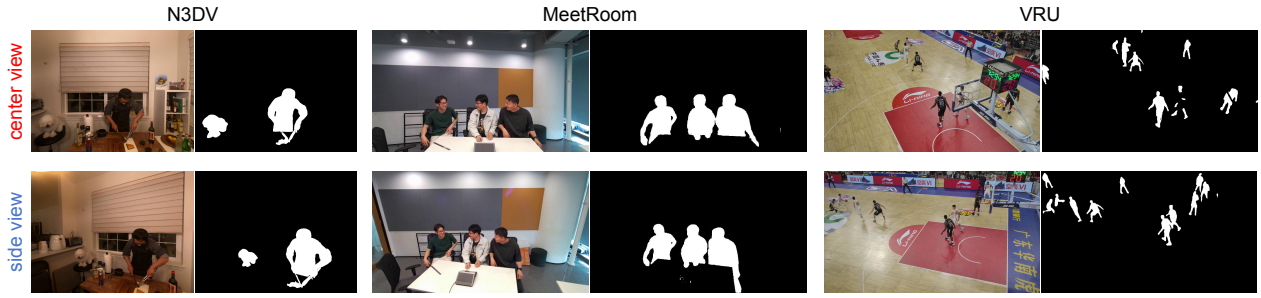

Figure 25. Examples of dynamic (white) masks on the N3DV, MeetRoom and VRU datasets manually annotated by Track-anything [53]. Since the audience are hard to annotate, we mask only explicitly moving athletes for evaluation on the VRU dataset.

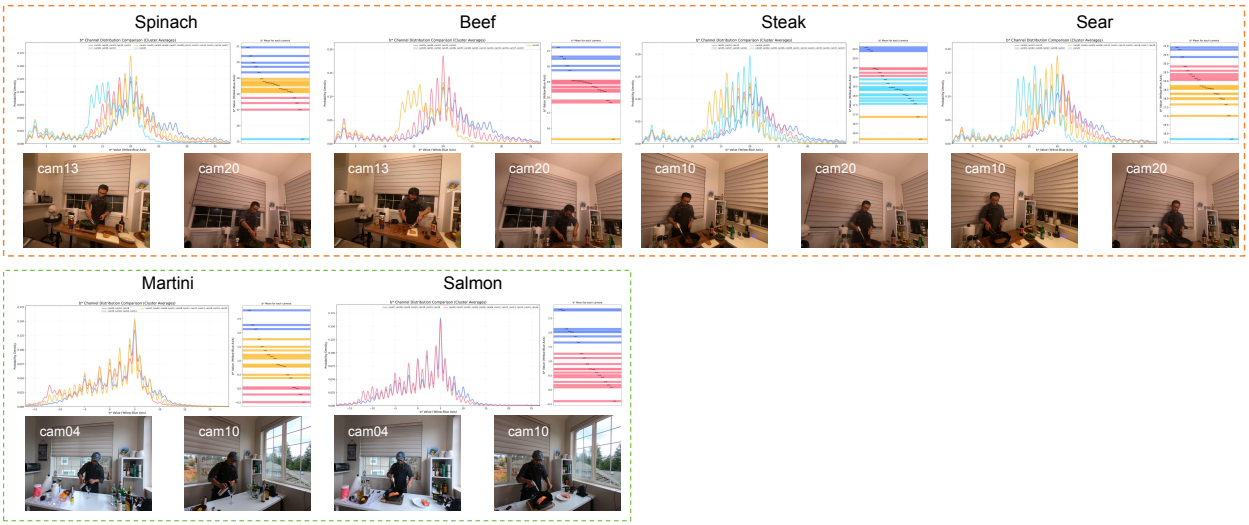

Figure 26. Quantitative visualization of color inconsistency issues in the N3DV datasets. For each scene, we extract the first frame of each camera, calculate the b-channel (yellow-blue axis) distribution in the LAB color space for each image (left panel), and perform clustering on the b-channel mean values of images from each viewpoint (right panel) to roughly evaluate the inter-viewpoint color consistency within the scene. Type A, B denotes two different shooting illumination conditions. We find that the four Type A scenes (*cook spinach*, *cut roasted beef*, *flame steak*, *sear steak*) exhibit strong inter-camera photometric inconsistency, while the two Type B scenes (*coffee martini*, *flame salmon*) show weak such inconsistency.

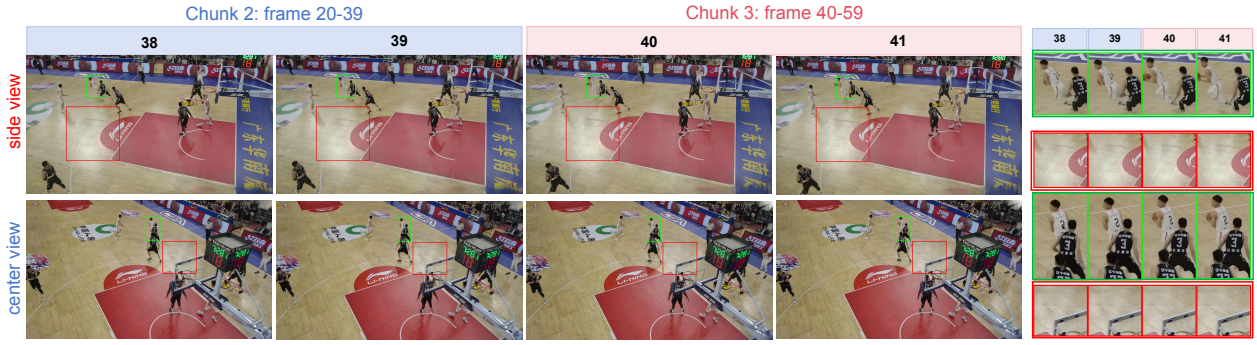

Figure 27. Detailed visualization of Adjacent Frames Across Two Test Views in VRU/dg scene. We mark the moving athletes and the floor specular highlights. Our method exhibits a certain degree of jitter on the side-view floor highlight regions, yet preserves fine consistency for the moving athletes.