MoRe: Motion-aware Feed-forward 4D Reconstruction Transformer

Supplementary Material

1. Training Details

We train our model using the AdamW optimizer [3] to minimize the overall loss function. The training is conducted for 100K iterations with a learning rate scheduler that includes a warm-up phase and a peak learning rate of 1×10^{-6}. At each iteration, we randomly sample 2–24 frames from each sequence with a temporal interval of 1–5. The input images are resized such that the longer side is fixed to 518 pixels, while the shorter side is randomly scaled by a factor of 0.8–1.2 for data augmentation. Training is performed on 64 NVIDIA A800 GPUs for approximately two days. We adopt bfloat16 precision and gradient checkpointing to reduce memory consumption and enable efficient large-scale training.

2. Motion Mask Extraction

2.1. Data Preparation

Most existing datasets lack reliable motion-mask annotations, making it difficult to obtain high-quality supervision for dynamic scene understanding. To address this issue, we propose a robust motion-mask extraction pipeline. Given raw images, we first apply SAM2 [6] to obtain semantic segmentation masks. The ego flow is computed from ground-truth camera poses and intrinsics, while SEA-RAFT [12] predicts the optical flow.

For each semantic region S_k, we compute the average flow discrepancy:

$$d_k = \frac{1}{|S_k|} \sum_{(i,j)\in S_k} \left\| \mathbf{F}^{\text{pred}}(i,j) - \mathbf{F}^{\text{ego}}(i,j) \right\|_2 . \quad (1)$$

A semantic region is considered moving if its discrepancy exceeds a statistical threshold:

$$d_k > \mu_d + 2\sigma_d. \quad (2)$$

Finally, the motion mask $M(u,v)$ is defined as:

$$M(u,v) = \begin{cases} 1, & \text{if } (u,v) \in S_k \text{ and } d_k > \mu_d + 2\sigma_d, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

2.2. Qualitative Results

We evaluate our method on the DAVIS [5] dataset. For visualization, we present both the raw outputs of our model and a refined version obtained by applying the image-level predictor of SAM2 [6]. As shown in Fig. 1, our approach consistently produces accurate motion-object segmentation across diverse scenes and object categories. After simple post-processing, the generated motion masks are sufficiently clean and robust to be directly used in downstream tasks such as dynamic scene reconstruction, moving-object removal, and motion-aware 4D generation.

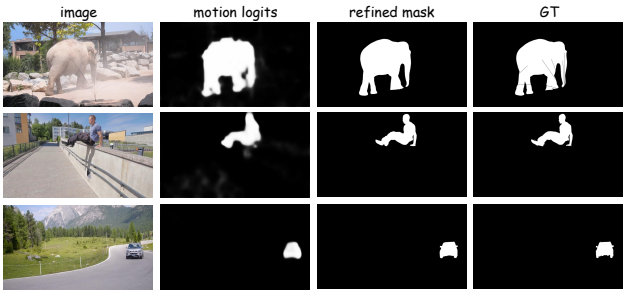

Figure 1. **Qualitative Results of Motion Mask Extraction.** Our method robustly captures moving objects across diverse scenes and objects.

3. Stream Inference

3.1. Implementation Details

Streaming generation has been widely adopted in large language models and related multi-modal systems to reduce latency and computational cost [4, 9, 14]. Inspired by this paradigm, we introduce streaming and causal attention mechanisms into MoRe, enabling real-time, constant-latency generation with image-wise KV caching. This design effectively avoids redundant computation by reusing the stored key–value pairs from previous steps. In addition, we incorporate a window-sliding strategy to prevent unbounded growth of the KV cache.

We employ two streamers: an input streamer for continuous image ingestion and an output streamer for delivering predictions. In the work flow, each new image enters an infinite decoding loop, where its hidden states are concatenated with cached keys and values – optionally applying window sliding – before passing through the stack of N transformer layers. The updated prediction is then immediately emitted through the output streamer, enabling continuous and low-latency streaming outputs.

3.2. Efficiency Test

We evaluate the inference speed of our method on the KITTI dataset at a resolution of 512×144 using an NVIDIA A800 GPU, ensuring consistency across all compared approaches except for Spann3R [**?**], which processes Stream inputs at

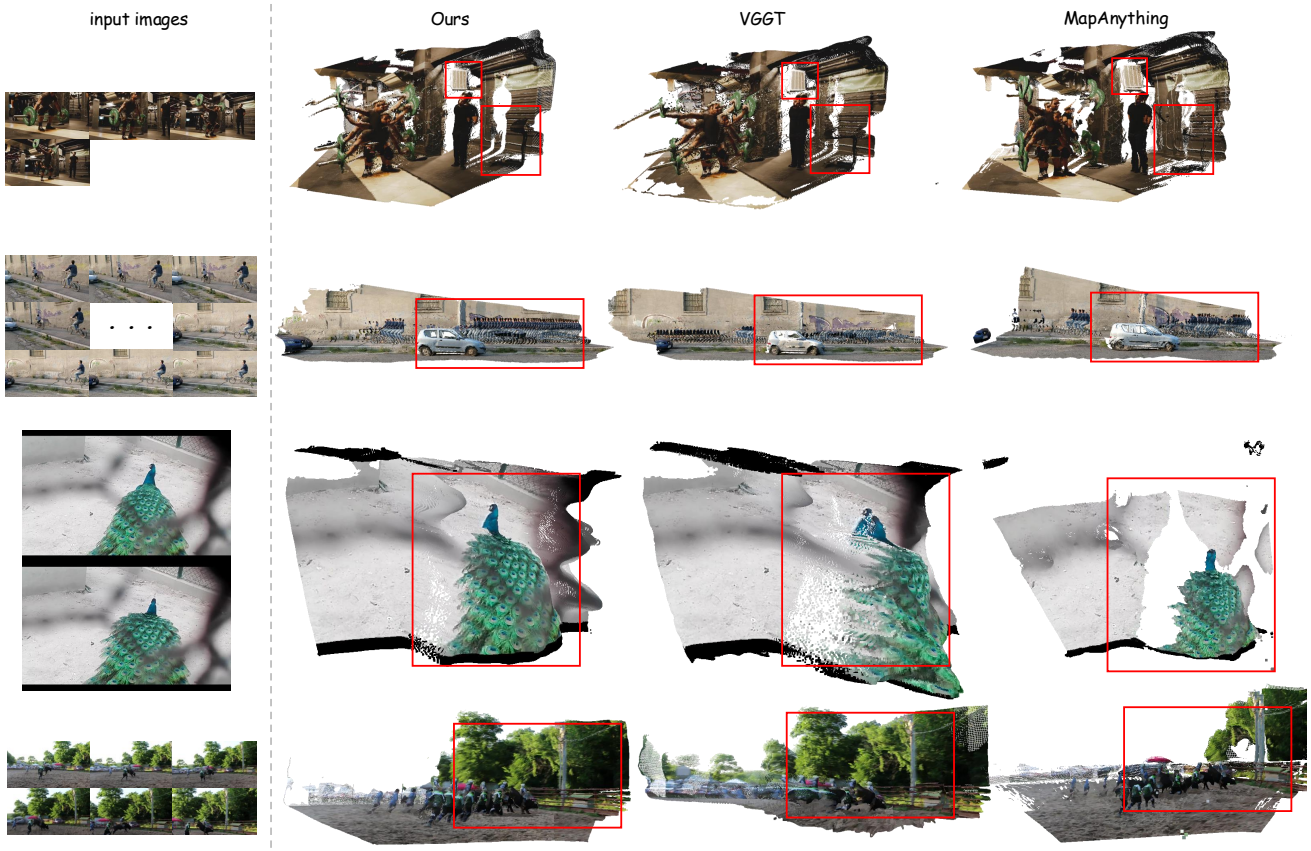

Figure 2. **Qualitative Comparison of Our Model with Other Methods.** We present an extensive set of visual results showcasing reconstruction quality, motion handling, and robustness under challenging dynamic scenarios.

Table 1. Inference speed comparison (FPS), tested on KITTI [1].

Method	**FPS**
VGGT [10]	7.32
Spann3R [**?**] (224×224)	13.55
CUT3R [11]	16.58
Stream3R$^{\alpha}$ [2]	23.48
Stream3R$^{\beta}$ [2]	12.95
Stream3R$^{\beta}$-W[5] [2] (window=5)	32.93
MoRe	30.09

a resolution of 224×224. The performance of other baselines follows the Stream3R[2] evaluation results. In addition, we report the FPS of our model employing a sliding window attention mechanism with a window size of 5. As shown in Tab. 1, our method achieves inference speeds within the fastest tier among all evaluated methods, outperforming most baselines while maintaining competitive reconstruction accuracy. This demonstrates that our approach provides an excellent trade-off between speed and performance, making it highly suitable for real-time 4D reconstruction systems and applications.

3.3. Qualitative Results

To qualitatively evaluate the reconstruction quality of our approach, we further visualize the dynamic 4D scenes reconstructed from monocular video sequences. As shown in Fig. 2, our method effectively captures both static scene geometry and dynamic object motion with high fidelity and temporal coherence. The detailed geometry and consistent motion trajectories demonstrate the robustness of our model in handling complex dynamic environments. These visual results further validate the effectiveness of our approach for practical 4D reconstruction applications. In addition to the stream inference, we also provide more examples of the full attention model. For some methods, slight deviations in the rendered viewpoint occur because their reconstructed point clouds have different scales.

Table 2. Camera Pose Estimation Comparison on Co3Dv2 [7].

Method	Co3Dv2		
	RRA@30↑	RTA@30↑	AUC@30↑
Fast3R [15]	97.49	91.11	73.43
CUT3R [11]	96.19	92.69	75.82
FLARE [16]	96.38	93.76	73.99
VGGT [10]	98.96	97.13	88.59
π^3 [13]	99.05	97.33	88.41
MoRe	**99.49**	**98.11**	**91.42**

4. Motion Aligned Attention

4.1. Quantitative Results

We further evaluate our model on the Co3Dv2 [7] dataset to verify its capability in static scene reconstruction, and we additionally compare against a broader range of baselines. The results are summarized in Tab. 2. As discussed in the main text, our full-attention variant achieves the best overall performance and surpasses all baselines, including the state-of-the-art π^3 method. These results demonstrate that, although our architecture and training strategy are primarily designed for modeling dynamic scenes and suppressing motion-induced ambiguities, the model retains excellent reconstruction accuracy on purely static scenarios. This highlights the strong robustness and generalization ability of our approach: the motion-aware design does not compromise performance when no motion is present, and instead enables the model to effectively capture both dynamic and static structural cues in a unified framework.

4.2. Visualization

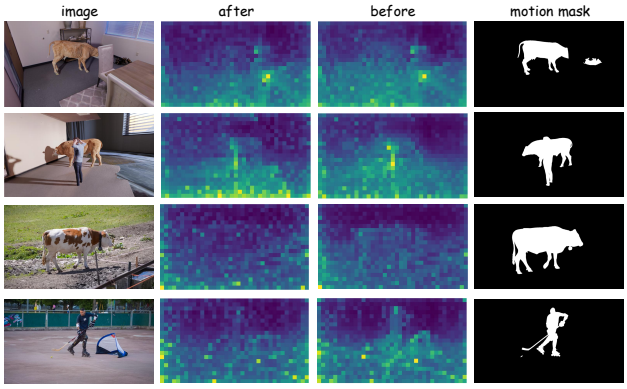

Figure 3. **Attention Map Comparison.** We visualize the attention map on Dynamic Replica [8] and DAVIS [5] dataset. Our motion-aligned training suppresses undesired attention from camera tokens to dynamic objects, yielding cleaner and more structured attention patterns.

To better illustrate the effectiveness of our motion-aligned training strategy, we visualize and compare the attention weight maps of camera token before and after training. Specifically, we select the last attention layer of our model and compute the average attention weight across all heads to obtain a stable and interpretable heatmap representation. As shown in Fig. 3, the attention distribution becomes significantly more structured. The attention weight from camera tokens to dynamic objects is notably suppressed, while attention toward static regions becomes more concentrated and semantically coherent. This indicates that the model has learned to reserve camera tokens for representing stable scene information, preventing dynamic content from leaking into the global representation. The resulting separation leads to cleaner latent features, more stable motion reasoning, and ultimately more accurate 4D reconstruction. These observations validate that our training strategy effectively regularizes the attention behavior and enforces the intended representational roles of different token types.

4.3. Loss Design

We initially explored divergence-based formulations, such as applying a KL divergence to align the attention distribution to the motion-score distribution. While this approach appears principled, it implicitly normalizes attention into a probability distribution, which tends to introduce an undesirable inductive bias in static scenes. In static regions, the correct behavior should allow attention weights to remain largely unconstrained, whereas KL-based losses force all tokens to contribute to a normalized distribution even when no motion exists, leading to degraded performance. The constant C serves as a neutral baseline representing the default attention level. During training, tokens with high motion scores ($\hat{a}_i$ large) are encouraged to deviate from this baseline, while tokens associated with static content ($\hat{a}_i \approx 0$) receive minimal gradient updates. This yields a motion-adaptive behavior that avoids imposing constraints where no motion is present. The multiplicative term $(\alpha_i - C)\,\hat{a}_i$ acts as a gating mechanism, in which motion-relevant tokens receive stronger supervision and motion-irrelevant tokens are softly ignored. This formulation provides flexibility and avoids the normalization issues inherent to divergence losses. We conducted ablations comparing the proposed loss with a KL-based alternative. As shown in Tab. 3, the KL formulation performs worse in both static and low-motion scenarios, confirming that the proposed motion-gated design better matches the nature of the task and provides more stable training behavior.

Table 3. Ablation on Loss Function for Motion Alignment.

	Sintel			TUM-dynamics		
Method	ATE↓	RPE_{trans}↓	RPE_{rot}↓	ATE↓	RPE_{trans}↓	RPE_{rot}↓
w/ KL loss	0.185	0.084	0.707	0.029	0.015	0.350
Ours	**0.147**	**0.082**	**0.616**	**0.026**	**0.013**	**0.320**

5. Limitations

Despite demonstrating strong performance in dynamic scenes, our method has several limitations. First, although our method achieves strong results in dynamic scene modeling, it heavily depends on the accuracy and quality of motion mask annotations. Since the motion masks provide critical supervision to distinguish moving regions from static background, any errors, noise, or inconsistencies in these masks can propagate through the training process, leading to degraded reconstruction quality and less reliable motion reasoning. This reliance poses a limitation, especially when high-quality motion mask labels are unavailable or difficult to obtain in real-world scenarios. Future work could explore more robust or self-supervised techniques to mitigate the impact of imperfect motion supervision and reduce dependency on manual or heuristic mask extraction. Second, while the feed-forward architecture enables efficient and real-time inference, it may struggle to capture very long-term temporal dependencies and complex dynamic interactions that extend beyond the modeled temporal window. Third, the model may exhibit reduced robustness in scenes with extremely fast or non-rigid motions, where motion patterns are highly irregular and difficult to disentangle. In addition, our model can fail in heavily motion-blurred scenarios, where rapid camera movement or fast object motion leads to severely degraded visual cues. In such cases, attention alignment becomes unreliable, causing inaccurate depth, unstable poses, or distorted geometry. Lastly, our current approach does not explicitly handle occlusions or severe appearance changes over time, which can lead to artifacts or inconsistencies in the reconstructed 4D scenes. Addressing these challenges is an important direction for future research.

References

[1] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013. 2

[2] Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. STream3R: Scalable sequential 3D reconstruction with causal transformer. In *ICLR*, 2026. 2

[3] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[4] Zhenyu Ning, Jieru Zhao, Qihao Jin, Wenchao Ding, and Minyi Guo. Inf-mllm: Efficient streaming inference of multimodal large language models on a single gpu. *arXiv preprint arXiv:2409.09086*, 2024. 1

[5] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 1, 3

[6] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1

[7] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 3

[8] Edgar Sucar, Zihang Lai, Eldar Insafutdinov, and Andrea Vedaldi. Dynamic point maps: A versatile representation for dynamic 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7295–7305, 2025. 3

[9] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[10] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2, 3

[11] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025. 2, 3

[12] Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *European Conference on Computer Vision*, pages 36–54. Springer, 2024. 1

[13] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025. 3

[14] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023. 1

[15] Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3R: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Conference*, pages 21924–21935, 2025. 3

[16] Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gor-

don Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Conference*, pages 21936–21947, 2025. 3