LIM: Large Interpolator Model for Dynamic Reconstruction

Supplementary Material

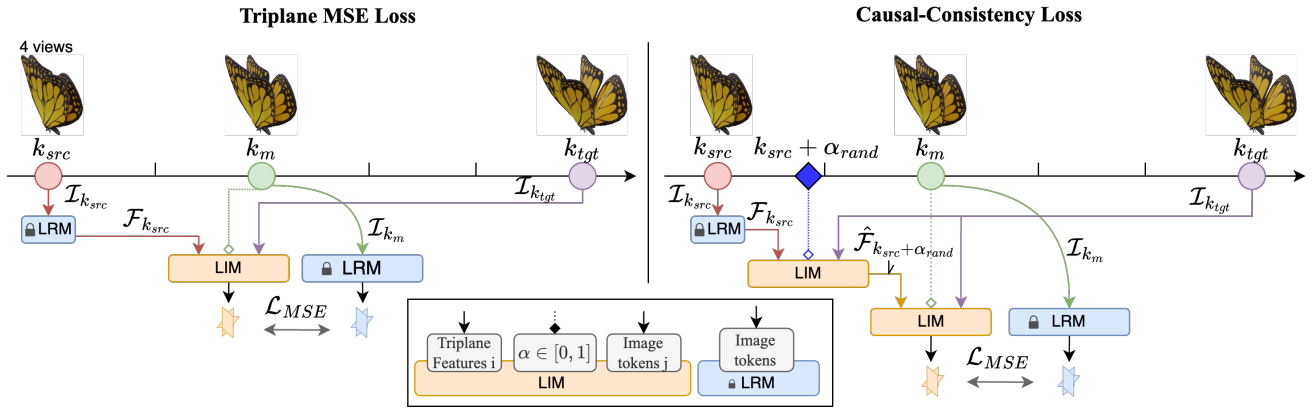

Figure 7. **LIM training losses**. (Left) The triplane MSE loss $\mathcal{L}_{\mathcal{T}}$ only supervises LIM on keyframes k_m. (Right) The causal consistency loss $\mathcal{L}_{\text{causal}}$ samples in-between keyframes with an additional forward-pass to LIM. Note that the second pass of LIM takes as input the intermediate features from LIM instead of the intermediate features from LRM.

A. Additional Evaluations

We recommend looking at the project page [https://remysabathier.github.io/lim.github.io], to see the video results. In particular, the webpage contains video result of RGB interpolation, XYZ canonical tracking, monocular reconstruction and mesh reconstruction.

Evaluation on OOD data We provide qualitative results on the Consistent4D eval set, which includes *real-world scenes*, in Table 5.

B. Additional Method Insights

Weight Initialization. The composition of blocks in LIM and LRM is presented in Fig. 2. We initialize LIM with LRM to take advantage of the learned 3D intermediate representation. More specifically, the intermediate-features cross-attention layers are derived from the self-attention layers from LRM. Furthermore, the image cross-attention layers are initialized using the image cross-attention layers from LRM, and the self-attention layers are initialized from the self-attention layer of LRM. Initialization is similar for $\overline{\text{LRM}}$ and $\overline{\text{LIM}}$ (presented in Fig. 9).

Model size. We ablate the choice of the number of layers in Tab. 6. We observe that LIM accuracy is proportional to the number of blocks in the architecture. However, adding more blocks in LIM slows down the interpolation. We set $N_{layer} = 6$ as a good trade-off between speed and accuracy.

Dataset details Our *3D* dataset includes 142,123 assets, while the *4D* dataset comprises 6,052 rigged models, each with 16 to 128 keyframes. We render the keyframes using Blender and the Cycles engine.

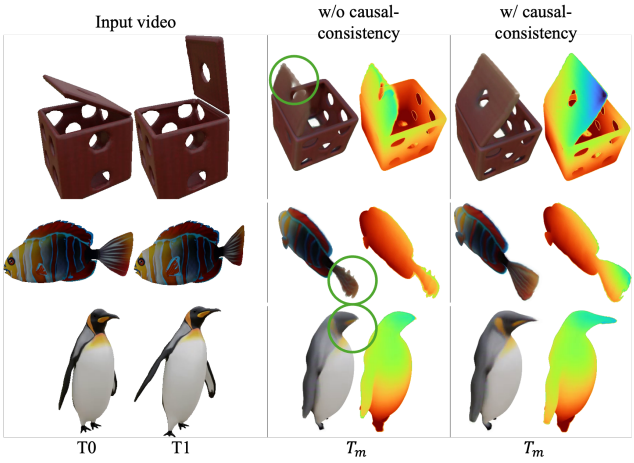

Figure 8. **Causal-loss ablation.** We show triplane interpolation result from LIM models trained either with the triplane MSE loss $\mathcal{L}_{\mathcal{T}}$ only, or with both $\mathcal{L}_{\mathcal{T}}$ and the causal-consistency loss $\mathcal{L}_{\text{causal}}$.

Table 5. **Monocular reconstruction (out of distribution OOD).**

	Inf. Time	Consistent4D set	
		LPIPS	FVD
Consistent4D	~90 min	0.428	1134.7
TripoSR	~0.5 min	0.497	1428.2
LIM (Our)	~3 min	**0.114**	**781.9**

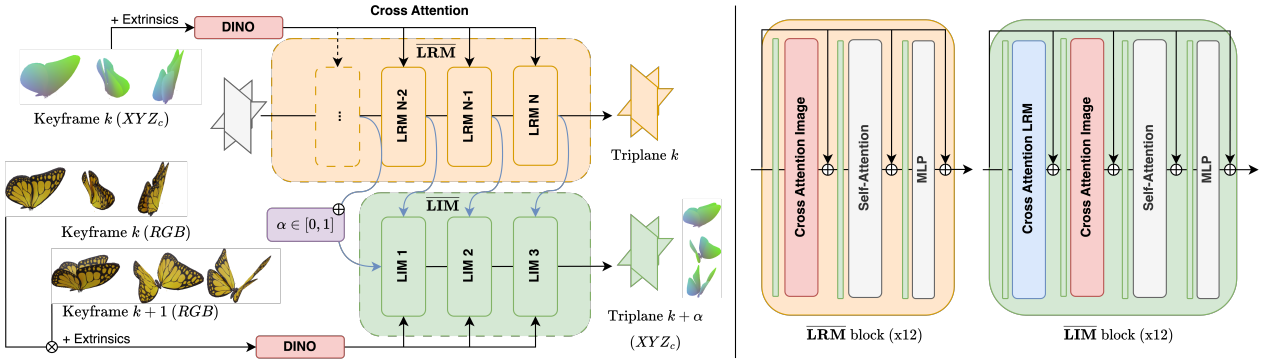

Figure 9. $\overline{\text{LIM}}$ **framework**. (Left) Given multi-view RGB images on 2 timesteps k and $k+1$ and XYZ canonical renders on timestep k, $\overline{\text{LIM}}$ interpolates any intermediate 3D representation of the XYZ canonical coordinate at $k+\alpha, \alpha \in [0,1]$. This gives direct correspondences in 3D space between the source shape at k and the interpolated shape at $k+\alpha$. In practice, our $\overline{\text{LIM}}$ architecture has 6 blocks and $\overline{\text{LRM}}$ 12 blocks. (Right) Block structure of $\overline{\text{LRM}}$ and $\overline{\text{LIM}}$. We include layer normalization before each module in blocks.

	PSNR ↑	PSNR_{FG} ↑	LPIPS ↓
LIM- 3 layers	22.35	14.56	0.079
LIM- 8 layers	23.19	16.2	0.075
LIM	23.11	16.12	0.075

Table 6. **Performance as a function of # layers** reporting interpolation accuracy of LIM while varying the number of transformer blocks in the architecture.

Causal consistency loss. We illustrate in Fig. 7 the behavior of the triplane MSE loss $\mathcal{L}_{\mathcal{T}}$ and the causal-consistency loss $\mathcal{L}_{\text{causal}}$ (see Sec. 3). $\mathcal{L}_{\mathcal{T}}$ involves a single pass of LIM and two passes of LRM, while $\mathcal{L}_{\text{causal}}$ involves 2 passes of LRM and 2 passes of LIM. Note that during LIM training, the weights of LRM are frozen. In practice, we discovered that the causal consistency loss was essential to achieve precise and accurate interpolation over a range of shapes and motions. We show interpolation results (in the same setting as Sec. 4.1) in Fig. 8, with a LIM model trained either with $\mathcal{L}_{\text{causal}}$ activated or deactivated.

Positional Encoding We apply positional encoding to the interpolation time $\alpha \in [0,1]$ with $\phi : \mathbb{R} \to \mathbb{R}^{2D}$, such that $\forall i \in [1, D], \phi(\alpha)[2i] = \cos(\alpha f_{2i}); \phi(\alpha)[2i+1] = \sin(\alpha f_{2i+1})$, and $f_i = \exp[-\frac{\log 10.000}{D}.i]$; we set $D = 512$ so that $2D$ matches the LRM embedding dimension.

4D reconstruction with ARAP regularization . We observe that our mesh-tracking framework can incorporate ARAP regularization to mitigate issues like triangle inversion or self-intersection. Instead of relying solely on direct matching through nearest neighbor search in the space of canonical coordinates (refer to Section Sec. 3.4), we implement a concise optimization loop. This loop incorporates both canonical-coordinate matching and ARAP energy as objectives to minimize.