Inferring Compositional 4D Scenes without Ever Seeing One

Supplementary Material

A. Training Details

We provide details of the full training losses described in Sec. 3.2. Specifically, eq. (3) provides the equation for $\mathcal{L}_S$ and thereafter describes $\mathcal{L}_T$ in text. Here, we write $\mathcal{L}_T$ formally:

$$\mathcal{L}_T = \mathbb{E}\left[\sum_{f=1}^{F} \left\|({}^f\boldsymbol{\epsilon} - {}^f\mathbf{z}_0) - \mathbf{v}_\theta({}^f\mathbf{z}_{t_f}, t_f, {}^f\mathbf{y})\right\|^2\right]. \quad (4)$$

The above expression in eq. (4) for the temporal loss is almost identical to the spatial expression in eq. (3). The obvious change is the the frame index f replacing the object index i, as each data: $({}^f\mathbf{z}_0, {}^f\mathbf{y})$ is sampled from DeformingThings [42]. Additionally, the conditional image embeddings ${}^f\mathbf{y}$ are separate for each frame among F, unlike in the spatial loss expression, where all N objects use the same image embedding $\mathbf{y}$.

Finally for completeness, we formally write the regularization loss, *i.e.*, the TripoSG [43] loss[1] as follows:

$$\mathcal{L}_R = \mathbb{E}\left[\left\|(\boldsymbol{\epsilon} - \mathbf{z}_0) - \mathbf{v}_\theta(\mathbf{z}_t, t, \mathbf{y})\right\|^2\right]. \quad (5)$$

Note that, each sample in eq. (5) only consists of one object with no temporal evolution. Samples $(\mathbf{z}_0, \mathbf{y})$ are obtained from the Objaverse training set [15]. Finally, the overall loss is $\mathcal{L}_{S/T/R}$: $\mathcal{L}_S, \mathcal{L}_T, \mathcal{L}_R$ are sampled with a ratio of $0.35 : 0.35 : 0.3$ respectively. The regularization and its sampling ratio of 0.3 is also used by PartCrafter [47] and MIDI [29].

B. User Study Details

To quantitatively evaluate our method's perceptual quality, we conducted a user preference study. The study was administered via Google Forms and compared our full model (with Attention Mixing) against an ablation baseline (without Attention Mixing).

Procedure. As shown in Fig. 8, participants were presented with a 2D input sequence indicating intended object motion and two corresponding 3D animated samples (labeled (1) and (2)). For each comparison, they answered the question: *"Which sample better matches the input in terms of object placement, motion, and scene structure?"* by rating their preference on a 5-point Likert scale (1: Sample 1 is better, 3: Both are about the same, 5: Sample 2 is better). To prevent bias, the assignment of our method to Sample (1) or (2) and the order of the scenes were randomized for each participant.

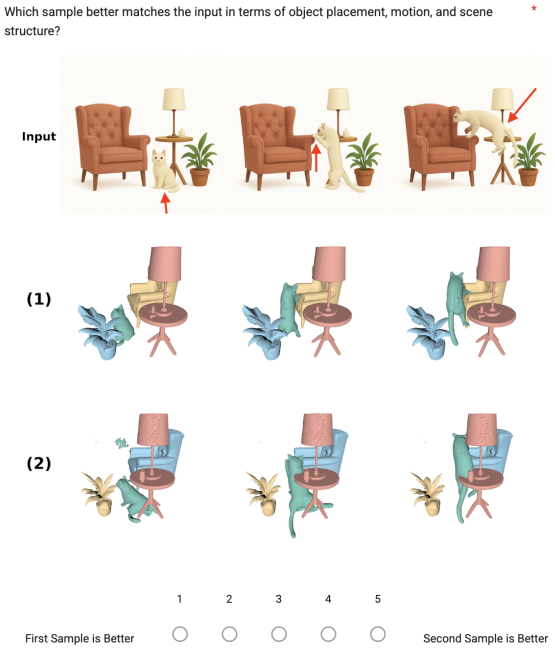

Figure 8. **Our user study interface, run on Google Forms.** Participants viewed a 2D input sequence (top) and two 3D results (middle, bottom), then rated their preference on a 5-point scale.

Data Analysis. All ratings were included in our final analysis. The preference scores reported in the main paper show the complete distribution of judgments across the 5-point Likert scale.

C. Evaluation on CMU Panoptic Dataset

To quantitatively assess the temporal consistency and motion accuracy of our model, we performed an evaluation on the CMU Panoptic dataset [32]. This section details our protocol for preparing the data and computing the metrics.

Ground Truth Point Cloud Generation. We first generated ground truth (GT) point clouds from the raw RGB-D Kinect data provided in the dataset. To ensure a clean and fair comparison, we pre-processed these GT clouds in two steps:

1. **Ground Removal:** The ground plane was removed using a simple height threshold.

[1] The training loss for DiT is not mentioned explicitly in the reference. Please refer to Algorithm 1 in [50] for the rectified flow loss.

2. **Denoising:** We applied a statistical outlier removal filter to eliminate stray, floating points in the cloud.

Alignment and Metric Computation. A key challenge in evaluating generative models is that their outputs are not inherently aligned with the GT coordinate system; they may have an arbitrary scale, rotation, and translation. To address this, we adopted a first-frame alignment protocol.

For each sequence, we independently registered the initial generated mesh (frame 1) from both our full model and the baseline (without Attention Mixing) to the corresponding ground truth point cloud. This one-time alignment transformation (capturing scale, rotation, and translation) was then applied uniformly to all subsequent frames generated by that method for the entire sequence.

Finally, we computed the Chamfer Distance (CD) between our transformed per-frame reconstructions and the GT point clouds. This metric effectively measures how much the predicted motion deviates from the ground truth over time, given an initial registration. A lower accumulated CD indicates a more accurate and temporally consistent motion prediction. Visual examples of these per-frame alignments are shown in Fig. 9, Fig. 10 and Fig. 11. We can observe that not only registered 1^{st} frame, but also the remaining frames align well with the GT pointcloud despite the motion, with surprisingly small drift.

Figure 9. Ablation study on mixing components across five time steps in the CMU Panoptic [32] sample. The top row shows the ground truth frames, followed by two views with our mixing strategy and two views without. Gray points denote the ground truth point cloud.

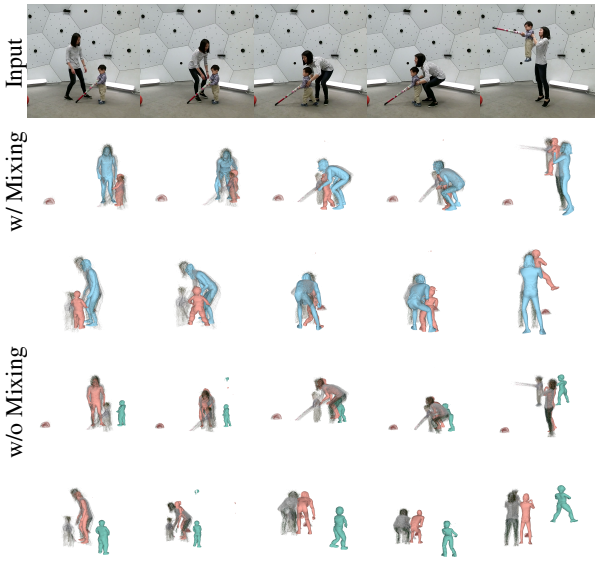

Figure 10. Ablation study on mixing components across five time steps. The top row shows the ground truth frames, followed by two views with our mixing strategy and two views without. Gray points denote the ground truth point cloud.

Figure 11. Ablation study on mixing components across five time steps. The top row shows the ground truth frames, followed by two views with our mixing strategy and two views without. Gray points denote the ground truth point cloud.

D. Scalability with Respect to Object Count

COM4D supports denoising with up to 8 parts, and up to 16 parts with additional finetuning. In the static dataset, scenes contain on average 6.19 objects, with the most common configuration being 5 objects (16.67%).

We observe that performance is moderately sensitive to

both the number of dynamic objects and the complexity of the static scene. As the number of interacting objects increases, the compositional reasoning task becomes more challenging, leading to gradual degradation in reconstruction quality.

E. Additional Qualitative Results on Compositional 4D

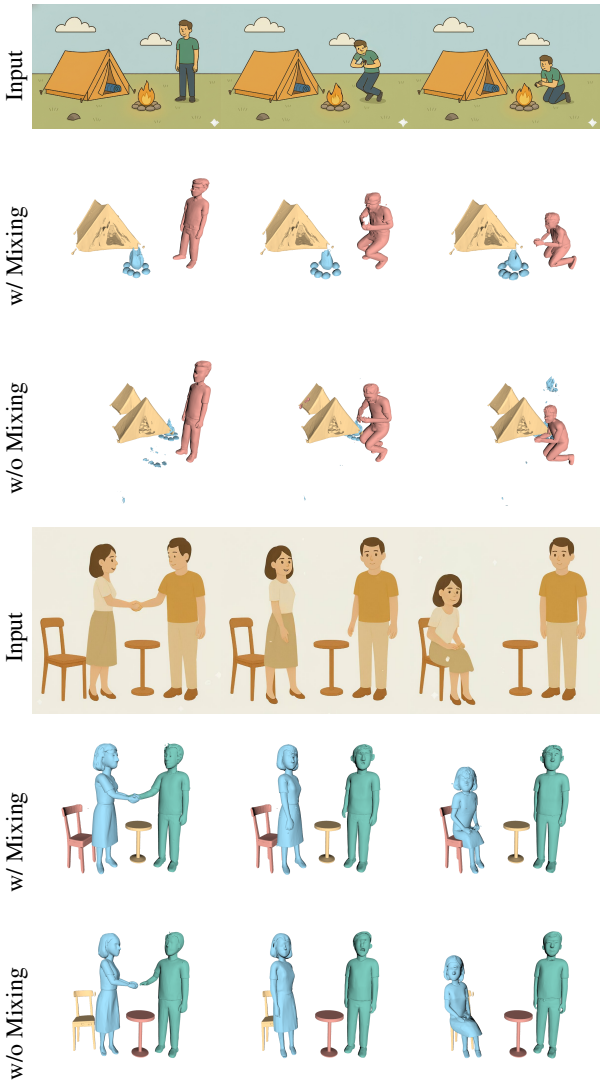

Figure 12. Ablation study on mixing components for various sequences. For each sample, we show input frames, results with our mixing strategy, and results without. In particular, the chair pose and the interaction between the dynamics (the lady shaking hands with the man) and the dynamic and static (lady and the chair) are captured incorrectly without mixing.

F. Additional Qualitative Results with Moving Camera

COM4D assumes a fixed camera when reconstructing compositional scenes containing both static and dynamic objects, as camera motion is inherently entangled with static scene geometry. In such settings, disentangling camera motion from scene structure is fundamentally ambiguous without additional constraints or supervision.

That said, COM4D is not inherently restricted to static-camera scenarios. In cases where the scene consists only of dynamic objects, the model can still recover consistent relative 4D structure and motion over time. This is because COM4D relies on learning relative spatial relationships and temporal coherence, rather than absolute camera-referenced geometry.

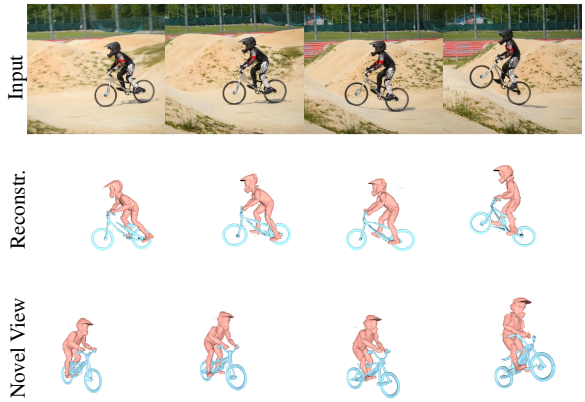

Figure 13. Qualitative results for a sample with dynamic camera.

G. Physical Plausibility and Object Interactions

COM4D is a fully data-driven framework and does not explicitly enforce physical constraints such as contact or collision avoidance. Instead, interactions between multiple objects are modeled implicitly through attention mechanisms in the latent space.

To assess whether this implicit modeling yields physically plausible reconstructions, we evaluate mesh interpenetration across 8 video sequences by measuring the intersection-over-union (IoU) between reconstructed object meshes. We observe a low average overlap of IoU $= 0.0096$, indicating minimal interpenetration between interacting objects.

For reference, we report an IoU of 0.0018 on 3D-FRONT, where ground-truth scenes exhibit little to no physical contact between objects (see Tab. 2). The small gap between these values suggests that COM4D maintains a comparable level of physical plausibility despite not explicitly modeling physical constraints.

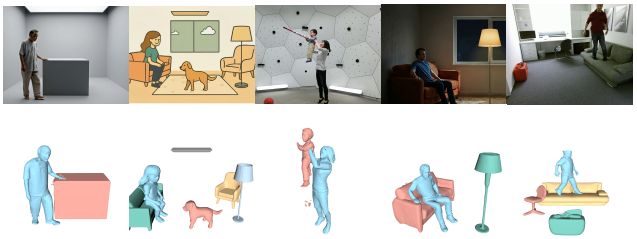

Figure 14. Qualitative results on interaction scenarios. Top row shows input frames, and bottom shows 4D reconstructions.

H. Comparison with Optimization-Based Baselines

We compare COM4D against the recent optimization-based method *Shape of Motion (SOM)* [80]. SOM struggles to recover complete geometry under occlusions, often producing incomplete or noisy reconstructions (see Fig. 15).

Quantitatively, SOM achieves a Chamfer Distance (CD) of **11.30 cm**, whereas COM4D achieves **7.42 cm** on [32, 33]. This gap highlights the advantage of our generative formulation, which leverages learned spatial and temporal priors to produce more complete and coherent reconstructions in challenging compositional settings.

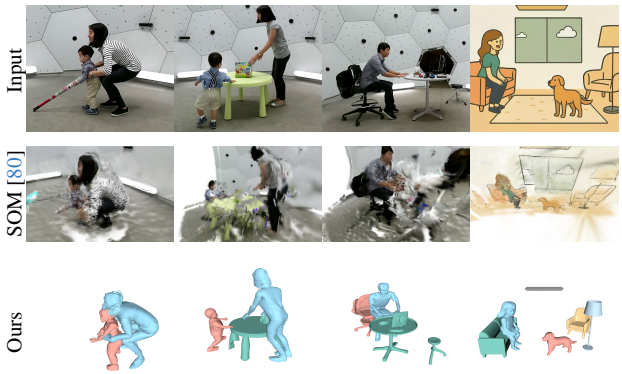

Figure 15. Qualitative comparison across methods.

I. Evaluation Using a Synthetic Dataset

We additionally evaluate our method on a synthetic 4D dataset since a suitable real-captured 4D scene benchmark is unavailable. We first generate candidate synthetic scenes by composing one or more dynamic humanoid motion sequences with several static objects. Each dynamic object is selected from a motion sequence, sub-sampled with a fixed frame stride, and placed into a shared normalized world coordinate system. Static assets are then scaled, rotated around the vertical axis, and translated so that they fit around the dynamic trajectories without excessive overlap. For every time step, we export a combined GLB that contains the static geometry and the current pose of each dynamic object, and we also save metadata describing the source assets, transforms, per-frame bounds, and global centering applied to the scene.

For visualization, each frame is rendered from a fixed orthographic camera with consistent framing across the entire sequence. The renderer keeps the scene centered and uses the same camera and lighting conventions for all frames within a sample, which makes frame-to-frame motion directly comparable.

For quantitative evaluation, we compare the generated scene-level mesh against the static part of the reference scene and compute dynamic scores per moving object and per frame. For static geometry, we compute one CD/F-score pair per scene and report the arithmetic mean over scenes. For dynamic content, we first average all valid dynamic per-frame scores within each sample, and then average these per-sample dynamic means across the whole set. This prevents samples with more moving objects or more valid frames from dominating the final measure.

Component	CD ↓	F-score ↑
Static scene geometry	0.293	0.418
Dynamic objects	0.284	0.434

Table 4. Quantitative results on the synthetic 4D evaluation set. Lower CD is better; higher F-score is better.

These synthetic-scene scores are worse than the corresponding single-object 4D and static compositional 3D scene results. The main reason is that, although the reconstructed objects themselves and their coarse arrangements are often visually consistent with the input, the relative positioning of objects is still somewhat inaccurate. Small errors in placement and pairwise spatial relationships accumulate at the full-scene level and are directly penalized by both CD and F-score, leading to lower quantitative performance than in simpler single-object or static-only settings.

J. Additional Qualitative Results on Synthetic Compositional 4D Dataset

In Fig. 16, Fig. 17, Fig. 18 we show additional qualitative results on our synthetic compositional 4D dataset.

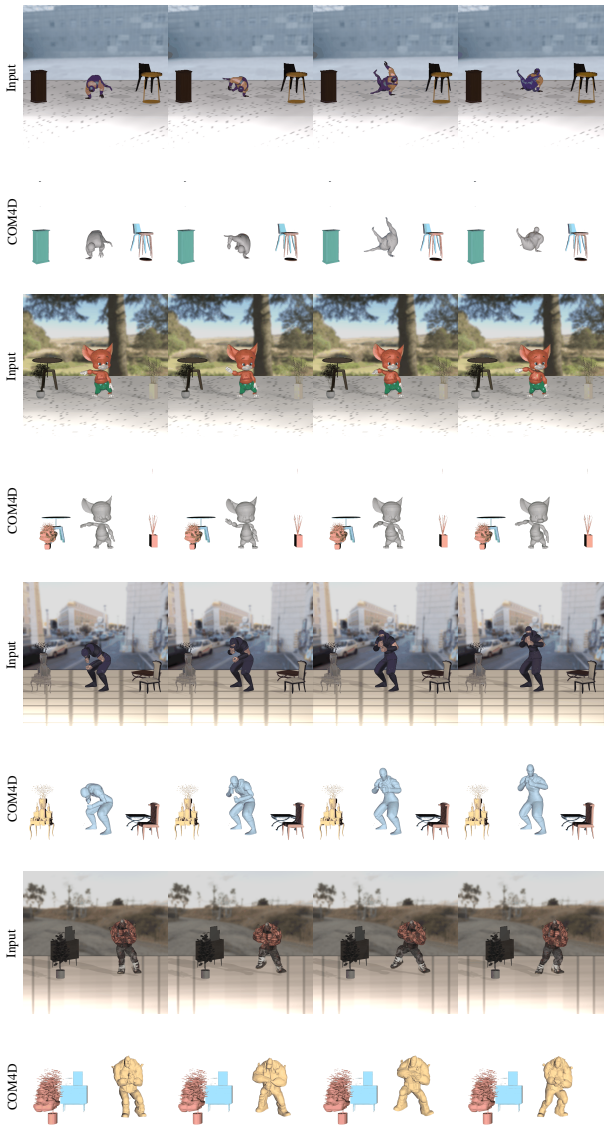

Figure 16. Additional qualitative results on the synthetic compositional 4D dataset.

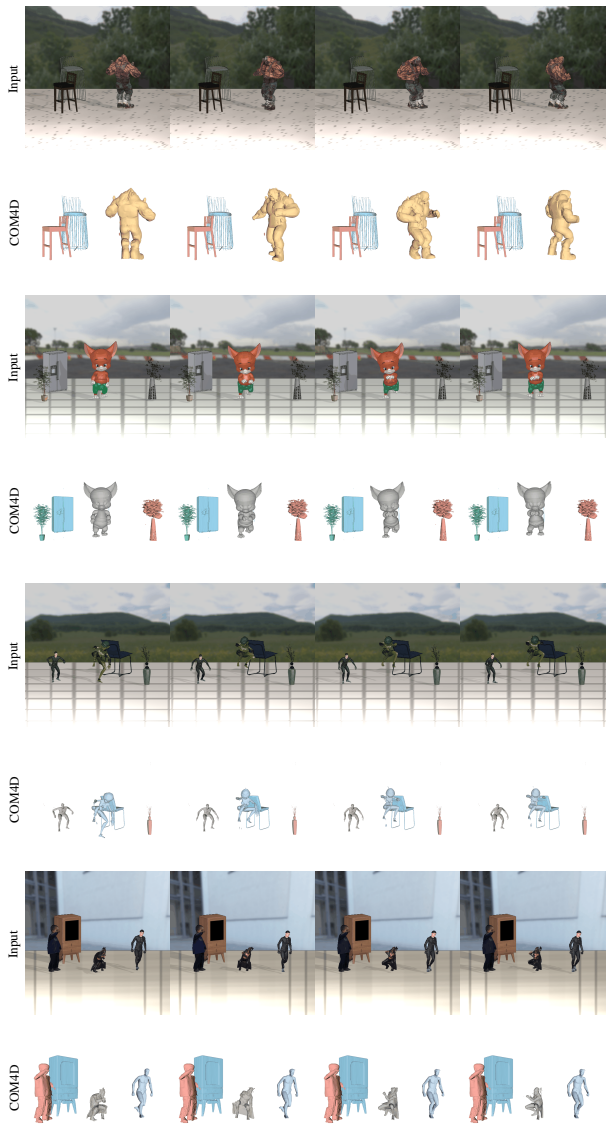

Figure 17. Additional qualitative results on the synthetic compositional 4D dataset.

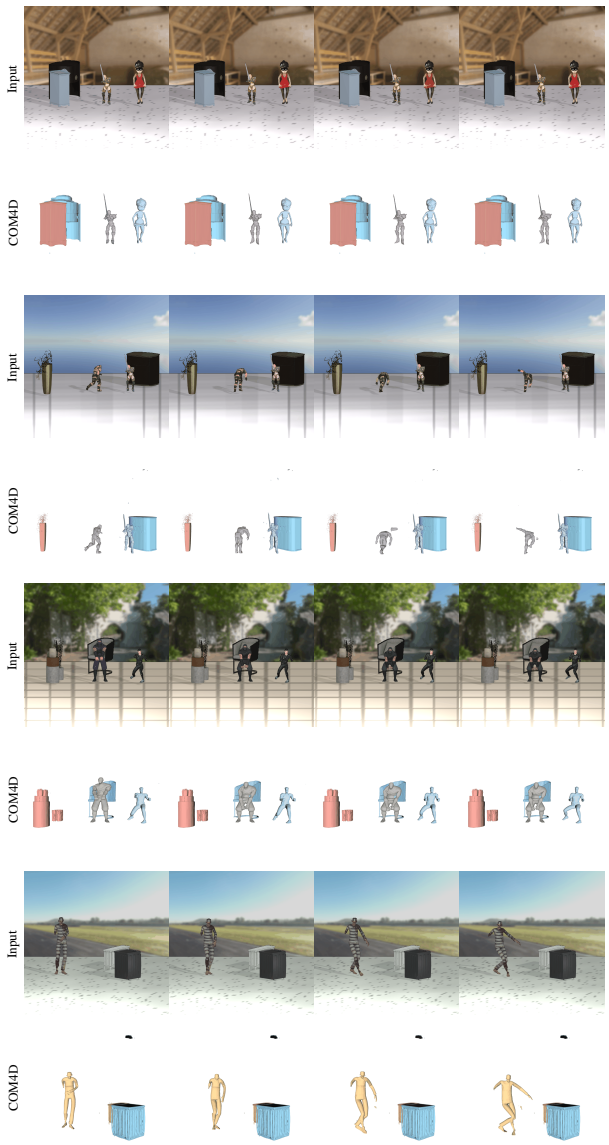

Figure 18. Additional qualitative results on the synthetic compositional 4D dataset.

K. Additional Qualitative Results on 4D Object Reconstruction

In Fig. 19, Fig. 20 and Fig. 21 we show more qualitative results of our model and baselines on single object 4D object reconstruction.

Figure 19. Qualitative 4D generation comparisons from Objaverse [15] showing three subjects at two time steps. For each model, we show the reconstructed input view (top) and a rendered novel view (bottom). We display the ground truth novel view in the bottom left.

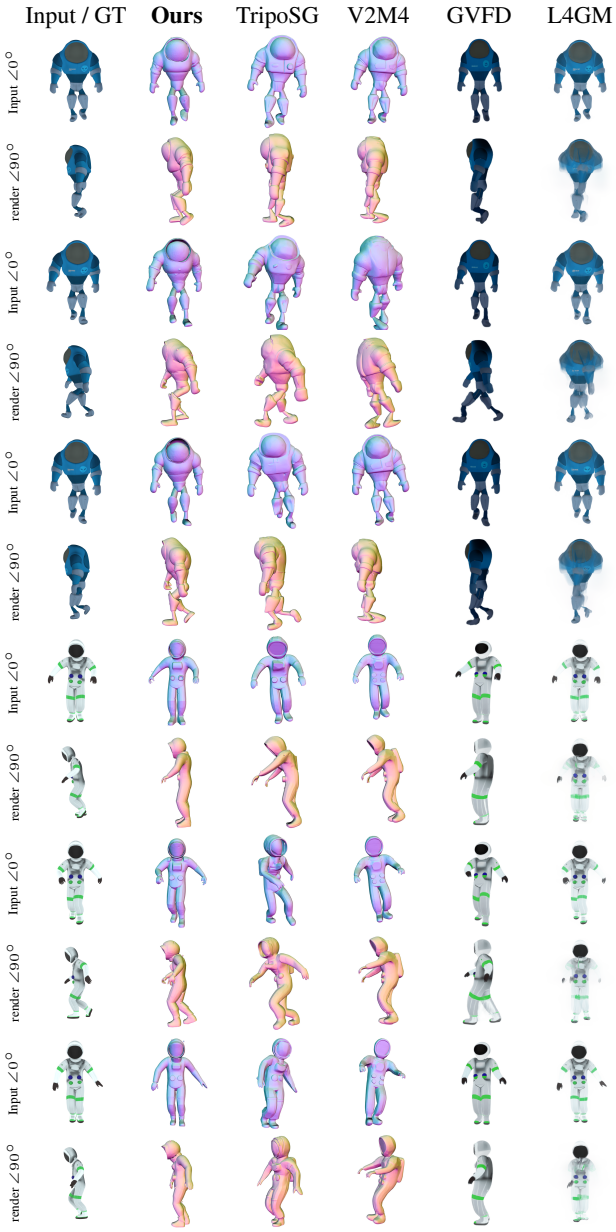

Figure 20. Further qualitative 4D generation comparisons from Objaverse [15] showing two subjects of Objaverse [15], at three time steps. For each model, we show the reconstructed input view (top) and a rendered novel view (bottom). We display the ground truth novel view in the bottom left. The novel view in particular highlights the discrepancies of the methods' outputs from the ground truth. Both TripoSG and V2M4 show moderate and consistent misalignment. Similar misalignment, particularly in rotation and skeletal pose is apparent in GVFD, while L4GM often fails to provide good novel views. Typically, our method shows far less shape or pose misalignment, as reflected in the quantitative metrics.

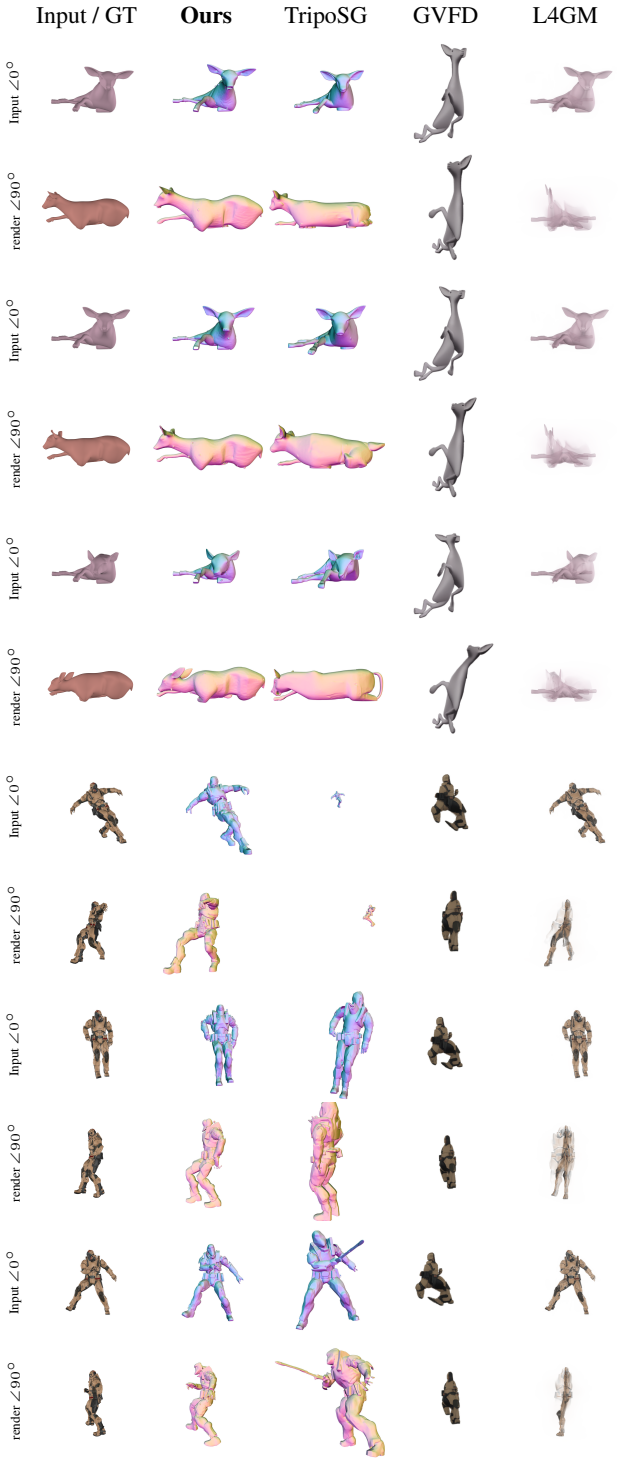

Figure 21. Qualitative 4D generation comparisons from DeformingThings [42] on two subjects at three time steps. For each model, we show the reconstructed input view (top) and a rendered novel view (bottom). We display the ground truth novel view in the bottom left. Compared to Fig. 20, the sequences contain stronger motion thus showing even larger difference in performances of our method. In particular, V2M4 fails completely due to the large motion.

L. Additional Qualitative Results on 3D Scene Generation

Figure 22. Qualitative comparison across ours, PartCrafter [47] and MIDI [29]. We show qualitatively how our method shows better performance, as shown by the quantitative metrics. Largely, this is due to the consistent reconstruction of all parts and their accurate composition. Both PartCrafter [47] and MIDI [29] often miss large objects, *e.g.*, the bed.