SC4D: Sparse-Controlled Video-to-4D Generation and Motion Transfer Supplemental File

Zijie Wu[1,2⋆], Chaohui Yu[2], Yanqin Jiang[2], Chenjie Cao[2], Fan Wang[2], and Xiang Bai[1†]

[1] Huazhong University of Science and Technology
[2] DAMO Academy, Alibaba Group
{zjw1031,xbai}@hust.edu.cn,
{huakun.ych,jiangyanqin.jyq,caochenjie.ccj,fan.w}@alibaba-inc.com
https://sc4d.github.io/

1 More Implementation Details

Video-to-4D Generation. In the coarse stage, we perform densification and pruning every 100 iterations in the first 1,000 iterations with a gradient threshold of 0.01 and opacity threshold of 0.01. Then in the remaining 500 iterations of the coarse stage, we only perform pruning with an interval of 100 iterations. The sampled timestep in the coarse stage is between [600, 800] in a linear decrease manner to ensure shape plausibility. In the fine stage, we only perform pruning every 1,000 iterations instead of densification. And the sampled timestep decreases from 800 to 200 for detailed texture. In both stages, the learning rates for Gaussians' rotation, opacity, scaling, and color equal 0.005, 0.05, 0.05, and 0.01, respectively. The learning rate for control points' control radius is set to 0.005. The Gaussians' position learning rate in the coarse stage decreases linearly from 0.01 to 0.0002 and remains 0.0002 in the fine stage. The learning rate for deformation MLP parameters equals 0.0002 in the coarse stage, and then gradually decreases to $2e^{-6}$ in the fine stage.

Motion Transfer Application. During the first 1,000 training iterations, the sampled timestep undergoes a linear decrease from 980 to 20. In the following 1,000 iterations, the timestep reduces from 200 to 20. The dense Gaussians' position learning rate is set to 0.0002, while other trainable Gaussian parameters are consistent with those used in in the video-to-4D pipeline.

Temporal Error. As described in the main paper, we utilize Temporal Error as a measurement for temporal consistency and motion fidelity. To do so, we adopt the RAFT [5] model pretrained on KITTI2015 [4] dataset provided by MMFlow [1] as the optical flow estimator. Given a video of the generated dynamic object projected from the reference view and the corresponding reference video, we first feed the initial two frames of the reference video into the optical

⋆ Work done during internship at DAMO Academy, Alibaba Group
† Corresponding author

flow estimator to obtain the predicted forward optical flow. Subsequently, we use the predicted optical flow to warp the first frame of the synthesized video. The warped frame is then compared against the second frame of the synthesized video by computing the L2 distance between them, which serves as the final result. We employ this procedure to calculate the corresponding outcomes for all adjacent frames of the video, and the average of these results is used as the final Temp Error (Temp).

2 Details of User Study

In the user study questionnaire, we choose 10 videos with large motion from the Consistent4D [2] dataset, which are: *triceratops*, *guppie*, *robot*, *rabbit*, *patrick*, *jack*, *ironman*, *elephant*, *egret*, *aurorus*, respectively. Utilizing the above videos as references, we train the compared methods to obtain the dynamic objects. We then project each object from a reference viewpoint and a random viewpoint, yielding $2\times3 = 6$ options for each question. We ask each participant to choose one option from the reference viewpoint and one from the novel viewpoint according to the reference view alignment, spatio-temporal consistency, and motion fidelity. The options for each question are randomly shuffled to avoid biases.

3 More Quantitative Comparisons

Following Consistent4D [2], we use the same testset and evaluation script to measure the performance of the compared methods. The results are shown in Tab. 1. Different from the evaluation in our main paper, the testset in Consistent4D contains four novel view ground truth renderings, apart from the reference video frames. The metrics are calculated between projections of the generated dynamic object and the corresponding rendering under the same viewpoint. We argue that this evaluation approach is less reasonable since the novel view information of the compared methods originates from Zero123 [3], and the prior knowledge in Zero123 might not align with the real data. Moreover, the values in Tab. 1 are too close, which misaligns with the qualitative results and user study, further revealing the necessity of the evaluations in our main paper. Additionally, we observe that the test samples from the consistent4D dataset exhibit relatively minor movements. In order to better assess and compare the robustness of various methods in learning motion, we have supplemented with some examples that demonstrate larger movements.

Table 1: Quantitative evaluations according to Consistent4D [2].

Method	CLIP ↑	LPIPS ↓
Consistent4D [2]	0.91	**0.14**
4DGen [6]	0.91	**0.14**
SC4D(Ours)	**0.92**	**0.14**

4 More Ablation Studies

4.1 Motion Transfer Application

The results in Fig. 1 illustrate the key factors of our design of motion transfer application, where *w/o depth cond* replaces depth-controlnet guidance with SD guidance, *w/o prog depth* utilizes the current depth of new object as condition, and *w/o prog guidance* fixes the guidance scale to 7.5 as in the original depth-controlnet.

Fig. 1: Motion transfer ablation study.

4.2 Additional Example on GA Loss and AG Init

We attach more examples on GA loss and AG Init, which are shown in Fig. 2. The qualitative results align with the conclusions of the main paper.

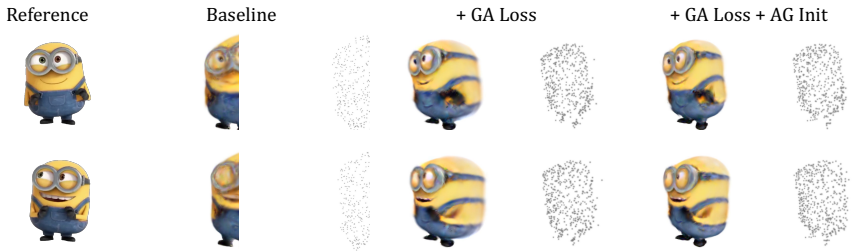

Fig. 2: Additional ablation study on GA Loss and AG Init.

4.3 Number of Control Points

We conduct experiments to verify the influences on the number of control points (denoted as M). We experiment on the example of *frog* here due to its complex motion. The results are shown in Tab. 2. As we can see, as M grows larger, the reference view alignment (PSNR, SSIM, LPIPS) and multi-view consistency (CLIP) of the generated dynamic object improve, which indicates that with more control points, the details of each view can be better restored. However, the temporal consistency degrades simultaneously, which reveals the degeneration of local rigidity when the number of control points is more significant. As a compromise, we choose $M = 512$ as our default setting.

Table 2: Ablation study on the number of control points (denoted as M).

M	PSNR ↑	SSIM ↑	LPIPS ↓	CLIP ↑	Temp ↓
128	26.46	0.934	0.084	0.890	**0.0175**
256	26.75	0.936	0.080	0.890	0.0176
512	27.00	0.936	**0.078**	**0.893**	0.0178
1024	**27.30**	**0.938**	0.079	**0.893**	0.0180

4.4 Comparison between GA Loss and the Chamfer Loss

As illustrated in Fig. 3, when utilizing the Chamfer loss to prevent the shape degeneration issue, the control points tend to cluster around certain points especially when the loss weight grows larger, resulting in an uneven distribution. In comparison, our proposed GA loss can prevent shape degeneration without hurting the overall distribution of control points.

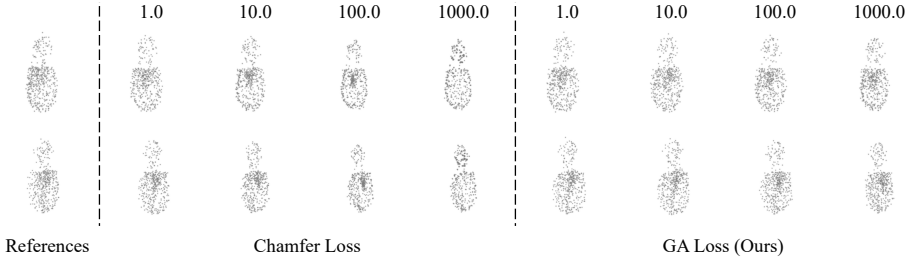

Fig. 3: Qualitative comparison on the effect of the proposed GA loss against the Chamfer loss. We attach the loss weight above each instance. The reference control points are the ones in the coarse stage from two different viewpoints.

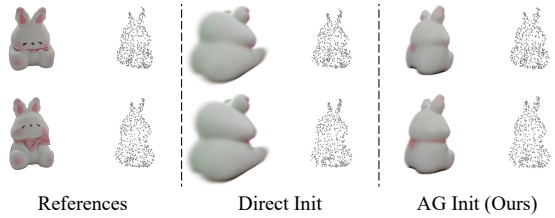

Fig. 4: Comparison between the proposed AG initialization and the Direct initialization. The latter represents directly utilizing the control Gaussians in the coarse stage as the initialization of the fine stage.

4.5 Comparison between AG Initialization and Direct Initialization

In the main paper, we compare the proposed AG initialization with *Ori init*, which represents initializing dense Gaussians within a sphere in the canonical space as in the coarse stage. In fact, there is another naive initialization approach that directly utilizes the control Gaussians as initialization of the fine stage, which we call *Direct init*. As shown in Fig. 4, *Direct init* also leads to shape degeneration issue, illustrated by the over-thickness of the deforming toy rabbit.

4.6 Necessity of the Coarse Stage

we conduct experiments that directly optimizing the control points and dense Gaussians from scratch. Fig. 5 shows that, without the proposed techniques, the generated dynamic object also encounters shape degeneration issue, similar to Fig. 7 of the main paper. Moreover, optimizing the control points and dense Gaussians from scratch also leads to motion misalignment of control points and the reference.

Fig. 5: Initializing control points and GS together comparison.

5 More Qualitative Comparisons

We show more qualitative comparisons in Fig. 6. The results illustrate the effectiveness and robustness of SC4D.

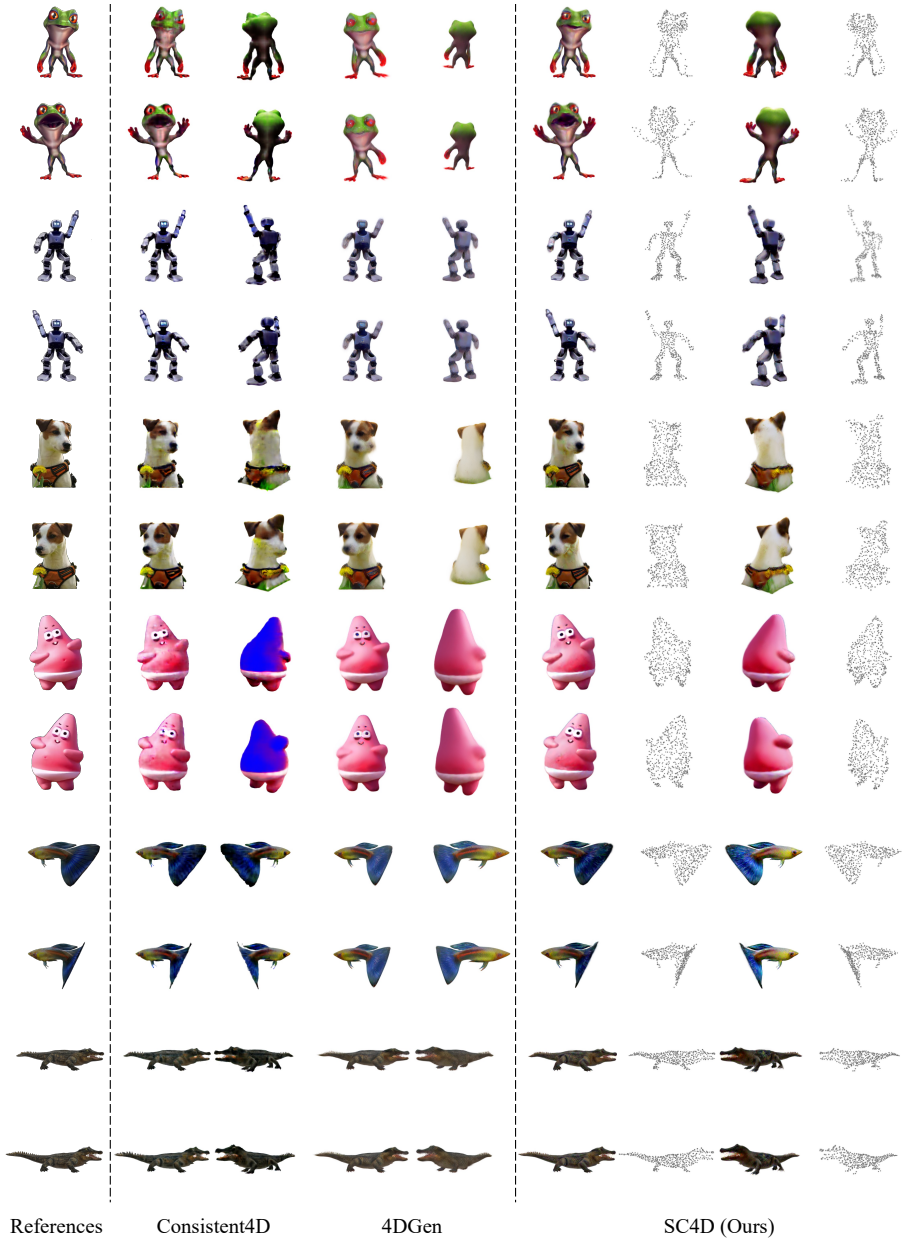

Fig. 6: Additional qualitative comparisons. (Best viewed when zoomed in.)

6 More Application Examples

We show more application examples in Fig. 7. These examples demonstrate the flexibility of the proposed motion transfer application.

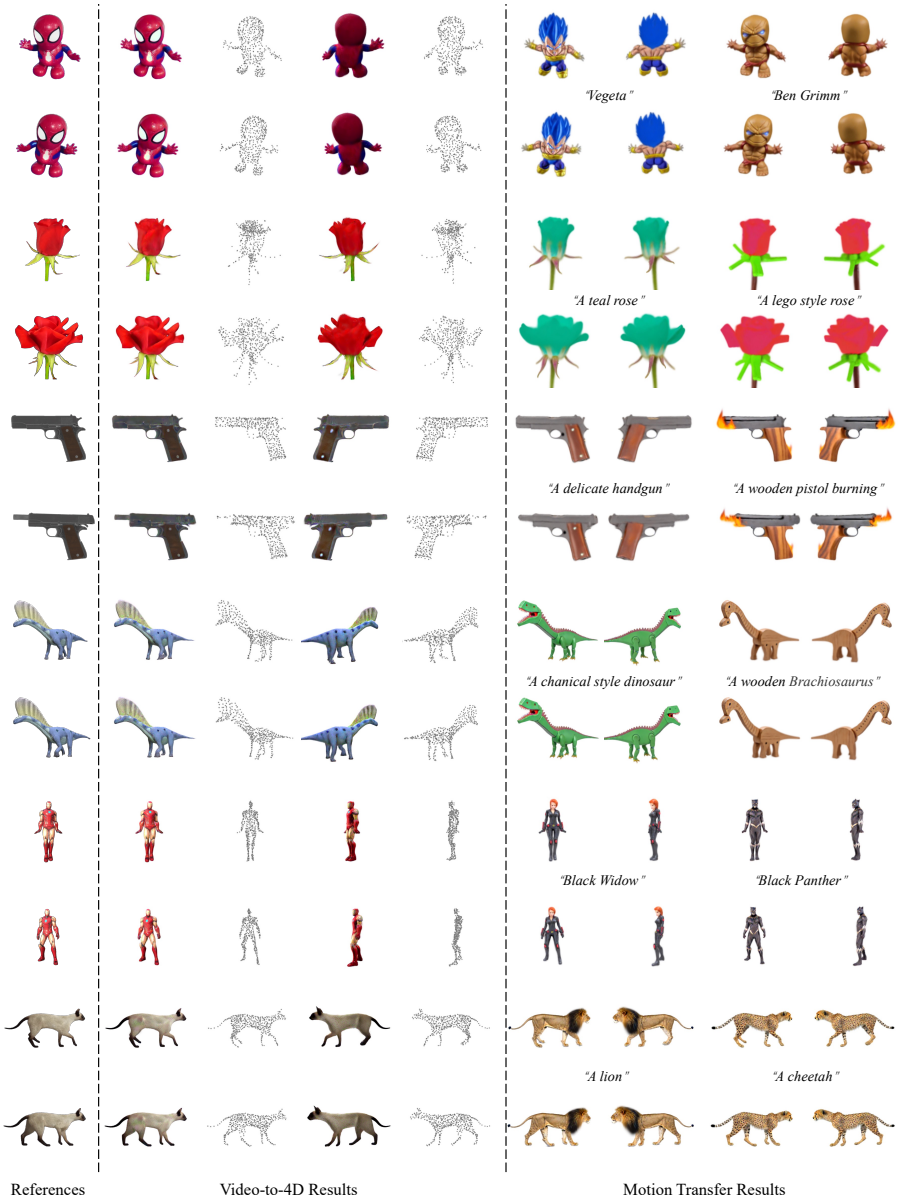

Fig. 7: Additional application examples. (Best viewed when zoomed in.)

References

1. Contributors, M.: MMFlow: Openmmlab optical flow toolbox and benchmark. `https://github.com/open-mmlab/mmflow` (2021)

2. Jiang, Y., Zhang, L., Gao, J., Hu, W., Yao, Y.: Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video. arXiv preprint arXiv:2311.02848 (2023)
3. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9298–9309 (2023)
4. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3061–3070 (2015)
5. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 402–419. Springer (2020)
6. Yin, Y., Xu, D., Wang, Z., Zhao, Y., Wei, Y.: 4dgen: Grounded 4d content generation with spatial-temporal consistency. arXiv preprint arXiv:2312.17225 (2023)