Supplementary Materials for STAG4D: Spatial-Temporal Anchored Generative 4D Gaussians

Yifei Zeng[1], Yanqin Jiang[2], Siyu Zhu[3], Yuanxun Lu[1], Youtian Lin[1], Hao Zhu[1], Weiming Hu[2], Xun Cao[1], and Yao Yao[1]

[1] Nanjing University
[2] CASIA
[3] Fudan University

1 Graident Distribution Analysis for Adaptive Densification

In Fig 1, we present a visual examination of the gradient distribution for the Duck and Bird scenarios across various training epochs. This analysis elucidates the rationale behind the ablation study of our adaptive densification approach. The graphical data indicates that the gradient distributions for different scenarios converge to disparate values, signifying a substantial numerical disparity when plotted on a logarithmic scale. For the Duck scenario, a suitable gradient threshold is approximately 40 ($e^{3.7}$), whereas for the Bird scenario, it is around 150 (e^5). Moreover, the gradient distribution maintains a consistent shape throughout the training phase for each case. Consequently, we introduce the adaptive densification mechanism that dynamically modulates the threshold in response to the specific scenario, which is to densify the Gaussian points that exhibit relatively higher gradient accumulation. This strategy is anticipated to enhance the robustness and quality of the generative performance.

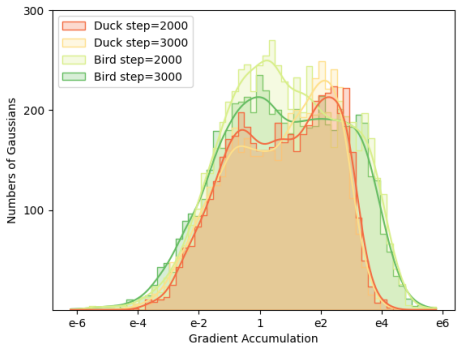

Fig. 1: Gradient Distribution Analysis for Adaptive Densification

2 User Study

2.1 Video-to-4D

In the conducted user study focusing on Video-to-4D content generation, we juxtaposed our methodology against Consistent4D [2] and 4DGen [5] across 14 test instances. Participants were solicited to appraise the methods according to visual quality (Vis.), temporal consistency (Cons.), and alignment with the input videos (Align.). As illustrated in Table 1, 4DGen did not receive any endorsements in the evaluated categories. In contrast, Consistent4D was acknowledged with 28.6% for Vis. and Cons., and 35.7% for Align. Predominantly, our method was favored, securing 71.4% of the votes for Vis. and Cons., and 64.3% for Align. These outcomes attest to the superior efficacy of our method in synthesizing high-fidelity and temporally coherent Video-to-4D content.

	Vis.	Cons.	Align.
Consistent4D	28.6%	28.6%	35.7%
4DGen	0%	0%	0%
Ours	**71.4%**	**71.4%**	**64.3%**

Table 1: User study on the best-performing Video-to-4D generation methods.

2.2 Text/Video-to-4D

A user study was conducted to evaluate the efficacy of Text&Video-to-4D content generation. For this purpose, 4Dfy [1] and DreamGaussian4D [3] were selected as comparative benchmarks. The study encompassed 14 test scenarios, and 30 evaluators were recruited to assess the methods based on visual quality (Vis.), temporal consistency (Cons.), and congruence with the input text (Align.). As delineated in Table 2, our approach garnered the highest ratings across all metrics. These findings underscore our method's preeminence and adaptability in the Text&Image-to-4D content generation domain.

	Vis.	Cons.	Align.
4Dfy	2.04	2.15	2.26
DG4D	2.91	2.92	3.26
Ours	**4.02**	**4.41**	**4.15**

Table 2: User study for Text&Image-to-4D generation methods.

3 Attention Mechanism for Normal Map

Building upon our attention mechanism, we have integrated our design with the newly introduced normal map model by Zero123++ [4]. This adaptation enables the generation of coherent multiview images with corresponding normal maps, as depicted in Figure 2. The successful application of our method to normal maps exemplifies the flexibility and applicability of our attention framework. This enhancement not only broadens the scope of our model but also demonstrates its potential for diverse 4D content generation scenarios.

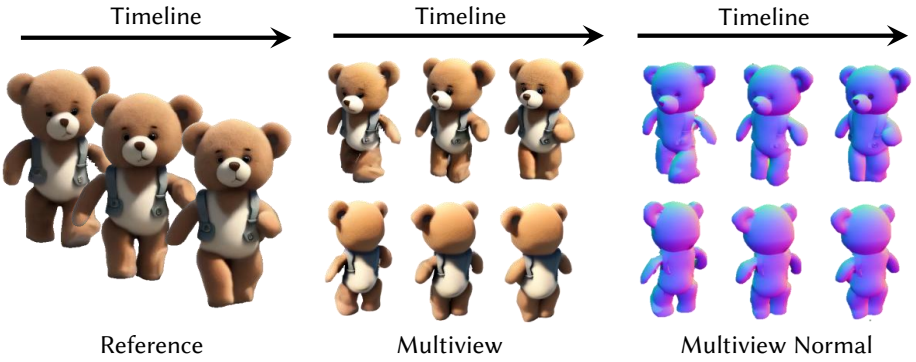

Fig. 2: Normal Map Illutstration

4 More Results for Multiview Generation

We show other results of our spatial and temporal consistent multiview images in Figure 3 & Figure 4. These results demonstrate the robustness and effectiveness of our method.

5 More Results for 4D Generation

We show more results of our 4D Generation in Figure 5. Our method could handle various objects and motions, which applies widely to different cases.

6 Discussion on the Potential Negative Impact

The generation process of our method involves various pre-trained diffusion models, which still have a controversy on the copyright of the generated result. This will remain an ethical issue until the relevant laws are matured.

References

1. Bahmani, S., Skorokhodov, I., Rong, V., Wetzstein, G., Guibas, L., Wonka, P., Tulyakov, S., Park, J.J., Tagliasacchi, A., Lindell, D.B.: 4d-fy: Text-to-4d generation using hybrid score distillation sampling (2023)
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. Ren, J., Pan, L., Tang, J., Zhang, C., Cao, A., Zeng, G., Liu, Z.: Dreamgaussian4d: Generative 4d gaussian splatting. arXiv preprint arXiv:2312.17142 (2023)
4. Shi, R., Chen, H., Zhang, Z., Liu, M., Xu, C., Wei, X., Chen, L., Zeng, C., Su, H.: Zero123++: a single image to consistent multi-view diffusion base model. arXiv preprint arXiv:2310.15110 (2023)
5. 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)

Fig. 3: More Results for Multiview Generation (Text input)

Fig. 4: More Results for Multiview Generation (Video input)

Fig. 5: More Results for 4D Generation