Supplementary Material of *GART*: Gaussian Articulated Template Models

Jiahui Lei[1] Yufu Wang[1] Georgios Pavlakos[2] Lingjie Liu[1] Kostas Daniilidis[1,3]
[1] University of Pennsylvania [2] UC Berkeley [3] Archimedes, Athena RC
{leijh, yufu, lingjie.liu, kostas}@cis.upenn.edu, pavlakos@berkeley.edu

Figure S.1. Additional application: Text-to-GART.

More implementation details are included in our code release at https://www.cis.upenn.edu/~leijh/projects/gart/. This document includes more results of our representation from the main paper, including an additional application – Text-to-3D generation in Sec. S.1 and more results on dogs in Sec. S.2. And provides more experiments and discussions in Sec. S.3.

S.1. Application: Text-to-GART Generation

GART is a general representation for articulated subjects and is not restricted to reconstruction from real monocular video. By changing the rendering $L1$ loss and SSIM loss in Eq.17 in the main paper to an SDS loss [6], we further demo an application – Text-to-GART. The input is a text describing the content the user aims to generate, and the output is an optimized *GART* representing this subject. The optimization loss becomes:

$$L = L_{\text{SDS}} + L_{\text{reg}}, \quad (1)$$

where L_{SDS} is computed via forwarding a fine-tuned Stable-Diffusion [7] model MVDream [9]. For more details on

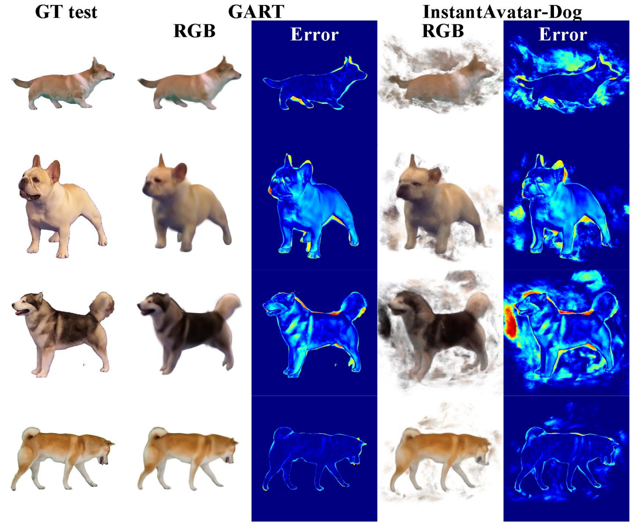

Figure S.2. Additional comparison on in-the-wild dog videos.

Data	Method	PSNR	SSIM	LPIPS
National Dog Show	InsAva-Dog	16.13	0.759	0.318
	GART	**17.86**	**0.825**	**0.238**
Adobe Stock	InsAva-Dog	20.62	0.834	0.227
	GART	**24.50**	**0.921**	**0.114**

Table S.1. Quantitative evaluation of view synthesis on ITW dogs.

L_{SDS}, please see Stable-Diffusion [7] and DreamGaussian [10]. Since there are no real poses estimated from video frames, we randomly sample some reasonable SMPL [2] template poses from AMASS [3] to augment *GART* during distillation. The generation results are shown in Fig. S.1. We observe that thanks to the efficiency of *GART*, the computation bottleneck of the generation is mainly in the 2D diffusion forwarding, and the typical generation time is around 10 minutes per subject on a single A40 GPU.

S.2. More Results on Dogs

GART can robustly reconstruct dogs from challenging in-the-wild videos. We further compare it to a NeRF-based approach [1, 4, 5], which we call InstantAvatar-Dog. We adapt the implementation of InstantAvatar [1] by changing the template model to D-SMAL [8] and applying it to the dog videos. Qualitative comparison from Fig S.2 shows that InstantAvatar-Dog produces ghostly artifacts similar to InstantAvatar's results on human bodies. These artifacts may be the result of inaccurate pose estimation and insufficient viewpoints in the training data, and they are more pronounced on the dogs due to the challenging in-the-wild sequences and the less accurate dog pose estimation [8]. An additional quantitative comparison is presented in Tab. S.1. *GART* has higher performance across all view synthesis metrics.

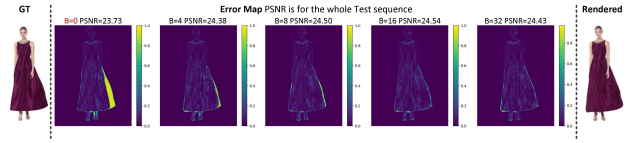

Figure S.3. Additional comparison on the effect of the number of latent bones.

	PSNR	SSIM	LPIPS*
UBC	25.65	0.9337	81.88
Sum1	**25.71**	**0.9347**	**76.93**
ZJU	32.22	0.9773	29.21
Sum1	**32.24**	**0.9774**	**29.11**
People	28.36	**0.9701**	46.49
Sum1	**28.37**	**0.9701**	**46.03**

Table S.2. Normalize the skinning weight to sum up to 1 on different datasets. The results reported in the main paper are colored yellow.

Figure S.4. More ablation of the smoothness regularization.

S.3. More Experiments and Discussions

S.3.1. Latent Bones

The main results in the paper use 32 latent bones for the UBC dataset. As shown in Fig. S.3, we further ablate the number of latent bones used for modeling challenging long cloths. Another limitation of our current latent bone method is the generalization to novel poses. Since the training poses are too limited, the latent bones tend to overfit the training poses and produce reasonable results only on similar poses. It's an open question of how to generalize this method to novel pose animation.

S.3.2. Skinning Weights

The learnable skinning weight in the main paper is not normalized to have a summation of 1 per Gaussian. We further verify this setup in Tab. S.2 and observe a slight but consistent improvement in the performance of our method.

S.3.3. More Ablation

We show more results for the ablation of voxel grid distilled skinning weights and the KNN regularization in Fig. S.4. And test the voxel grid resolution's effect on the performance in Tab. S.3.

References

[1] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60

vox res	RunTime(s)	PSNR	SSIM	LPIPS*
16	108.2	31.66	0.975	32.16
32	112.4	32.03	0.977	29.92
64	116.7	32.22	0.977	29.21
128	213.8	32.31	0.977	29.64

Table S.3. Different voxel grid resolution on ZJU dataset. The results reported in the main paper are colored yellow.

seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16922–16932, 2023. 2

[2] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2

[3] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019. 2

[4] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[5] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 2

[6] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2

[8] Nadine Rüegg, Shashank Tripathi, Konrad Schindler, Michael J Black, and Silvia Zuffi. Bite: Beyond priors for improved three-d dog pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8867–8876, 2023. 2

[9] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023. 1

[10] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 2