Supplementary Materials of Gloria: Consistent Character Video Generation via Content Anchors

Yuhang Yang[1], Fan Zhang[2], Huaijin Pi[3], Ailing Zeng[5], Shuai Guo[1], Guowei Xu[4]
Wei Zhai[1,†], Yang Cao[1], Zheng-Jun Zha[1]
[1] MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, USTC
[2] UNSW [3] HKU [4] UESTC [5] Independent Researcher
†Corresponding Author, https://yyvhang.github.io/Gloria_Page/

1. Appendix

1.1. Details of Data Pipeline

Here, we provide additional details on the pipeline for extracting anchor frames. For the global anchor, since we can control how each 5s training clip is segmented from a long video, we know the exact start and end timestamps of each clip. This allows us to extract global frames both inside and outside the training clip.

Regarding the viewpoint anchor, we leverage the world-grounded human motion recovery method GVHMR [11]. It represents human motion using the 3D body model SMPL [8] (a widely used parameterized human model [2–4, 7, 13–16]) and recovers human motion directly from videos. Specifically, it estimates the facing direction of the person, which can be represented by the Z-axis orientation of the SMPL root joint. In addition, such methods typically use SLAM [12] to estimate the camera pose. By computing the angle θ between the camera orientation and the human orientation, and determining a threshold based on empirical observations across multiple test cases, we can distinguish frames in which the person is facing front, backward, left, or right. These categorized frames are then extracted as viewpoint-anchor frames.

For expression anchors, we adopt a hybrid extraction strategy. We first use EmotiEffLib [9, 10] to scan video clips and identify frames containing any of the predefined eight expression categories, selecting only clips that include at least two distinct expressions. Although this method effectively identifies candidate clips, the expression category assigned to each extracted anchor frame may still be ambiguous. To address this issue, we further use an MLLM (*e.g.*, Gemini [1]) to verify whether the predicted expression matches the semantic content of each anchor frame. If the prediction is incorrect, we correct the frame to the appropriate category through the MLLM; if it does not belong to any defined category, we discard it. This refinement process increases the accuracy from 66% to 82%.

Table 1. Ratios of different categories in viewpoint-anchor.

Viewpoints	**Front**	**Back**	**Left**	**Right**
Ratio	29.6%	17.3%	27.9%	25.2%

1.2. Training and Inference Details

Here we provide additional training details for content anchors. First, video clips that simultaneously contain different expressions and viewpoints are relatively scarce. Thanks to our proposed "RoPE as Weak Condition", which binds each type of content anchor to a fixed position, the model can be trained in a mixed manner to learn how to extract information from different types of anchors. Specifically, we first balance the number of anchors across all categories, with the final ratio of viewpoint anchors is shown in Tab. 1. The ratio of expression anchors is reported in Tab. 2. Then, we ensure that training batches contain a balanced mixture of expression-anchor frames and viewpoint-anchor frames, and training with a batch size of 256 for 3000 steps. Afterward, we perform a fine-tuning stage (200 steps) on a small subset of data that contains both types of anchors, allowing the model to adapt to scenarios where multiple anchor types appear simultaneously. Note that the global anchor remains throughout the entire training process.

Regarding the linear blend in continuation generation, let the last n frames of the previous chunk be x_1, and the first n frames of the next be x_2, blended as: $w * x_1 + (1-w) * x_2$, $n = 4$, $w = [1, 0.67, 0.33, 0]$, applied at the clean latent level.

1.3. Subject Evaluation Protocol

Regarding the subject evaluation, we refer to certain dimensions of existing benchmarks [5, 6, 17, 18] and make an A/B testing protocol. For each evaluation dimension, annotators

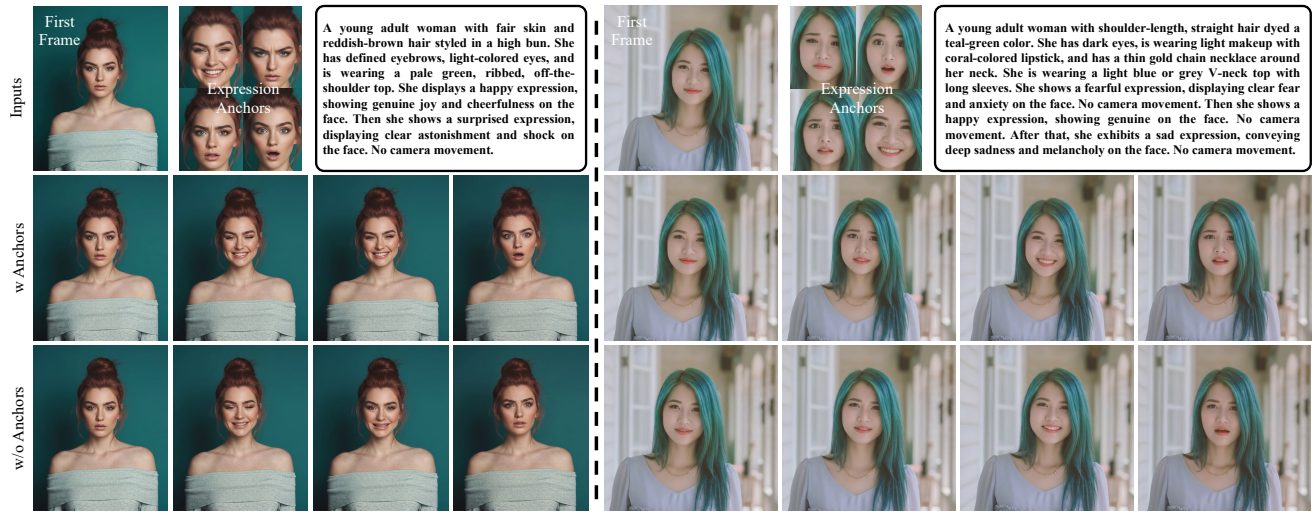

Figure 1. Qualitative results of the expression anchors. The first row shows the inputs, the second row shows the results with the expression anchor, and the third row shows the results without the anchors.

Table 2. Ratios of different categories in expression-anchor.

Expressions	**Surprise**	**Angry**	**Disgust**	**Fear**
Ratio	15.0%	8.2%	13.7%	10.1%
Expressions	**Contempt**	**Neutral**	**Happy**	**Sad**
Ratio	13.6%	8.1%	14.6%	16.7%

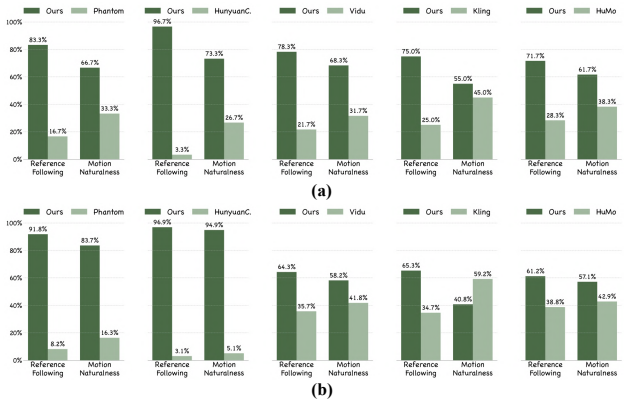

Figure 2. Subject Evaluation of **(a)** expressive identity consistency, and **(b)** multi-view appearance consistency.

are presented with a specific question to choose the better result, and they are also asked to provide the reason for the choice. Below, we list several example questions for reference.

Questions for reference-based aspects:

- Between the two videos below, which one better preserves the multi-view appearance details shown in the reference images?
- Between the two videos below, which one better follows the prompt-specified expression, with facial details more closely matching the reference image?
- Between the two videos below, which one exhibits more natural overall motion when reference frames are provided?

Questions for general aspects:

- Consider the following aspects for video color consistency: severity and frequency of overall color tone changes, background color shifts, and visual inconsistencies throughout the entire video. Which of the provided videos maintains the better overall color consistency?
- Consider the following aspects for movement naturalness: fluidity and continuity of actions and expressions, coordination between different body parts, naturalness of facial expression changes, realism and believability of movements, and overall naturalness of all movements throughout the entire video. Which videos have higher movement naturalness?
- Consider the following aspects for lip sync quality: severity and frequency of timing synchronization misalignment, pronunciation lip shape mismatch, and movement amplitude mismatch throughout the entire video. Which of the provided videos has higher lip sync quality?

1.4. More Experimental Results

We provide more qualitative results illustrating the effect of the expression anchor. As shown in Fig. 1, when the expression anchor is provided, the character produces more accurate facial expressions, and the facial details remain more consistent across different expressions.

In the main paper, we present the overall subject evalu-

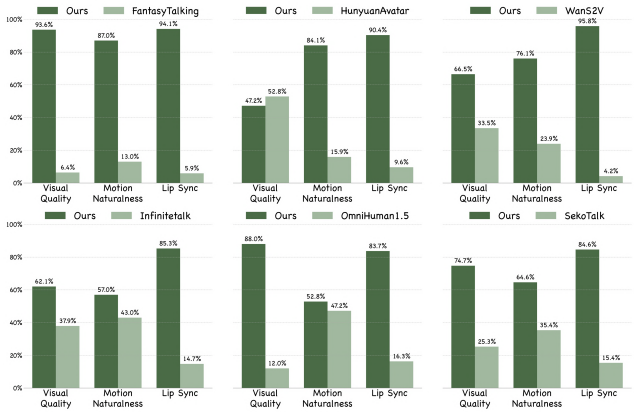

Figure 3. Comparison of subject evaluation results with human-centric models across multiple evaluation dimensions.

Table 3. Quantitative comparison results on CelebV-HQ with the human-centric methods. Fantasy. denotes FantasyTalking, Huny.A. is HunyuanAvatar and Infinite. indicates InfiniteTalk.

Method	IQA ↑	AES ↑	Sync-C ↑	Sync-D ↓	FID/FVD ↓
Fantasy.	3.48	2.26	2.88	10.39	37.43 / 108.92
Huny.A.	3.40	2.28	5.17	8.37	28.39 / 93.49
WanS2V	3.56	**2.33**	4.99	8.40	36.10 / 93.73
Infinite.	3.47	2.31	5.20	8.29	29.56 / **76.39**
Ours	**3.59**	**2.33**	**5.62**	**8.06**	**25.50** / 90.86

ation results. Here, we provide more detailed results along individual dimensions: identity consistency across expressions is shown in Fig. 2 (a) and multi-view appearance consistency is shown in Fig. 2 (b), and comparisons with human-centric models are shown in Fig. 3.

In addition to the results on the combined testset from human-centric baselines reported in the main paper, we also provide comparisons against these baselines on public datasets such as CelebV-HQ [19]. We use 100 cases in total, and the results are summarized in Tab. 3.

We also compare the 5B and 17B models during training (10M data): the 5B model is weaker than the 17B across all dimensions. Regarding the data, we report the performance (17B) under 2M and 10M training clips, shown in Tab. 4. This also aligns with the user study results.

Table 4. Performance in different model sizes (a) and different data scales (b).

Size	IQA ↑	AES ↑	Sync-C ↑	Arcface ↑
5B	4.58	3.54	4.93	0.583
17B	4.65	3.63	5.12	0.787

(a)

Data	IQA ↑	AES ↑	Sync-C ↑	Arcface ↑
2M	4.53	3.57	5.03	0.623
10M	4.65	3.63	5.12	0.787

(b)

References

[1] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1

[2] Ke Fan, Junshu Tang, Weijian Cao, Ran Yi, Moran Li, Jingyu Gong, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Freemotion: A unified framework for number-free text-to-motion synthesis. In *European Conference on Computer Vision*, pages 93–109. Springer Nature Switzerland Cham, 2024. 1

[3] Ke Fan, Jiangning Zhang, Ran Yi, Jingyu Gong, Yabiao Wang, Yating Wang, Xin Tan, Chengjie Wang, and Lizhuang Ma. Textual decomposition then sub-motion-space scattering for open-vocabulary motion generation. *arXiv preprint arXiv:2411.04079*, 2024.

[4] Ke Fan, Shunlin Lu, Minyue Dai, Runyi Yu, Lixing Xiao, Zhiyang Dou, Junting Dong, Lizhuang Ma, and Jingbo Wang. Go to zero: Towards zero-shot motion generation with million-scale data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13336–13348, 2025. 1

[5] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1

[6] Hongxiang Li, Yaowei Li, Bin Lin, Yuwei Niu, Yuhang Yang, Xiaoshuang Huang, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Long Chen. Gir-bench: Versatile benchmark for generating images with reasoning. *arXiv preprint arXiv:2510.11026*, 2025. 1

[7] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1

[8] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.*, 2015. 1

[9] Andrey Savchenko. Facial expression recognition with adaptive frame rate based on multiple testing correction. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 30119–30129. PMLR, 2023. 1

[10] Andrey V Savchenko, Lyudmila V Savchenko, and Ilya Makarov. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 2022. 1

[11] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view

coordinates. In *SIGGRAPH Asia Conference Proceedings*, 2024. 1

[12] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021. 1

[13] Yabiao Wang, Shuo Wang, Jiangning Zhang, Ke Fan, Jiafu Wu, Zhucun Xue, and Yong Liu. Timotion: Temporal and interactive framework for efficient human-human motion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7169–7178, 2025. 1

[14] Lixing Xiao, Shunlin Lu, Huaijin Pi, Ke Fan, Liang Pan, Yueer Zhou, Ziyong Feng, Xiaowei Zhou, Sida Peng, and Jingbo Wang. Motionstreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10086–10096, 2025.

[15] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Lemon: Learning 3d human-object interaction relation from 2d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16284–16295, 2024.

[16] Yuhang Yang, Wei Zhai, Chengfeng Wang, Chengjun Yu, Yang Cao, and Zheng-Jun Zha. Egochoir: Capturing 3d human-object interaction regions from egocentric views. *Advances in Neural Information Processing Systems*, 37: 54529–54557, 2024. 1

[17] Yuhang Yang, Ke Fan, Shangkun Sun, Hongxiang Li, Ailing Zeng, FeiLin Han, Wei Zhai, Wei Liu, Yang Cao, and Zheng-Jun Zha. Videogen-eval: Agent-based system for video generation evaluation. *arXiv preprint arXiv:2503.23452*, 2025. 1

[18] Ailing Zeng, Yuhang Yang, Weidong Chen, and Wei Liu. The dawn of video generation: Preliminary explorations with sora-like models. *arXiv preprint arXiv:2410.05227*, 2024. 1

[19] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*, 2022. 3