Lookahead Anchoring: Preserving Character Identity in Audio-Driven Human Animation

Junyoung Seo[1], Rodrigo Mira[2], Alexandros Haliassos[2], Stella Bounareli, Honglie Chen, Linh Tran[2], Seungryong Kim[1], Zoe Landgraf[2⋆], and Jie Shen[2]

[1] KAIST
[2] Imperial College London

Abstract. Audio-driven human animation models often suffer from identity drift during temporal autoregressive generation. A common remedy generates keyframes as intermediate anchors, but this needs an extra keyframe stage and can restrict natural motion. We propose **Lookahead Anchoring**, which places keyframes at future timesteps *ahead* of the current generation window rather than within it. Since pretrained video Diffusion Transformers treat distant conditioning frames as softer references, positioning keyframes ahead turns rigid boundary constraints into soft directional guidance that preserves identity while allowing expressive motion. This enables self-keyframing from the reference image alone and makes temporal distance a control parameter for the identity–expressiveness trade-off. Across three human animation models, Lookahead Anchoring improves lip synchronization, identity preservation, and visual quality. Video results are available at https://lookahead-anchoring.github.io.

Keywords: Audio-driven human animation · Video generation

1 Introduction

Audio-driven human animation generates realistic human videos synchronized with input audio, with applications in film production, virtual assistants, and content creation. Diffusion Transformers (DiTs) [12] have advanced this field [2, 16], but can only handle short clips at a time (typically ∼5s) due to their quadratic complexity. Segment-wise autoregressive generation extends length by conditioning each segment on preceding frames [3,6,10], but suffers from identity drift: errors compound across segments, so the appearance deviates from the reference. Adding reference image conditioning [3,6] creates a conflict, as the model must honor the rigid final frames of the previous segment while loosely following the reference, and typically prioritizes the immediate frames, losing identity (Fig. 2).

Keyframe-based methods such as KeyFace [1] mitigate drift by generating lip-synced sparse keyframes as segment endpoints, but require a two-step keyframe-generation-and-interpolation pipeline and are bounded by keyframe quality. We

⋆ Project lead.

instead ask whether keyframes must act as rigid boundaries. Observing that pretrained video DiTs treat distant conditioning frames as softer, less constraining references, we propose **Lookahead Anchoring**, which positions keyframes perpetually *ahead* of the current window, always visible but never reached. This turns the keyframe from a rigid constraint into soft directional guidance that preserves identity while allowing expressive, audio-driven motion.

This offers two advantages. First, the keyframe need not match the audio at its timestamp, enabling *self-keyframing*: using the reference image itself as a recursive anchor, removing the keyframe generation stage. Second, the temporal distance becomes a control parameter trading identity adherence for expressiveness, with an optimal lip-sync range. We introduce a fine-tuning strategy over a continuous range of distances and validate across three recent DiT-based models [2,3,6], improving lip synchronization, character consistency, and visual quality.

2 Method

We generate arbitrarily long, audio-driven human animation videos while preserving character identity. Given an audio sequence $\mathbf{a}$ and a reference image I_{ref}, the goal is a video that keeps consistent identity with I_{ref} while exhibiting natural motion synchronized to $\mathbf{a}$.

2.1 Preliminaries and Motivation

Autoregressive approaches divide the video into segments of length L [3,6]; segment V_i over frames $[iL, (i+1)L)$ is generated by $G_{\text{auto}}(\cdot)$ conditioned on the audio $\mathbf{a}_{iL:(i+1)L}$, the previous segment's final frames V_{i-1}^{end}, and the reference:

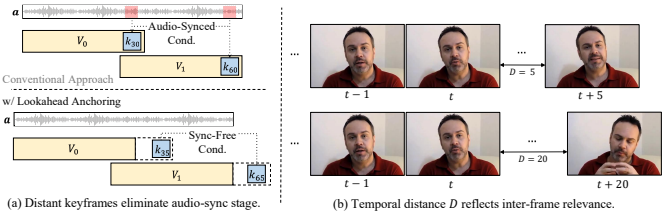
$$V_i = G_{\text{auto}}(\mathbf{a}_{iL:(i+1)L}, V_{i-1}^{\text{end}}, I_{\text{ref}}). \tag{1}$$

Errors in V_{i-1}^{end} accumulate across segments, causing identity drift, and reference conditioning (e.g., ReferenceNet [8] in Hallo3 [3], channel concatenation in OmniAvatar [6]) is insufficient for long-term consistency (Fig. 2). Keyframe methods such as KeyFace [1] instead add temporal anchors: a specialized model generates sparse audio-synced keyframes used as boundary constraints for an interpolation model. This reduces drift, but requires an extra keyframe model, forces exact poses at fixed timestamps (suppressing dynamics), and is upper-bounded by the keyframe quality.

2.2 Lookahead Anchoring

Instead of forcing the model to reach keyframes at segment boundaries, we position them perpetually ahead, always visible but never reached (Fig. 1). Our model conditions on a keyframe placed D frames beyond the segment endpoint, where the lookahead distance $D \in \mathbb{N}, D > 0$ is a control parameter:

$$V_i = G_{\text{LA}}(\mathbf{a}_{iL:(i+1)L}, V_{i-1}^{\text{end}}, k_{(i+1)L-1+D}). \tag{2}$$



(a) Distant keyframes eliminate audio-sync stage. (b) Temporal distance D reflects inter-frame relevance.

Fig. 1: Motivation. (a) Instead of using conditional keyframes as generation-window endpoints, we reposition them as temporally distant anchors beyond the window, eliminating constraints such as audio synchronization. (b) Models learn that longer temporal distances allow greater scene variation, so distant keyframes provide high-level guidance without strict physical constraints.

The model generates expressive motion that progressively "*chases*" the target, continuously receiving guidance but never reaching it. A shorter distance D grounds identity, while a longer one affords greater motion freedom, systematically trading facial consistency for dynamics. Since precise audio–pose alignment at the distant timestamp is unnecessary, any image can serve as the anchor; in particular, *self-keyframing* uses the reference itself ($I_{\text{trg}} := I_{\text{ref}}$), *eliminating the keyframe generation stage* while assigning I_{ref} a future temporal position that makes the model interpret it as a future target, creating a persistent identity pull.

2.3 Leveraging Lookahead Anchoring for Video DiTs

Video DiTs patchify compressed latents into tokens with positional embeddings $p_{x,y,t} = p_x \oplus p_y \oplus p_t$; $p_t[\ell]$ is the temporal embedding at latent frame ℓ (n latent frames per L original frames, $r = L/n$). During training, we append a distant clean latent to $\mathbf{z} = \{z_0, ..., z_{n-1}\}$ at position $n{-}1{+}d$ ($d = \lfloor D/r \rceil$), mapped by a projection layer into the noisy token space; at inference it is replaced by z_{trg}, the encoded latent of I_{trg}, keeping the distant embedding $p_t[n-1+d]$. Pretrained DiTs already yield larger motion at larger distances but degrade without training on distant pairs, so we sample the keyframe position $\ell \sim \mathcal{U}[0, n-1+d_{\max}]$ ($d_{\max} > 0$) spanning inside and outside the window: inside ($\ell < n$) the model learns accurate reconstruction, outside ($\ell \geq n$) a smooth attenuation of the keyframe's influence with distance.

3 Experiments

Setup. We integrate our approach into three audio-driven human animation DiTs: Hallo3 [3], HunyuanVideo-Avatar [2], and OmniAvatar [6], fine-tuning on the Hallo3 data plus 160 hours of collected portrait video; all baselines use the same data. We evaluate on the HDTF [19] and AVSpeech [5] test splits, sampling videos over 30s (average 48s).

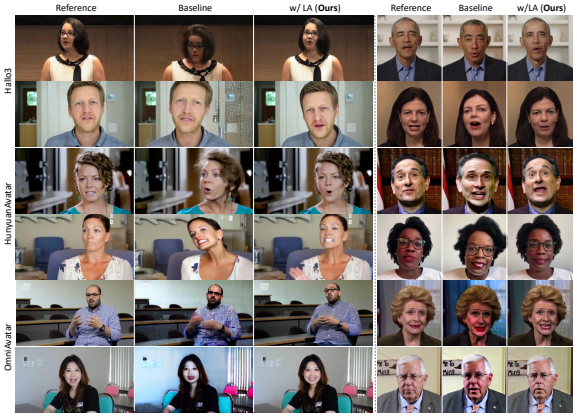


Fig. 2: Qualitative results. We compare our method with three audio-conditioned DiT baselines under the temporal segment-wise autoregressive framework on AVSpeech [5] and HDTF [19], presenting mid-sequence frames to demonstrate generation quality.

Table 1: Quantitative results on HDTF [19] under the temporal segment-wise autoregressive scheme. We compare DiT-based models with and without our approach, together with the concurrent DiT-based StableAvatar [14]. All baselines are fine-tuned on the same data. * HunyuanAvatar is modified to take 5 starting frames for the autoregressive scheme.

Models	Sync-D ↓	Sync-C ↑	Face-Con. ↑	Subj-Con. ↑	FID ↓	FVD ↓	MS. ↑	Dyn. ↑
StableAvatar [14]	11.29	4.34	0.9260	0.9843	11.76	137.82	0.9942	0.0850
Hallo3 [3]	7.89	7.79	0.8628	0.9774	21.61	180.30	0.9939	0.1042
+ **LA (Ours)**	**7.53**	**8.22**	**0.9267**	**0.9843**	**12.44**	**142.97**	**0.9939**	**0.1236**
HunyuanAvatar* [2]	7.77	8.27	0.6109	0.9470	37.84	501.80	0.9925	**0.1681**
+ **LA (Ours)**	**7.70**	**8.27**	**0.9135**	**0.9870**	**15.98**	**182.23**	**0.9946**	0.1451
OmniAvatar [6]	8.48	7.47	0.7904	0.9568	35.26	467.90	**0.9953**	0.0519
+ **LA (Ours)**	**7.19**	**9.28**	**0.8628**	**0.9686**	**24.09**	**302.71**	0.9923	**0.0793**

Quantitative comparisons. Tab. 1 reports HDTF [19] results. Following established protocols [16, 18], we report SyncNet [13] distance/confidence for lip sync, ArcFace [4] and DINO [17] similarity for face and subject consistency (VBench [9]), and FID [7], FVD [15], motion smoothness (MS) [9], and dynamic degree (Dyn.) [11] for overall quality, alongside the concurrent DiT-based StableAvatar [14]. Our approach consistently improves lip synchronization, facial consistency, and visual fidelity across all three baselines, and a user study among 34 participants strongly prefers it (64–84% overall quality, 71–89% character consistency).

Qualitative comparisons. Fig. 2 shows results on AVSpeech [5] and HDTF [19]: baselines exhibit gradual identity drift, while our method maintains consistent identity and superior visual quality throughout.

Ethics Statement

Our method synthesizes realistic human videos from audio and a reference image. Like other talking-head and human-animation techniques, it could be misused to produce deceptive or non-consensual content such as deepfakes. We advocate its use only with the consent of the depicted individuals and for legitimate applications, and we support forensic detection, provenance, and watermarking efforts to mitigate misuse. All data used in this work was collected and processed in accordance with the respective licenses and terms of use.

References

1. Bigata, A., Stypułkowski, M., Mira, R., Bounareli, S., Vougioukas, K., Landgraf, Z., Drobyshev, N., Zieba, M., Petridis, S., Pantic, M.: Keyface: Expressive audio-driven facial animation for long sequences via keyframe interpolation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5477–5488 (2025)
2. Chen, Y., Liang, S., Zhou, Z., Huang, Z., Ma, Y., Tang, J., Lin, Q., Zhou, Y., Lu, Q.: Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters. arXiv preprint arXiv:2505.20156 (2025)
3. Cui, J., Li, H., Zhan, Y., Shang, H., Cheng, K., Ma, Y., Mu, S., Zhou, H., Wang, J., Zhu, S.: Hallo3: Highly dynamic and realistic portrait image animation with diffusion transformer networks. arXiv e-prints pp. arXiv–2412 (2024)
4. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
5. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. arXiv preprint arXiv:1804.03619 (2018)
6. Gan, Q., Yang, R., Zhu, J., Xue, S., Hoi, S.: Omniavatar: Efficient audio-driven avatar video generation with adaptive body animation. arXiv preprint arXiv:2506.18866 (2025)
7. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
8. Hu, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8153–8163 (2024)
9. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
10. Kong, Z., Gao, F., Zhang, Y., Kang, Z., Wei, X., Cai, X., Chen, G., Luo, W.: Let them talk: Audio-driven multi-person conversational video generation. arXiv preprint arXiv:2505.22647 (2025)
11. Ma, P., Haliassos, A., Fernandez-Lopez, A., Chen, H., Petridis, S., Pantic, M.: Auto-avsr: Audio-visual speech recognition with automatic labels. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)

12. Peebles, W.S., Xie, S.: Scalable diffusion models with transformers. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 4172–4182 (2022), https://api.semanticscholar.org/CorpusID:254854389
13. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM international conference on multimedia. pp. 484–492 (2020)
14. Tu, S., Pan, Y., Huang, Y., Han, X., Xing, Z., Dai, Q., Luo, C., Wu, Z., Jiang, Y.G.: Stableavatar: Infinite-length audio-driven avatar video generation. arXiv preprint arXiv:2508.08248 (2025)
15. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
16. Xu, M., Li, H., Su, Q., Shang, H., Zhang, L., Liu, C., Wang, J., Yao, Y., Zhu, S.: Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. arXiv preprint arXiv:2406.08801 (2024)
17. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022)
18. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8652–8661 (2023)
19. Zhang, Z., Li, L., Ding, Y., Fan, C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3661–3670 (2021)