Supplementary Material for `SyncDreamer`: Controllable and Expressive Avatar Generation Beyond the Talking Head

Fatemeh Nazarieh[1] Zhenhua Feng[2*] Diptesh Kanojia[1] Josef Kittler[3,4]
Muhammad Awais[3,4]
[1]School of Computer Science and Electronic Engineering, University of Surrey, UK
[2]School of Artificial Intelligence and Computer Science, Jiangnan University, China
[3]Centre for Vision, Speech and Signal Processing, University of Surrey, UK
[4]Institute for People-Centered AI, University of Surrey, UK

1. Supplementary Materials

This document provides additional implementation details, dataset descriptions, evaluation metrics, and qualitative results for **SyncDreamer**. We also include ablation studies, visualization examples, and extended comparisons supporting the claims in the main paper.

2. Datasets and Evaluation Metrics

We construct a large-scale, high-quality dataset by aggregating publicly available sources that capture diverse motion styles, identities, and camera viewpoints. The collected data is categorized into three framing types, full-body, upper-body, and facial close-ups, covering a wide range of expressive speaking and performance scenarios.

- **Full-body:** Derived from **HumanVID** [17], containing approximately 19K YouTube videos. Although the dataset lacks synchronized audio, its rich visual diversity facilitates learning co-speech body dynamics and global motion patterns.
- **Upper-body:** Collected from **Hallo3** [4], which provides around 10K high-quality audio–video clips featuring expressive upper-body gestures and head movements.
- **Facial close-ups:** Sourced from **HDTF** [20] and **AVSpeech** [6], offering in-the-wild talking-face recordings with clear lip motion and fine-grained identity details.

To ensure quality and consistency, we perform audio–visual synchronization using SyncNet [2] and apply face quality filtering via InsightFace [5]. After filtering, we retain approximately 650K clips (each 5–30 seconds), standardized to a fixed spatial resolution and trimmed to single-speaker segments. The resulting dataset enables training for expressive, identity-consistent avatar generation across varied conditions, including emotional speech, singing, and performance-style motion.

We evaluate model performance using both perceptual and identity-based measures. Fréchet Inception Distance (FID) [9] and Fréchet Video Distance (FVD) [13] assess image realism and temporal coherence, respectively. SyncNet Confidence (Sync-C) and SyncNet Distance (Sync-D) [2] quantify audio–lip synchronization by measuring correspondence between speech and mouth motion. In addition, Peak Signal-to-Noise Ratio (PSNR) [10], Structural Similarity Index (SSIM) [16], and Cosine Similarity (CSIM) [8] evaluate pixel-level and perceptual similarity between generated and reference frames, reflecting visual fidelity and identity consistency. Finally, Image Quality Assessment (IQA) [18], computed from reference-based perceptual similarity, evaluates visual fidelity and identity preservation. Together, these metrics provide a comprehensive assessment of perceptual realism, temporal consistency, and audio–motion alignment in generated sequences.

3. Implementation Details

The attribute extraction module is implemented using the Qwen2-VL-7B-Instruct model [15], publicly released under the checkpoint Qwen/Qwen2-VL-7B-Instruct. It provides structured semantic cues such as attire, background layout, surrounding objects, body posture, and dynamic expressions like "tilting the head" or "holding a pen.".

For description generation, we use the Gemini 2.5 Pro model [3]. The model receives a single extracted video frame and outputs a concise (fewer than 77 tokens) description capturing salient visual cues such as attire, background, and surrounding context. These captions serve as text–image pairs for training the prompt enhancement module described in the main paper. To obtain image descriptions, we query the model with the following instruction: *"You are given a reference image. Describe the image*

*Corresponding Author: Zhenhua Feng

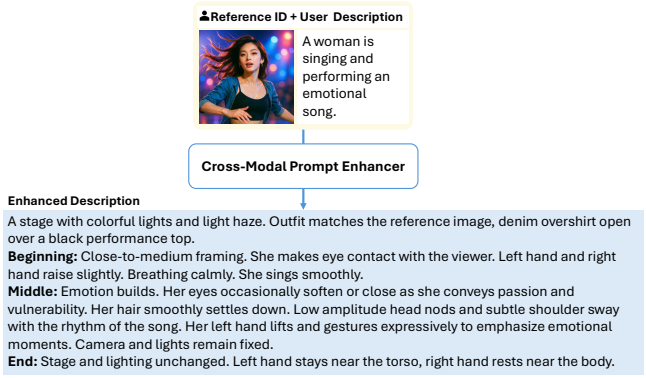

Figure 1. **User instruction enhancement.** The cross-modal parompt enhancer fuses visual cues from the reference image with the user's instruction to produce a motion-aware description.

clearly and concisely in fewer than 77 tokens. Do not include speculation, hidden details, or information not visible in the image. Only describe what can be directly observed."

4. Enhanced Prompt Evaluation Criteria

Generic reward functions are often inadequate for optimizing text prompts in talking-avatar generation, as they fail to account for the connections between visual grounding, temporal structure, and expressive semantics. To address this, we design a domain-specific composite reward function for GRPO training that evaluates enhanced prompts across five complementary dimensions: **Image Faithfulness**, **Instruction Consistency**, **Temporal Structure**, **Pose and Gesture Cues**, and **Conciseness**. Each component captures a distinct aspect of multimodal alignment, guiding the model toward generating prompts that are visually grounded, temporally coherent, and semantically consistent with user intent.

The **overall reward function** is defined as follows: for each candidate enhanced prompt y_i, generated from a reference image I and an input instruction x, the total reward is computed as:

$$\begin{aligned} R(y_i) = & \, w_{\text{img}} R_{\text{img}}(I, y_i) + w_{\text{inst}} R_{\text{inst}}(x, y_i) \\ & + w_{\text{time}} R_{\text{time}}(y_i) + w_{\text{pose}} R_{\text{pose}}(y_i) \\ & + w_{\text{len}} R_{\text{len}}(y_i), \end{aligned} \tag{1}$$

where each term $R_* \in [0, 1]$ represents a normalized score for one dimension, and the weights $\{w_*\}$ are empirically tuned (satisfying $\sum w_* = 1$) to balance visual fidelity, semantic alignment, and brevity. In the following sections, we explain each metric separately.

Image Faithfulness (R_{img}) To evaluate how accurately the enhanced prompt represents the visual content of the reference image, we employ a frozen multimodal LLM (Gemini [3]) as an image–text alignment evaluator. The model outputs a scalar score in $[0, 1]$ according to an evaluation scheme with four weighted criteria:

Figure 2. **Qualitative comparison with prior methods.** **`SyncDreamer`** generates facial outputs with consistent identity, smooth motion transitions, and precise audio–lip synchronization.

- **Identity and Appearance:** Consistency in hairstyle, clothing, and visible accessories.
- **Pose and Body Configuration:** Agreement on posture, head orientation, and hand visibility.
- **Scene and Background:** Alignment of environmental context, such as desk, lighting, and nearby objects.
- **Motion Plausibility:** Whether described actions are physically consistent with the image.

This score captures high-level semantic alignment between the image and the prompt. To complement this global assessment, we additionally employ an attribute-level match score to extract structured visual attributes (e.g., posture, gaze, accessories) from both the image and the enhanced prompt. The attribute-level match score is computed as:

$$R_{\text{attr}}(y_i) = \frac{\text{\# matched attributes}}{\text{\# known attributes in image}}, \tag{2}$$

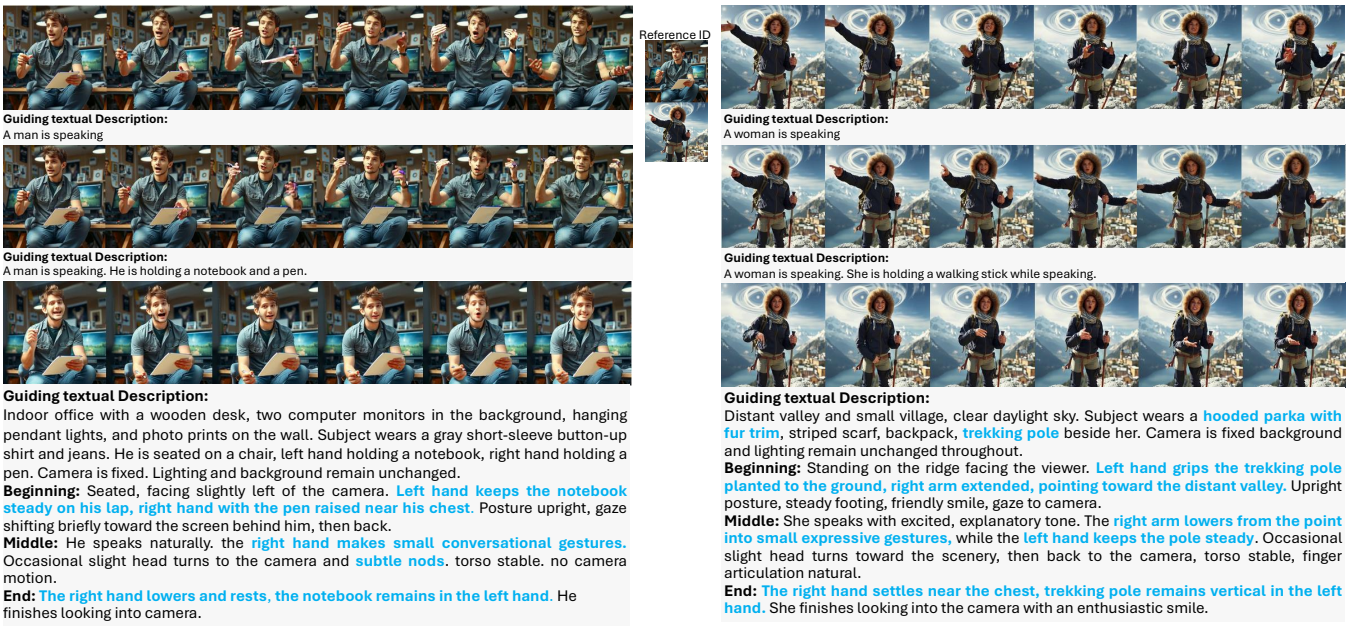

Figure 3. **Impact of the Cross-Modal Prompt Enhancer on generation quality.** We compare avatars generated from (top to bottom): (1) a minimal user prompt, (2) a manually refined prompt, and (3) an enhanced prompt produced by our image-grounded Prompt Enhancer. Minimal prompts result in generic motion and object disappearance (e.g., missing notebook or pen), while enhanced prompts lead to improved gesture naturalness, object continuity, and identity-scene coherence across the video.

and for list-type attributes, overlap is measured by Jaccard similarity:

$$R_{\text{Jaccard}}(A, B) = \frac{|A \cap B|}{|A \cup B|}, \tag{3}$$

where A and B are sets of attribute values from the image and prompt.

$$R_{\text{hybrid}} = \lambda R_{\text{Jaccard}} + (1 - \lambda) R_{\text{attr}}, \quad \lambda = 0.3 \tag{4}$$

Finally, the overall image faithfulness metric combines Gemini's global alignment with attribute-level grounding:

$$R_{\text{img}} = \beta R_{\text{Gemini}} + (1 - \beta) R_{\text{hybrid}}, \quad \beta = 0.6 \tag{5}$$

This formulation balances semantic coherence and fine-grained visual consistency, ensuring enhanced prompts remain faithful to the subject's identity and context.

Instruction Consistency (R_{inst}) This component evaluates how faithfully the enhanced prompt preserves the semantic intent of the user's original instruction. We employ the Gemini model [3] to compare the two texts and produce a normalized scalar score in the range $[0, 1]$, where higher values indicate stronger semantic alignment and lower values reflect omissions or unintended alterations.

Temporal Structure (R_{time}) This component assesses whether the enhanced prompt describes motion with a clear temporal progression (e.g., *beginning* $\rightarrow$ *middle* $\rightarrow$ *end*). We implement two complementary versions:

- A lightweight rule-based evaluator that assigns discrete values $R_{\text{time}} \in 0.0, 0.5, 1.0$ based on the presence of explicit temporal cues.
- An LLM-based evaluator that judges temporal coherence on a continuous scale within $[0, 1]$.

Pose and Gesture Cue Score (R_{pose}) This component evaluates whether the enhanced prompt includes fine-grained, human-specific cues, such as head orientation, gaze direction, and hand gestures, that contribute to expressive motion. We compare these cues with pose features extracted from the reference image to compute a normalized alignment score in the range $[0, 1]$.

Conciseness Penalty (R_{len}) To encourage concise prompts compatible with the text encoder, we apply an exponential decay penalty based on the prompt length $T(y_i)$ relative to a token budget B, no more than 220 tokens:

$$R_{\text{len}}(y_i) = \exp\left(-\alpha \cdot \max(0, T(y_i) - B)\right), \tag{6}$$

where $\alpha \in [2, 5]$. Prompts within the budget receive full score ($R_{\text{len}} = 1$), while overly long ones are softly penalized.

4.1. User Instruction Enhancement Example

Figure 1 shows an example of how a user's initial prompt is refined by our Cross-Modal Prompt Enhancer into a more structured and visually grounded description. This enhanced description incorporates visual cues and temporal details that are typically missing from generic inputs.

5. Additional Experimental Results

5.1. Additional Qualitative Comparisons

Figure 2 shows qualitative comparisons with existing audio-driven avatar models. Compared to prior diffusion-based methods [1, 4, 7, 14], **`SyncDreamer`** produces temporally coherent results with clearer facial details and consistent gaze from a given speech. The integration of the Visual Adapter, Attention Localization Loss, and Audio Dynamics Encoder enables smoother motion transitions, accurate details, and aligned audio–lip synchronization, resulting in consistently realistic and emotionally aligned avatar behavior. Notably, *V-Express*, *Hallo*, *Hall2*, and *EchoMimic* are based on standard diffusion models, while *Hallo3*, *FantasyTalk*, *Hunyuan Video Avatar*, and *OmniAvatar* adopt Transformer–based diffusion architectures for improved expressiveness.

5.2. Effect of the Visual Adapter on Identity Preservation

We provide additional qualitative comparisons to illustrate the role of the visual adapter in preserving identity. Figure 4 compares results generated with and without the visual adapter. Without the visual adapter, the model relies only on latent visual representations extracted using a *3D VAE* from the reference image. In this case, identity-related features are not explicitly enforced during generation, which leads to noticeable identity drift and inconsistencies in facial attributes across frames. By contrast, the proposed visual adapter injects identity-aware features through attention-based conditioning, enabling the model to maintain consistent facial characteristics and fine-grained identity cues throughout the generated sequence.

5.3. Effect of Attention Localization Loss.

Figure 5 illustrates the extended qualitative comparison for the Attention Localization Loss ablation. Without this loss, attention becomes spatially diffused, leading to identity drift and misplaced features during body motion. By constraining cross-attention to semantically relevant regions (e.g., eyes, mouth, hands), our loss enforces spatial consistency and preserves fine-grained details across frames. To further validate the contribution of this component, we additionally report quantitative ablation results in Table 1. The results show consistent degradation in identity preservation

Figure 4. Qualitative comparison illustrating the effect of the visual adapter. Without the visual adapter, facial identity drifts and identity-related details become inconsistent, whereas our method maintains stable identity characteristics.

and audio–visual synchronization when the loss is not applied, demonstrating its importance for stable identity control and improved audio–visual alignment.

Method	DINO% ↑	Face.sim% ↑	SyncNet ↑
w/o Attention Localization Loss	73.8	69.3	7.91
w Attention Localization Loss	**74.2**	**71.5**	**8.26**

Table 1. Quantitative ablation of the Attention Localization Loss.

5.4. Effect of the Cross-Modal Prompt Enhancer.

Figure 3 presents an extended comparison of motion generation under three input conditions: minimal, manually refined, and enhanced prompts. Minimal prompts produce generic motion with missing contextual details, while manual refinements offer only partial improvements. In contrast, prompts enhanced by our Cross-Modal Prompt Enhancer, which integrates visual cues from the reference image, yield context-aware gestures, stable object handling, and improved identity–scene coherence across time. These results highlight the importance of multimodal grounding for achieving semantically rich and visually consistent avatar motion.

5.5. Additional Results on Text-Guided Motion Control

Prior works often rely on pose trajectories or 3D landmarks to drive body motion [11, 12, 19], which restricts flexibility and generalization. Other audio-driven methods [1, 4, 7, 14] incorporate text as an auxiliary condition, yet our analysis (Figure 6) reveals that it functions primarily for scene reconstruction rather than actual motion control. When provided with identical enhanced prompts from our *Cross-Modal Prompt Enhancer*, these baselines fail to align gestures or gaze with textual semantics, often producing static or inconsistent body movements and disrupted object continuity. In contrast, **`SyncDreamer`** eliminates reliance on pose trajectories by using language as an active motion control signal, generating smooth, semantically grounded gestures and gaze shifts synchronized with both

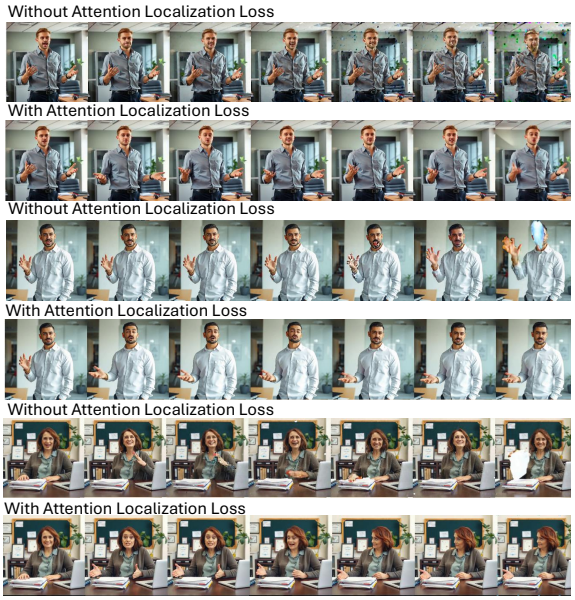

Figure 5. **Impact of the Attention Localization Loss.** The attention localization loss guides the diffusion model to focus on key facial regions relevant to speech dynamics. Without this loss, the model produces blurred or spatially inconsistent features, while including it enhances visual sharpness, identity consistency, and temporal coherence across frames.

the speech rhythm and described behavior. This demonstrates the model's ability to achieve flexible, text-driven control beyond portrait-based settings.

5.6. Ablation on the Audio Dynamics Encoder

To further evaluate the contribution of the proposed *Audio Dynamics Encoder*, we conduct detailed ablation studies on both upper-body and full-body settings under expressive audio conditions. Figures 8 and 7 provide qualitative visualizations comparing: (1) recent diffusion-based methods (Hallo3 [4], OmniAvatar [7], HunyuanVideo, Avatar [1]) (2) a variant without audio dynamics and (3) our full model. When the audio dynamics encoder is removed, avatars display flattened facial expressions, minimal rhythmic motion, and limited head–body coordination. In contrast, our full model generates synchronized and emotionally rich behaviors, effectively capturing subtle audio-driven transitions such as beat-aligned head motion or tempo-consistent body gestures. This effect is particularly noticeable in full-body scenarios, where rhythmic cues drive expressive hand and torso movement.

6. Applications and Limitations

`SyncDreamer` enables controllable avatar generation from audio, text, and a single reference image, supporting a wide range of applications in virtual assistants, education, digital performance, and creative media production. Its ability to synthesize full-body motion and interpret natural language prompts allows for expressive, flexible, and user-guided animation across both portrait and full-body scenarios. However, the system remains sensitive to input quality and prompt specificity. Ambiguous or underspecified textual instructions may lead to less precise motion synthesis, while noisy audio or overlapping speech can degrade synchronization and expressiveness.

References

[1] Yi Chen, Sen Liang, Zixiang Zhou, Ziyao Huang, Yifeng Ma, Junshu Tang, Qin Lin, Yuan Zhou, and Qinglin Lu. Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters, 2025. 4, 5

[2] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016. 1

[3] Gheorghe Comanici, Eric Bieber, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. 1, 2, 3

[4] Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21086–21095, 2025. 1, 4, 5

[5] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 1

[6] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Trans. Graph.*, 2018. 1

[7] Qijun Gan, Ruizi Yang, Jianke Zhu, Shaofei Xue, and Steven Hoi. Omniavatar: Efficient audio-driven avatar video generation with adaptive body animation, 2025. 4, 5

[8] Safouane El Ghazouali, Umberto Michelucci, Yassin El Hillali, and Hichem Nouira. Csim: A copula-based similarity index sensitive to local changes for image quality assessment, 2024. 1

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017. 1

[10] Alain Horé and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369, 2010. 1

[11] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation, 2024. 4

[12] Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. Echomimicv2: Towards striking, simplified, and semi-body

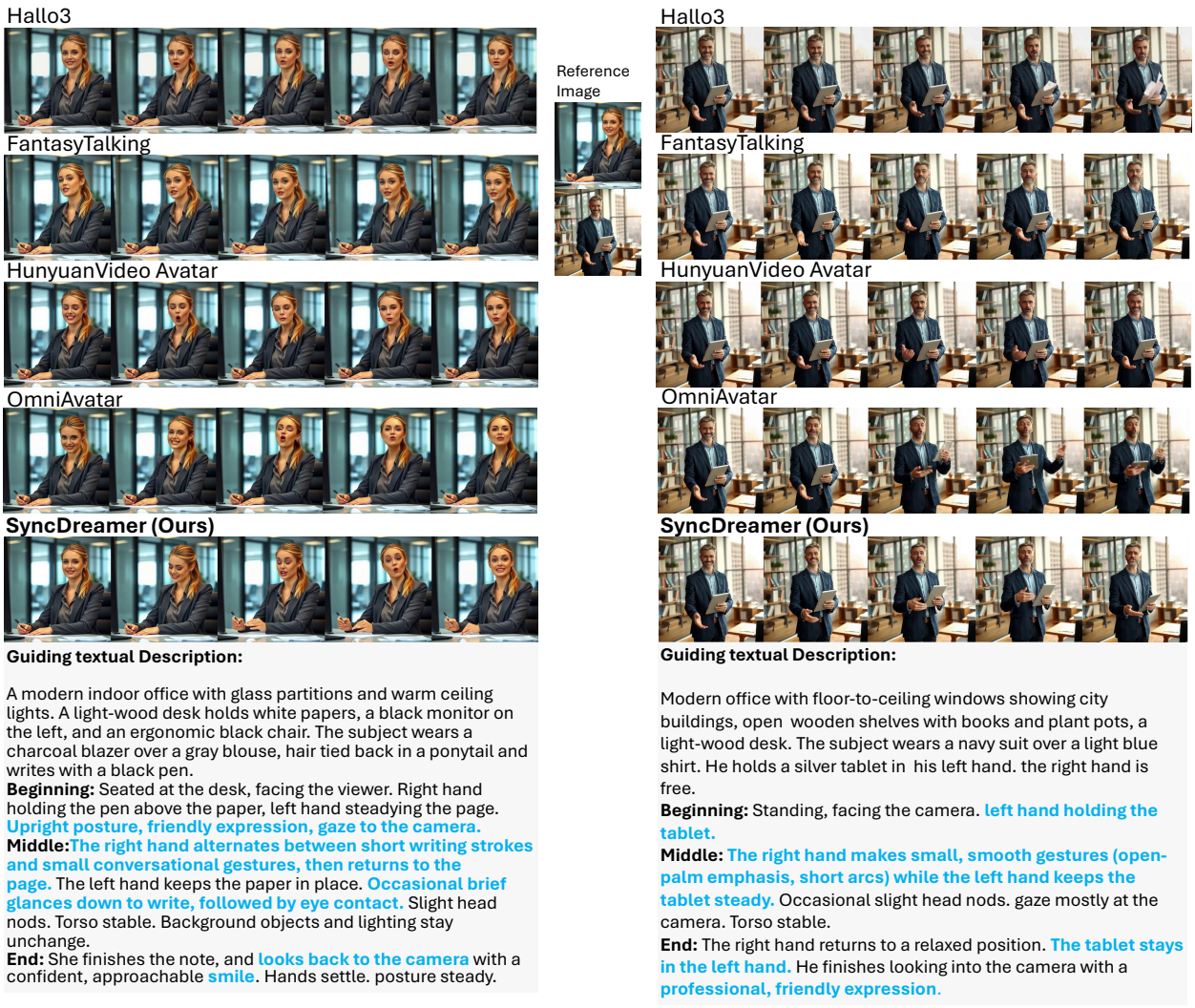

Figure 6. **Detailed comparison of text-guided upper-body motion generation.** Baseline diffusion-based models fail to maintain gesture–text alignment and object continuity, producing abrupt or inconsistent movements. In contrast, **`SyncDreamer`** maintains semantic coherence and smooth transitions, accurately reflecting the described actions and rhythmic speech cues.

human animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5489–5498, 2025. 4

[13] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges, 2019. 1

[14] Mengchao Wang, Qiang Wang, Fan Jiang, Yaqi Fan, Yunpeng Zhang, Yonggang Qi, Kun Zhao, and Mu Xu. Fantasytalking: Realistic talking portrait generation via coherent motion synthesis, 2025. 4

[15] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. 1

[16] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, pages 600–612, 2004. 1

[17] Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, and Dahua Lin. Humanvid: Demystifying training data for camera-controllable human image animation. In *NeurIPS*, 2024. 1

[18] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching lmms for visual

Figure 7. **Ablation on the Audio Dynamics Encoder (full-body, Singing Avatar).**Generating realistic full-body avatars with synchronized motion is highly challenging. Compared with Hallo3, HunyuanVideo Avatar and OmniAvatar, our model produces expressive, rhythm-aware full-body animation that aligns naturally with the musical beat. Removing the Audio Dynamics Encoder leads to less expressive and unsynchronized motion.

scoring via discrete text-defined levels, 2023. 1

[19] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. In *International Conference on Machine Learning*, 2025. 4

[20] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 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*

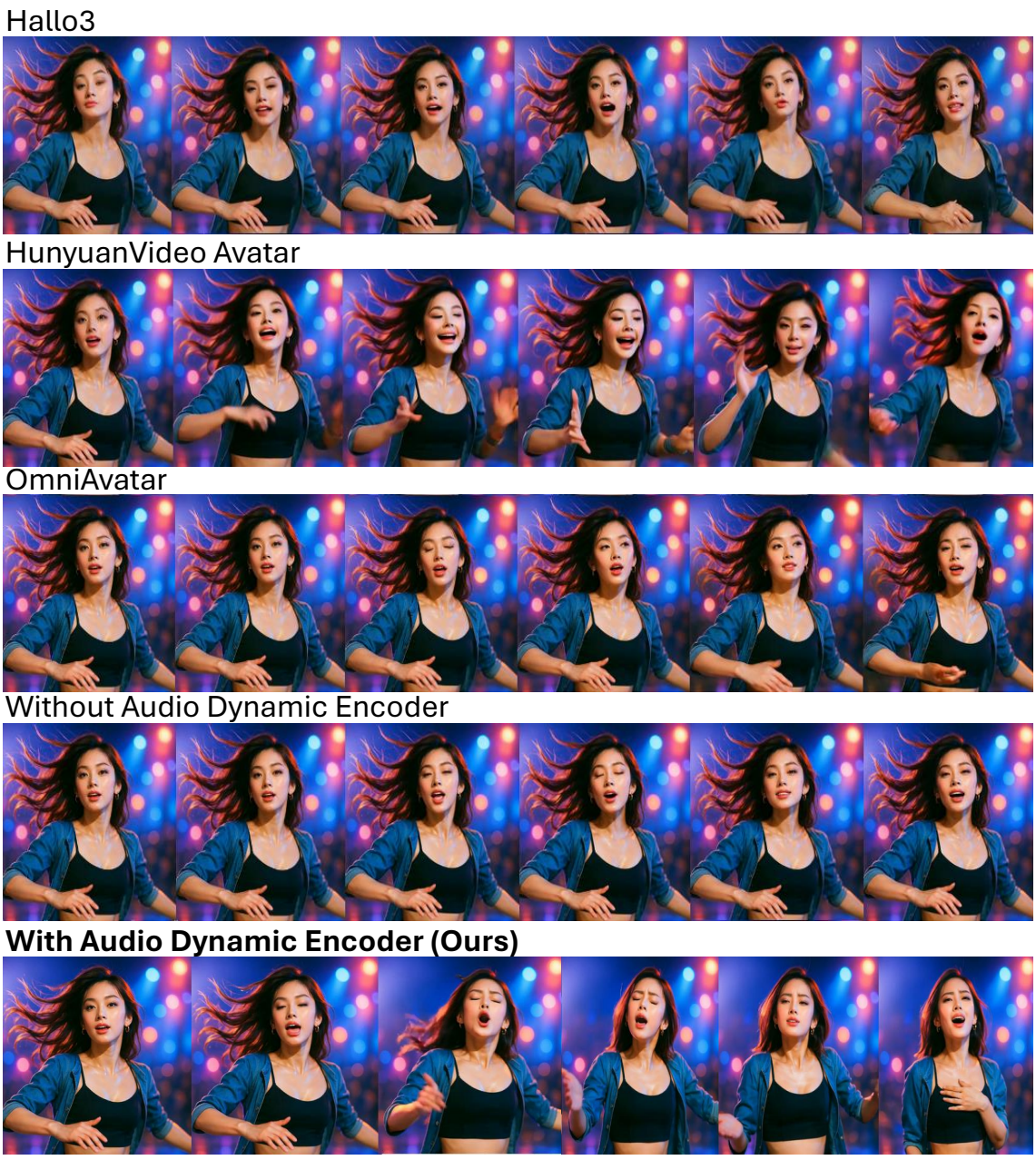

Figure 8. **Ablation on the Audio Dynamics Encoder (Upper-body, Singing Avatar).** Removing the audio dynamics encoder results in limited expressivity and weak temporal rhythm, while the full model generates emotionally rich, rhythm-synchronized motion aligned with the song's energy and beat.

Recognition, pages 3661–3670, 2021. 1