

Appendix of Paper:

“Physically Grounded Avatar Generation”

This appendix presents essential elements to enhance understanding of our PhysAvatar: the data curation process detailed in Appendix A, the physical state token extraction and projection in Appendix B, additional analysis related to MLLM in Appendix C, detailed information about the visual Turing test scoring standard in Appendix D, more qualitative results presented in Appendix E, and a discussion of the proposed PhysAvatar in Appendix F.

A DATASET CURATION

Our dataset builds on a curated subset of AVSpeech (Ephrat et al., 2018) and is further enriched with high-quality, self-collected videos covering diverse scenarios such as speeches, interviews, and news reports. To guarantee consistency and reliability for downstream training, we design a rigorous multi-stage filtering pipeline to exclude low-quality or unsuitable samples.

Specifically, we remove clips with abrupt scene transitions using PySceneDetect (Castellano, 2025), as discontinuities disrupt video temporal coherence. Speaker identity is verified with Insight-Face (DeepInsight, 2025), and only single-speaker clips are retained to avoid complications from speaker alternation. To ensure precise audio–visual alignment, we further discard clips with poor audio–lip synchronization using SyncNet (Chung & Zisserman, 2016). The filtered clips are then standardized into uniform segments: single-person, upper-body shots that focus on the most informative regions—facial expressions, lip movements, and upper-body gestures. Each video is cropped and resized to 512×512 resolution, 25 fps, and a duration of 3–15 seconds.

The dataset consists of approximately 200 hours of meticulously curated high-quality audio–visual content. Its scale, diversity, and standardization make it a robust benchmark for advancing audio-driven avatar video generation. To accelerate training, we employ an offline feature extraction pipeline. Whisper is used to derive audio tokens that capture speech content and prosody, while X-Pose provides detailed physical state tokens spanning the body, face, and hands.

B PHYSICAL STATE TOKENS EXTRACTION AND PROJECTION

Physical state tokens extraction. To integrate human behavioral dynamics into PhysAvatar, we employ the SOTA pose estimator X-Pose (Yang et al., 2024) to extract essential physical state information, including body, facial, and hand movements. As illustrated in Figure S1, X-Pose is an end-to-end multimodal pose estimation framework that can accurately detect **any** keypoints in complex real-world scenarios. It accepts an image or video clip as input and processes it through an encoder and an enhancer to improve feature extraction by leveraging information from various modalities. Subsequently, it employs two different levels of decoders, *i.e.* the **object-level** and **keypoint-level** decoders, to generate the final keypoints. The body, face, and hand keypoints are represented by the numbers 17, 68, and 21, respectively, with a detailed description in Figure S2.

In this context, we define physical state tokens using the output tokens from the object-level decoder. Specifically, we select the top- k tokens after the object-level decoding process: the top-1 token for both body and face, along with the top-2 tokens for each hand. Given that each token is 256-dimensional, these selections constitute the final 1024-dimensional physical state tokens for each video frame. Notably, we can input predicted physical state tokens along with their keypoint descriptions into the keypoint-level decoder during the denoising process to achieve visualization.

Physical state token projection. In this work, we introduce a physical state projector that utilizes a lightweight MLP architecture designed for translating each video latent into four corresponding physical state tokens while maintaining adherence to the temporal compression ratio established by the VAE. For example, in the $f = 21$ -latent configuration of the VACE framework, consistent with other models in the Wan series, the first latent token represents the initial frame. The subsequent latents are derived by compressing four consecutive frames from a total of $F = 1 + 4 \times 20$ video frames, resulting in $F = 81$. To facilitate the discrete loss calculation, we replicate the physical state token from the first frame four times, resulting in a total of 84 *reference physical state tokens*. During

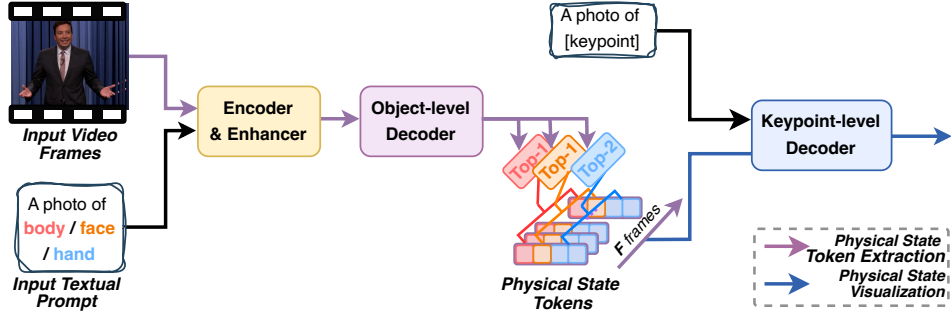


Figure S1: Illustration of pose token extraction and visualization.

Body (person) : [nose, left eye, right eye, left ear, right ear, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle, right ankle]

Face : [right cheekbone 1, right cheekbone 2, right cheek 1, right cheek 2, right cheek 3, right cheek 4, right cheek 5, right chin, chin center, left chin, left cheek 5, left cheek 4, left cheek 3, left cheek 2, left cheek 1, left cheekbone 2, left cheekbone 1, right eyebrow 1, right eyebrow 2, right eyebrow 3, right eyebrow 4, right eyebrow 5, left eyebrow 1, left eyebrow 2, left eyebrow 3, left eyebrow 4, left eyebrow 5, nasal bridge 1, nasal bridge 2, nasal bridge 3, nasal bridge 4, right nasal wing 1, right nasal wing 2, nasal wing center, left nasal wing 1, left nasal wing 2, right eye eye corner 1, right eye upper eyelid 1, right eye upper eyelid 2, right eye eye corner 2, right eye lower eyelid 2, right eye lower eyelid 1, left eye eye corner 1, left eye upper eyelid 1, left eye upper eyelid 2, left eye eye corner 2, left eye lower eyelid 2, left eye lower eyelid 1, right mouth corner, upper lip outer edge 1, upper lip outer edge 2, upper lip outer edge 3, upper lip outer edge 4, upper lip outer edge 5, left mouth corner, lower lip outer edge 5, lower lip outer edge 4, lower lip outer edge 3, lower lip outer edge 2, lower lip outer edge 1, upper lip inter edge 1, upper lip inter edge 2, upper lip inter edge 3, upper lip inter edge 4, upper lip inter edge 5, lower lip inter edge 3, lower lip inter edge 2, lower lip inter edge 1]

Hand : [wrist, thumb root, thumb's third knuckle, thumb's second knuckle, thumb's first knuckle, forefinger's root, forefinger's third knuckle, forefinger's second knuckle, forefinger's first knuckle, middle finger's root, middle finger's third knuckle, middle finger's second knuckle, middle finger's first knuckle, ring finger's root, ring finger's third knuckle, ring finger's second knuckle, ring finger's first knuckle, pinky finger's root, pinky finger's third knuckle, pinky finger's second knuckle, pinky finger's first knuckle]

Figure S2: Detailed descriptions of keypoints for each body part.

the discrete diffusion process involving physical state supervision, we employ a noise-based masked modulation strategy to randomly mask 21 audio tokens. Following this, we employ the physical state projector to predict the corresponding 84 physical state tokens. Together, the reference and predicted tokens form the basis for robust supervision in our discrete loss calculations, enabled by generated audio masks that specifically target masked positions.

C MORE MLLM ANALYSIS

MLLM input prompt. Figure S3 illustrates the specific input prompt designed for the Qwen2.5-Omni Thinker. This MLLM input prompt performs two primary functions: (i) analyzing the audio input and the corresponding human image, denoted as [AUDIO] and [IMAGE] contents, respectively; and (ii) facilitating the planning of future avatar videos, represented as [VIDEO] content, while adhering to a limited word count constraint.

You are a creative director. **Analyze** the audio and reference image **together**. Your output must consist entirely of concise phrases and follow this fixed format:

- **[AUDIO]**: [Analyze **the speaker's vocal tone, emotional nuances, pacing, rhythm, intonation, and volume**. Limit to 3-5 concise phrases without questions.]
- **[IMAGE]**: [Describe **the person's key appearance, facial and pose expression, outfit, and environmental setting**. Limit to 3-5 concise phrases without questions.]
- **[VIDEO]**: [Focus on **how audio influences future human changes**: posture shifts, gaze alterations, and expression dynamics. **Indicate emotional states. Limit to 3-5 concise phrases without questions.**]
- **Constraints**: All content combined must not exceed 80 words. No filler words, conjunctions, or articles are allowed.

Figure S3: Input MLLM prompt.

Text Editing Capability with MLLM. Figure S4 presents examples of emotion-related text editing, indicating that the inherent text-conditioned editing capabilities are preserved even when the original T5 model is replaced with our MLLM-based guider. Specifically, to achieve precise text editing, we can simply append *the desired modifications as descriptions* at the end of the MLLM input prompt. Additionally, the generated video examples reveal that our PhysAvatar effectively produces realistic gestures and plausible facial expressions in text-edited scenarios, highlighting its versatility.



Figure S4: Emotion-related text editing results.

D VISUAL TURING TEST SCORING STANDARD

To evaluate our PhysAvatar more comprehensively beyond conventional quantitative evaluation, we conducted blinded visual Turing tests with 15 participants. They assessed 30 randomly selected avatar videos from three distinct test sets, rating each video on five key dimensions:

- **Gesture Plausibility:** Assesses the semantic compatibility of gestures with the accompanying audio. 5 – Gestures align naturally and convincingly with the audio. 4 – Gestures generally reflect the audio’s intent, with minor inconsistencies; 3 – Gestures show partial or ambiguous relevance to the audio; 2 – Gestures appear unrelated or incongruent with audio; 1 – Gestures are absent, static, or clearly contradict the audio.
- **Expression Appropriateness:** Evaluates how well facial expressions convey the affective content of the audio. 5 – Expressions consistently and convincingly reflect emotional prosody; 4 – Generally appropriate with minor mismatches; 3 – Partially aligned but often ambiguous; 2 – Weak or inconsistent affective cues; 1 – Expressions absent or clearly misaligned with the audio.
- **Visual Quality:** Captures the perceived realism and rendering quality of the generated video. 5 – High perceptual realism with minimal artifacts; 4 – Visually convincing with minor imperfections; 3 – Moderate quality with visible artifacts; 2 – Noticeable degradation in texture or stability; 1 – Severe visual artifacts and low overall quality.
- **Identity Consistency:** Assesses whether the avatar’s *facial* identity remains stable throughout the video. 5 – Identity is consistently preserved across all frames; 4 – Mostly stable with minor temporal variations; 3 – Occasional drift but generally recognizable; 2 – Frequent identity instability; 1 – Significant identity loss or distortion.
- **Lip Synchronization:** Assesses how accurately lip movements correspond to the audio. 5 – Precise phoneme-level alignment with natural articulation; 4 – Generally well-synchronized with minor deviations; 3 – Adequate synchronization but with noticeable timing issues; 2 – Frequent desynchronization that disrupts perception; 1 – Severe mismatch between lip motion and speech.

E MORE QUALITATIVE RESULTS

Figure S5 presents more qualitative results that demonstrate the effectiveness of our PhysAvatar in generating realistic avatar videos exhibiting physically grounded human behavior. However, we acknowledge that some avatars still struggle to produce high-quality, fully articulated *hands*, which may exhibit artifacts during complex gestures, rapid movements, or occlusions due to the constraints of the pretrained model and the relatively limited number of model parameters employed.

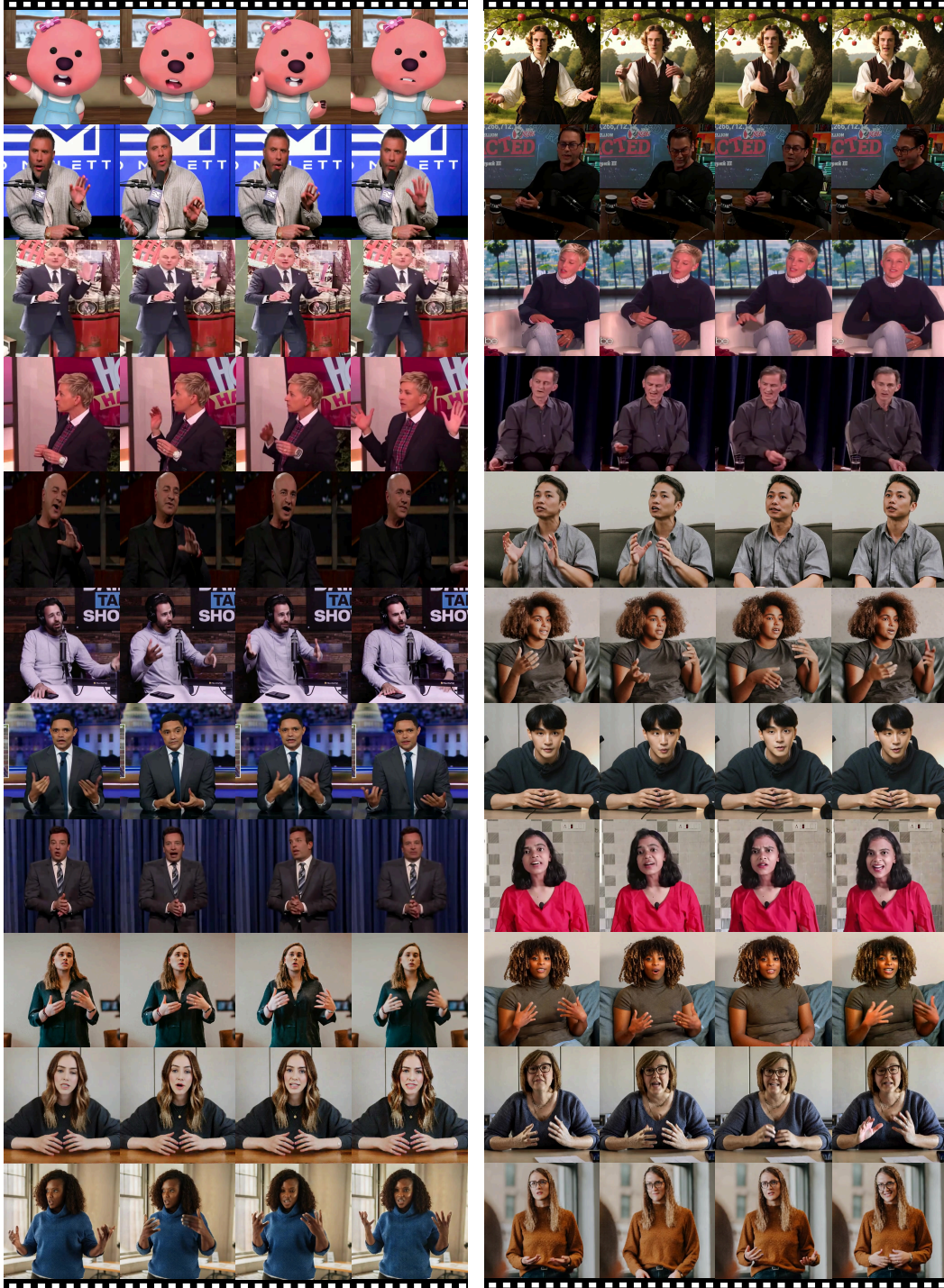


Figure S5: More qualitative results of our PhysAvatar.

F DISCUSSION

This section discusses the selection of X-Pose, the advantages and limitations of our PhysAvatar, social impact, and several aspects for future improvements.

Discussion on the selection of X-Pose. X-Pose provides a more comprehensive representation of human dynamics by jointly predicting both coarse-grained object information and fine-grained keypoints. This dual-level approach enables nuanced encoding of global object semantics and keypoint geometric structures through intermediate features that act as *physical state tokens*. In contrast, traditional methods like OpenPose (Cao et al., 2017) and DWPose (Yang et al., 2023) are limited to generating keypoint coordinates alone, which primarily focus on local geometry and fine-grained positioning, lacking the holistic understanding provided by X-Pose. Moreover, our preliminary experimental results indicate that the final keypoints produced by X-Pose significantly surpass those generated by OpenPose and DWPose.

Discussion on advantages and limitations. Our PhysAvatar framework outperforms existing SOTA baselines in both generative quality and behavioral realism, consistently producing avatars that are more physically grounded, expressive, and lifelike. However, we acknowledge a primary limitation: the quality of hand generation, which can exhibit artifacts during complex gestures, rapid movements, or occlusions due to constraints of the base VACE-1.3B model. Future iterations could benefit from utilizing larger or stronger base models to further improve the realism and fluidity of hand animations.

Discussion on social impact. Our PhysAvatar unlocks substantial commercial opportunities as virtual emotional companions, educational assistants, and live-streaming hosts, greatly enhancing user engagement and creating diverse revenue streams. However, these advancements also bring significant social implications. The use of advanced generative models raises concerns about misinformation, emotional manipulation, and ethical treatment of digital identities. As users form attachments to these digital entities, issues such as dependency and the erosion of real-world relationships become critical. Moreover, without proper regulatory frameworks, there is a heightened risk of exploitation or misuse. Therefore, it is essential to establish robust governance and ethical guidelines to ensure responsible deployment, balancing commercial benefits with the imperative to safeguard societal well-being and foster trust in AI technologies.

Discussion on future improvements. In summary, there are several key areas for improving our PhysAvatar in the future. **(i) Enhancements in Generative Performance.** Given the constraints of limited training datasets and computational resources, we utilized parameter-efficient methods, such as LoRA, for fine-tuning. However, we can leverage more advanced models, such as Wan2.2 or Wan2.2-based VACE, along with other robust base models, to achieve better performance. Additionally, expanding the dataset would allow for full model weight fine-tuning, further enhancing performance and fully taking advantage of the scalability of transformers Peebles & Xie (2023). **(ii) Adapting for Real-Time Scenarios:** Furthermore, tailoring PhysAvatar for real-time applications introduces challenges in balancing computational efficiency with visual fidelity. Future work should emphasize optimizing inference speed through methods such as Meanflow (Geng et al., 2025) or video distillation (Huang et al., 2025), as well as improving resource utilization to enhance the practicality of PhysAvatar in interactive environments.

REFERENCES FOR APPENDIX

- Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pp. 7291–7299, 2017.
- Brandon Castellano. PySceneDetect: Python and OpenCV-based scene cut/transition detection program & library, 2025. URL <https://www.scenesdetect.com>.
- Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *ACCV*, pp. 251–263, 2016.
- DeepInsight. InsightFace: State-of-the-art 2D and 3D face analysis project, 2025. URL <https://github.com/deepinsight/insightface>.

270 Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T
271 Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent
272 audio-visual model for speech separation. *arXiv:1804.03619*, 2018.
273
274 Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for
275 one-step generative modeling. In *NIPS*, 2025.
276
277 Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging
278 the train-test gap in autoregressive video diffusion. *arXiv:2506.08009*, 2025.
279
280 William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pp.
281 4195–4205, 2023.
282
283 Jie Yang, Ailing Zeng, Ruimao Zhang, and Lei Zhang. X-Pose: Detecting any keypoints. In *ECCV*,
284 pp. 249–268, 2024.
285
286 Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with
287 two-stages distillation. In *ICCV*, pp. 4210–4220, 2023.
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323