

HUMANEXPERT: UNIFIED MULTIMODAL HUMANOID GENERATION

Anonymous authors

Paper under double-blind review

This appendix presents extended results and implementation details for HumanExpert, our unified multimodal generative model for text, speech, and behavior. We include qualitative visualizations and comparative results across modalities (Sec. A), along with additional experiments analyzing training objectives, expert module configurations, and behavior modeling strategies (Sec. B). Detailed inference-time measurements are also reported (Sec. C). To support standardized evaluation, we describe the protocol used for assessing humanoid generation, including task definitions, dataset preparation, and metric computation (Sec. D). We also provide further architectural details of HumanExpert, including module specifications, integration of modality-specific experts, and key implementation choices (Sec. E).

Video. We have provided the supplemental video to illustrate our results. In this video, we show 1) multimodal generation results, 2) comparisons on talking head synthesis, 3) comparisons on whole-body synthesis, and 4) more diverse outputs across prompts and identities. We suggest watching this video for dynamic motion results.

Code is available in supplements. We provide example code files, which include the process of the training and evaluation of our HumanExpert models, as well as several example results. We can hardly upload our large model files, but all codes, data, and pre-trained models will be fully released.

A QUALITATIVE RESULTS

We visualize some qualitative results on the comparison of portrait animation (*cf.* Fig. 4), command execution (*cf.* Fig. 5). For more results, please refer to the video in the supplementary materials.

B ADDITIONAL EXPERIMENTS

To further analyze the design choices and training dynamics of HumanExpert, we conduct a series of controlled experiments. We begin by examining the impact of discrete representations or continuous representations (*cf.* Sec. B.2). Next, we investigate the integration of low-rank adaptation (LoRA) into our diffusion framework, demonstrating how modular expert design supports flexible adaptation without compromising core language capabilities (*cf.* Sec. B.3). Finally, we evaluate the scalability and design of the motion expert by varying its parameter size and connection structure (*cf.* Sec. B.4).

B.1 AUDIO-BEHAVIOR ALIGNMENT IN SELF- AND CROSS-REENACTMENT

We evaluate audio-behavior alignment under two protocols. In *self-reenactment*, for each test video the first frame serves as the reference image; the remaining frames provide the driving signals and also act as ground-truth targets, enabling a like-for-like assessment of alignment quality. In *cross-reenactment*, we pair a reference portrait with driving signals from an *in-the-wild* source video of a different identity, stressing identity preservation while transferring speech and motion patterns. Table 4 reports results for both settings. HUMANEXPERT achieves competitive alignment compared with baselines and, uniquely, also supports broader generation modes—including text-to-audio and silent-behavior synthesis—that baseline methods do not provide.

B.2 DISCRETE OR CONTINUOUS?

We investigate the impact of behavior representation by comparing discrete and continuous training objectives under the same expert architecture. Specifically, we implement two variants: (1) a VQ-VAE-based model following Jiang et al. (2023), where behavior sequences are quantized into discrete



Figure 4: Comparison on audio-conditioned portrait animation.

Table 4: Comparison of audio-behavior alignment. HumanExpert achieves competitive results while supporting broader capabilities beyond baselines, such as text-to-audio and silent behavior generation.

Method	Self-Reenactment			Cross-Reenactment			
	SSIM \uparrow	LPIPS \downarrow	FVD \downarrow	ID-SIM \uparrow	EMO-SIM \uparrow	Sync-C \uparrow	Sync-D \downarrow
JoyVasa (Cao et al. (2024))	0.619	0.204	656.9	0.832	0.209	5.826	14.33
EchoMimic (Chen et al. (2025))	0.445	0.368	776.1	0.842	0.133	4.800	13.87
Memo (Zheng et al. (2024))	0.613	0.216	505.2	0.858	0.305	6.324	14.29
Sonic (Ji et al. (2024))	0.583	0.225	490.4	0.832	0.229	6.634	13.45
HumanExpert	0.619	0.203	514.3	0.839	0.245	4.936	14.52

tokens and predicted using a cross-entropy loss, and (2) a continuous regression model that replaces diffusion with a simple L2 loss for direct behavior prediction. To evaluate these alternatives, we conduct experiments on the audio-conditioned portrait animation task, following the same evaluation protocol as in Tab. 2. Results in Tab. 5 indicate that the diffusion-based objective outperforms both the discrete (VQ-VAE) and continuous (L2) baselines, achieving better alignment and motion realism.

B.3 ABLATION ON EFFECTIVENESS OF BEHAVIOR EXPERT

We further evaluate the effectiveness of our behavior expert design by comparing it to a LoRA Hu et al. (2022) adaptation variant, where the language model remains frozen and behavior capabilities are introduced through low-rank adaptation. In addition to the language understanding results shown in

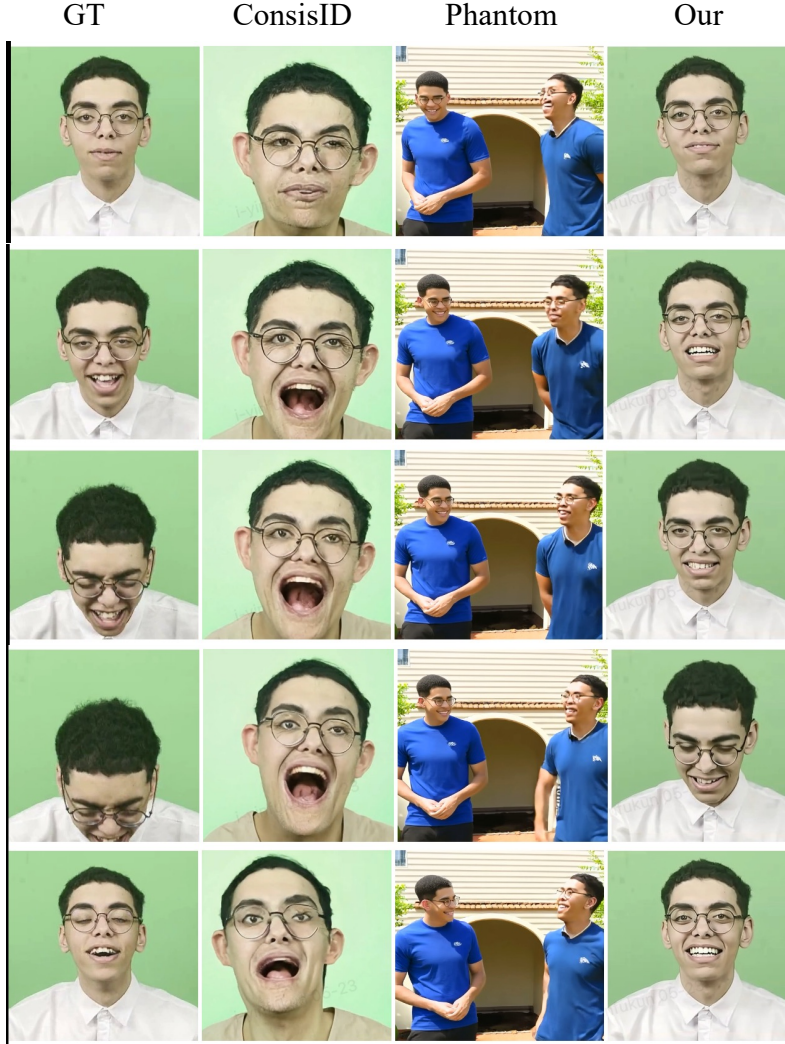


Figure 5: Comparison on command-conditioned portrait animation.

Table 5: Ablation of Training Target.

Method	SSIM \uparrow	LPIPS \downarrow	FVD \downarrow	ID-SIM \uparrow	EMO-SIM \uparrow	Sync-C \uparrow	Sync-D \downarrow
Cross-entropy	0.510	0.259	758.2	0.833	0.212	4.682	15.11
L2	0.476	0.266	611.1	0.829	0.156	4.721	16.96
HumanExpert	0.619	0.203	514.3	0.839	0.245	4.936	14.52

Table 3, we assess behavior generation performance on the command-conditioned portrait animation task, using the same evaluation protocol as in Table 1. As shown in Tab. 6, our full expert design outperforms the LoRA-based variant due to a more flexible attention and a hidden size design.

Table 6: Ablation on Effectiveness of Behavior Expert.

Method	SSIM \uparrow	LPIPS \downarrow	FVD \downarrow	ID-SIM \uparrow	EMO-SIM \uparrow
LoRA	0.434	0.565	1139.4	0.692	0.054
HumanExpert	0.538	0.306	792.6	0.723	0.137

B.4 ABLATION ON DESIGN OF BEHAVIOR EXPERT

We further examine the design of the behavior expert through two sets of ablation studies. Tab. 7 compares models with different hidden sizes. Increasing the capacity from Small to Base leads to a significant improvement in behavior prediction accuracy. In Tab. 8, we investigate how the placement of cross-modal attention affects performance. In the "Cross-Attn." setting, we control whether the behavior expert integrates cross-modal signals (from text and audio experts) at the beginning ("First Block") or end ("Last Block") of its transformer stack. We also vary the depth of cross-attention integration: either the first half or last half of layers are allowed to attend to other modalities, while the remaining layers perform modality-specific self-attention only. Results show that introducing cross-modal attention earlier in the behavior expert leads to better performance, suggesting that early fusion is more beneficial for aligning behavioral outputs with linguistic and acoustic cues.

Table 7: Ablation of Different Sizes of Behavior Expert.

Method	Hidden Size	Trainable Params.	L2 ↓
HumanExpert-Small	512	599M	0.056
HumanExpert-Base	2048	2.4B	0.021

Table 8: Ablation of Different Attention Designs of Behavior Expert.

Method	Cross-Attn.	L2 ↓	Cross-Attn.	L2 ↓
HumanExpert-Base	First Block	0.034	First Half Blocks	0.026
HumanExpert-Base	Last Block	0.035	Last Half Blocks	0.030

B.5 ABLATION ON CLASSIFIER-FREE GUIDANCE

We selected CFG hyperparameters via grid search on a held-out validation set. For the guidance scale, we ran an ablation sweep and chose 1.5, which achieved the lowest L1 error among tested settings (Table 9). For history conditioning, we tuned relative weightings with the same procedure, jointly considering accuracy (L1) and throughput (FPS).

C TRAINING AND INFERENCE EFFICIENCY

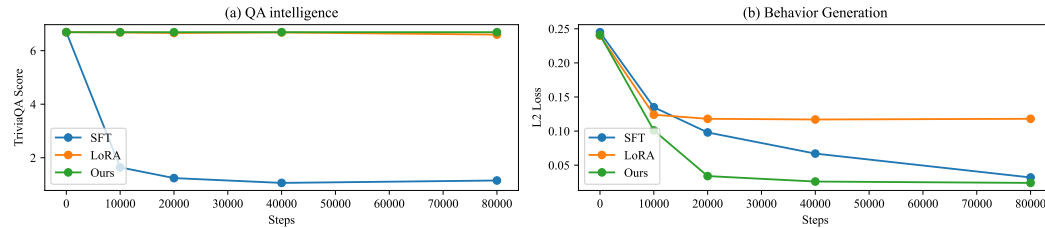


Figure 6: Ablation on training efficiency.

C.1 TRAINING EFFICIENCY

We quantify training efficiency along two axes. First, Fig. 6(a) tracks GPT-Score on WebQA over optimization steps to assess language quality. Second, Fig. 6(b) measures behavior accuracy for talking-head generation. Together, these curves indicate fast convergence and effective multimodal learning in HumanExpert.

Table 9: Ablation over CFG guidance scales on the validation set. Lower is better for L1. We select 1.5.

Guidance Scale	L1	Guidance Scale	L1
1.0	0.56	2.5	0.61
1.5	0.55	3.5	0.75

Table 10: Evaluation of inference time costs on text-driven motion generation. We evaluate the Frames Per Second (FPS) by averaging our generated frames for each second, under the same 1 Tesla A100, the smaller model size gets the faster FPS.

	HumanExpert	UNet Render	DiT Render
FPS	2.95	0.75	0.16

C.2 INFERENCE EFFICIENCY

To enable low-latency streaming, we employ an *interleaved sampling* strategy that processes text, speech, and behavior in small, interleaved chunks rather than full sequences; module-level throughput is summarized in Tab. 10. On a single A100 GPU without engineering optimizations, our diffusion-based renderer achieves ~ 3 FPS with ~ 17 s initial latency (Table 8). A faster autoregressive variant reaches ~ 13 FPS with ~ 2 s latency, but—as shown by our ablations (Table 4)—this speed comes with reduced behavior quality. Closing this performance–quality gap is a key direction for future work, where we plan to explore more efficient architectures and streamlined behavior representations to approach real-time performance without sacrificing fidelity.

D EVALUATION PROTOCOLS ON HUMANOID GENERATION

We propose a protocol to evaluate HumanExpert on a range of humanoid generation tasks, including talking head synthesis, latent command execution, and fine-grained pose or facial motion. Based on curated and cleaned multimodal datasets, we construct standardized benchmarks with consistent input-output formatting and modality alignment. Evaluation is performed using ID, emotion, and sync metrics for talking heads, as well as GPT-Score for language relevance.

D.1 DATASET STATISTICS

Our hybrid dataset comprises two components. Text–Speech–Video triplets are built by aggregating public corpora Zhang et al. (2021); Xie et al. (2022); Kirschstein et al. (2023) under their respective licenses; audio is transcribed with Whisper to obtain text supervision. Command–Video is a newly collected corpus in which professional actors execute distinct, human-centric behavioral commands. A statistical summary is presented in Table 11.

D.2 PRE-TRAINING TALKING DATASET

The pre-training talking dataset is constructed from two primary sources: (1) high-quality, custom-produced question-answering videos and (2) internet videos authorized for use, acquired from data providers. Advanced algorithms are employed to create a high-volume dataset that aligns audio, text, and human motion. During this phase, the model learns essential information to bridge modalities, with a particular emphasis on facial movements (e.g., emotional expressions, lip synchronization) and hand gestures, correlated with audio features such as tone, semantic content, and speech rate. The dataset comprises over 4 million clips, totaling approximately 16,732 hours, primarily consisting of conversational interview-based videos.

D.3 INSTRUCTION-FOLLOWING DATASET

To develop an interactive virtual human capable of answering questions and responding to user instructions, we focus on instruction-following capabilities. This objective is divided into two key

Table 11: Statistical overview of the curated dataset used for training and evaluation.

Datasets	Hours	Clips	Duration Range (s)	Avg. Words (ASR)	Subjects	Commands
Text-Speech-Video (Train)	94	35,764	1.12–19.4	14.43	—	—
Text-Speech-Video (Test)	0.3	100	2.72–26.0	14.95	—	—
Command-Video (Train)	13	5,432	1.12–105.8	—	24	38
Command-Video (Test)	0.26	100	1.96–63.36	—	5	10

aspects: (1) facial-focused instruction-following tasks, such as executing specific head movements or facial expressions, and (2) body-movement-focused tasks, involving specific body gestures. To support these goals, we curated two datasets: (1) a facial dataset comprising 163.1 hours of high-quality, custom-produced content, and (2) a body movement dataset totaling 2,597.46 hours, also derived from high-quality, custom-produced sources.

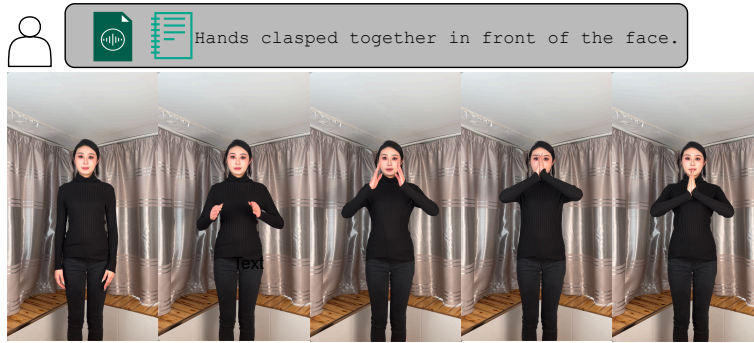


Figure 7: Example of body movement instruction-following tasks.

D.4 SPEECH DATASET

To maintain the model’s intelligence level, we prepared an auxiliary audio-only conversation dataset to fine-tune domain-specific capabilities during training. This dataset, used for Low-Rank Adaptation (LoRA) and full supervised fine-tuning (SFT), includes 13,450 hours of Chinese speech and 10,682 hours of English speech. Although not directly utilized in the primary training pipeline, this dataset enhances the model’s conversational abilities.

D.5 DATA PROCESSING PIPELINE

We developed an end-to-end data preprocessing pipeline to prepare the datasets. For audio data, the pipeline includes audio separation, automatic speech recognition (ASR), and large language model (LLM)-based question synthesis to create paired data consisting of questions in text form and answers in both audio and text formats. For visual data, the pipeline incorporates scene separation, face detection, body detection, and latent feature extraction for both face and body. Additionally, we applied lip synchronization algorithms to video segments and implemented video quality filtering to ensure high-quality outputs.

E DETAILS ON HUMANEXPERT MODELS

E.1 PIPELINE OVERVIEW

We provide the whole pipeline of our HumanExpert in Fig. 9.

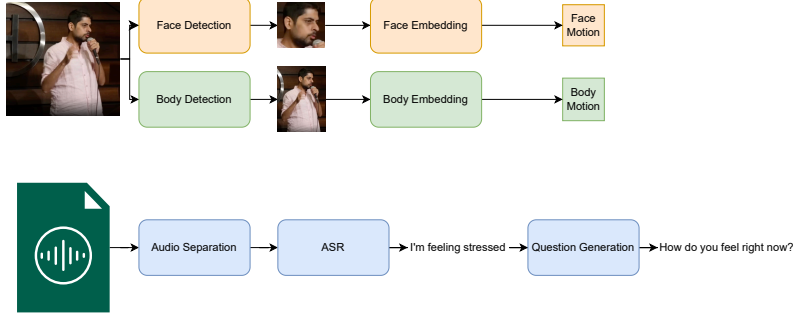


Figure 8: Overview of the data preprocessing pipeline.

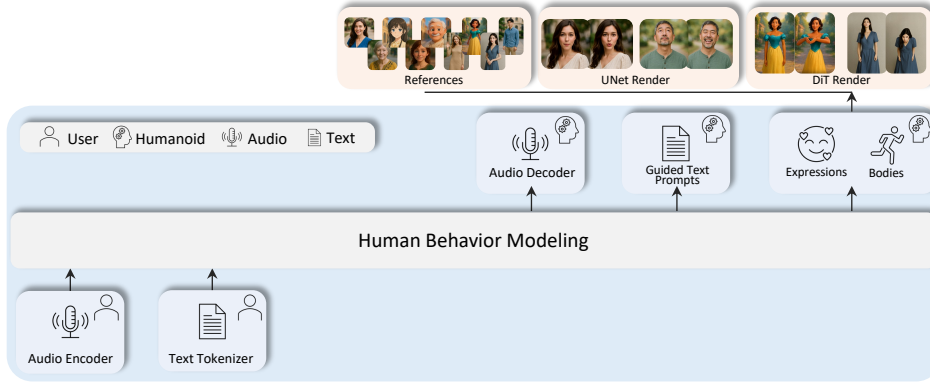


Figure 9: Overview of HumanExpert Pipeline.

E.2 RENDERER DETAILS

Our framework converts the generated *behavior* representations into photorealistic videos through task-specific rendering interfaces adapted from state-of-the-art models. For whole-body animation, we follow a pose-driven pipeline akin to UniAnimate-DiT Wang et al. (2025): HumanExpert’s normalized keypoints are first de-normalized and rasterized into a sequence of skeleton maps, which—together with a reference image—condition a diffusion-based generator to synthesize the final video. For head animation, we adopt a motion-driven renderer inspired by X-NemMo Zhao et al.: HumanExpert directly outputs a sequence of motion embeddings that, combined with a static identity reference, are injected via cross-attention into a diffusion backbone to produce the animated head.

E.3 IMPLEMENTATION DETAILS

The current effective generation horizon of our method is 20s, reflecting the training data distribution. Concretely, the behavioral sequence is trained with a maximum length of 20s (500 frames at 25fps). Although Rotary Positional Encoding (RoPE) permits extrapolation to longer sequences at inference, we observe quality degradation—most notably in lip-sync—beyond 20s, and expanding this horizon via longer-interaction datasets is a focus of future work. By contrast, the rendering stage imposes no inherent time limit: it synthesizes video with a sliding temporal window to avoid GPU memory issues, enabling arbitrarily long outputs as long as a coherent behavior sequence is provided. We provide more details of our HumanExpert in Tab. 12.

Table 12: Hyperparameters for different sizes of HumanExpert.

HumanExpert	Small	Base
Backbone	GLM-4-Voice	GLM-4-Voice
Training Batch Size	8	4
Total Params.	10.1B	11.9B
Trainable Params.	0.6B	2.4B
Behavior Dim	262/512	262/512
Behavior Hidden Size	512	2048
Behavior FFN Hidden Size	6848	6848
Behavior Chunk Size	52	52
Text Vocabulary	151328	151328
Audio Vocabulary	16384	16384
Audio Chunk Size	26	26
Hidden Size	4096	4096
FFN Hidden Size	13696	13696

REFERENCES

- Xuyang Cao, Guoxin Wang, Sheng Shi, Jun Zhao, Yang Yao, Jintao Fei, and Minyu Gao. Joyvase: portrait and animal image animation with diffusion-based audio-driven facial dynamics and head motion generation. *arXiv preprint arXiv:2411.09209*, 2024.
- Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2403–2410, 2025.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- Xiaozhong Ji, Xiaobin Hu, Zhihong Xu, Junwei Zhu, Chuming Lin, Qingdong He, Jiangning Zhang, Donghao Luo, Yi Chen, Qin Lin, et al. Sonic: Shifting focus to global audio perception in portrait animation. *arXiv preprint arXiv:2411.16331*, 2024.
- Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023.
- Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023.
- Xiang Wang, Shiwei Zhang, Longxiang Tang, Yingya Zhang, Changxin Gao, Yuehuan Wang, and Nong Sang. Unianimate-dit: Human image animation with large-scale video diffusion transformer. *arXiv preprint arXiv:2504.11289*, 2025.
- Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 657–666, 2022.
- 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 recognition*, pp. 3661–3670, 2021.
- Xiaochen Zhao, Hongyi Xu, Guoxian Song, You Xie, Chenxu Zhang, Xiu Li, Linjie Luo, Jinli Suo, and Yebin Liu. X-nemo: Expressive neural motion reenactment via disentangled latent attention. In *The Thirteenth International Conference on Learning Representations*.
- Longtao Zheng, Yifan Zhang, Hanzhong Guo, Jiachun Pan, Zhenxiong Tan, Jiahao Lu, Chuanxin Tang, Bo An, and Shuicheng Yan. Memo: Memory-guided diffusion for expressive talking video generation. *arXiv preprint arXiv:2412.04448*, 2024.