

To provide a more complete understanding of our proposed U-Mind system, this supplementary material presents (1) detailed descriptions of our data construction pipeline, including reasoning-augmented annotations and multimodal synthesis preparation; (2) expanded implementation details, covering model architecture, tokenizer training, and experimental setup; and (3) additional qualitative results that illustrate its strengths in multimodal dialogue. These materials aim to clarify key design choices and support the reproducibility of our work.

1. Data Construction Details

To support the training of U-Mind across dialogue, instruction-following, and full-stack video generation, we construct a comprehensive multimodal dataset that integrates both benchmark resources and proprietary corpora. The data pipeline is organized into two primary components: core datasets for multimodal alignment and reasoning, and auxiliary corpora for supporting text2speech (TTS) and video rendering.

1.1. Multimodal Training Datasets

Multimodal Dialogue Supervision. We build a reasoning-enhanced dialogue dataset based on BEAT v2 [13], which provides paired speech and motion data. For each utterance or gesture segment, we prompt a large language model (Qwen3) to generate three question-answer-reasoning triplets, injecting explicit CoT-style traces into the original corpus. The generated textual answers are then converted to audio using a high-quality neural TTS system (Orpheus-TTS [2]), resulting in coherent multimodal dialogue samples. For original speech recordings, we apply Whisper [15] to obtain accurate transcriptions. This pipeline yields a dataset containing 10k reasoning-enhanced S2M examples and over 30k multimodal QA-reasoning pairs, enabling symbolic and physical grounding in dialogue settings.

Instruction-Following Supervision. We construct a CoT-augmented instruction-following dataset based on HumanML3D [5], which provides natural language descriptions paired with 3D motion data. For each description, we prompt a large language model (Qwen3) to generate three question-answer-reasoning triplets, enabling hierarchical task decomposition and intermediate planning. This process yields 16k high-quality T2M samples, each enriched with explicit symbolic reasoning signals. To convert symbolic supervision into multimodal alignment targets, all Qwen3-generated responses are synthesized into speech using Orpheus-TTS [2], forming synchronized text-speech-motion triples.

1.2. Auxiliary Corpora for Speech and Video Modules

Conversational Reasoning and TTS. To further enhance symbolic planning and speech generation capabilities, we incorporate two external datasets: OpenOrca [14], a high-quality open-domain corpus rich in multi-turn conversational reasoning, and Common Voice [1], a multilingual crowd-sourced speech dataset that improves the TTS model’s prosody, phonemic diversity, and naturalness. By mixing these resources into our training pipeline, the model benefits from both abstract reasoning supervision and diverse acoustic grounding, leading to more fluent, expressive, and semantically coherent multimodal outputs.

Video Renderer Training. To enable photorealistic and temporally aligned video generation, we adopt two complementary rendering pipelines. First, we collect a 500-hour proprietary human video corpus annotated with 2D keypoints using DWPose [20], which is used to train a diffusion-based renderer based on the WAN [18] backbone. Second, we train a Gaussian Splatting renderer on the TaoAvatar [4] dataset, utilizing 3D SMPL-X annotations to enable direct pose-driven 3D video synthesis. These two renderers provide complementary views: while diffusion models produce high-fidelity 2D renderings from 2D pose inputs, Gaussian-based renderers directly generate consistent 3D avatar videos from SMPL-X poses, supporting diverse real-time generation scenarios. These components together support U-Mind’s full-stack, instruction-following multimodal interaction capabilities.

2. Additional Implementation Details

This section provides implementation details for the tokenizer modules, model configuration, training strategy, and inference settings used in the U-Mind system.

2.1. Motion Tokenizer

We employ a residual VQ-VAE architecture [16, 24] to discretize 3D motion sequences into compact token representations. Specifically, we use a 4-layer residual quantization scheme with a temporal downsampling ratio of 4. To enhance numerical stability and preserve geometric continuity, each input sequence of SMPL-X parameters is first converted into a 6D continuous pose representation before encoding. These continuous poses are then encoded into discrete codebook indices, enabling token-level modeling of motion. This design ensures a balance between fidelity and compression, facilitating efficient multimodal learning.

2.2. Speech Tokenizer

For speech representation, we utilize the SpeechTokenizer [23] module to discretize waveform audio into acoustic tokens. This tokenizer captures phonetic and prosodic infor-

Do you think higher salaries always lead to greater happiness in a job?



<think>The question inquires about the relationship between salary and happiness at work, implying a direct link between the two. To address this, I reflect on my belief that while money can make life easier, it doesn't inherently lead to happiness. I recall instances where individuals with high-paying jobs still felt unfulfilled, suggesting that intrinsic factors, such as alignment with one's passions, values, and the nature of the work itself, play a more significant role in job satisfaction. From this reasoning, I conclude that true happiness at work arises from doing meaningful work that you are passionate about, rather than solely from a high salary.</think>

🗣️: "In my opinion, a higher salary doesn't always equate to greater happiness at work. While earning more money can certainly provide comfort, it's not a guarantee of job satisfaction. If the salary is reasonable and the work offers personal fulfillment, that's ideal. However, if the salary is too low, it's hard for anyone to maintain a positive outlook on their job."

Ref Image



Can you share a memory of a special place you visited and a wish you made there?



<think>The question asks about a memorable place and the wish associated with it. I begin by recalling a location that holds personal significance, and a temple comes to mind, possibly connected to my cultural or spiritual practices. I remember specifically wishing for a job that would be easy to find during my visit there. In structuring the answer, I focus on the connection between the place, the wish, and my desire for a smooth career path, while also conveying emotions like hope and optimism for the future.</think>

🗣️: "I visited a temple and made a wish for a job that would be easy to find. I placed my hopes and fortune there, wishing for a smooth future where luck would guide me toward good opportunities.."

Ref Image



Figure 1. More Multimodal Dialogue Results.

mation suitable for both speech synthesis and alignment. During multimodal training, the speech encoder, like all modality encoders, is kept frozen to preserve stable representations and avoid catastrophic forgetting.

2.3. Model Backbone and Token Prediction

Our multimodal language model is based on the LLaMA2-7B [17] architecture, initialized with the pretrained weights from AnyGPT [21]. We extend the tokenizer's codebook to include motion and speech tokens and train the model using a unified next-token prediction objective. Given a prompt sequence, the model autoregressively generates text, acoustic, and motion tokens in a unified token stream.

2.4. Training Details

Motion Tokenizer. The motion tokenizer is trained using a 4-layer residual VQ-VAE architecture with a temporal downsampling ratio of 4. To ensure robust learning and smooth quantization, SMPL-X pose parameters are first converted into 6D continuous representations. The tokenizer is optimized using the AdamW optimizer with a learning rate of $1e-5$, β values of $[0.9, 0.999]$, and no weight decay. We apply a learning rate scheduler with milestones at epochs 50, 150, and 250, decaying the learning rate by a factor of 0.4 at each milestone. Exponential moving average (EMA) with a momentum of 0.99 is employed to stabilize training. The model is trained with a batch size of 256 on 8 H100 GPUs.

U-Mind Backbond. We adopt a two-stage training pipeline to balance perceptual grounding with instruction alignment.

In the pre-training stage, we use a learning rate of $1e-4$, a batch size of 4, a maximum sequence length of 2048, and a cosine learning rate scheduler. Training is conducted on 8 H100 GPUs using bfloat16 precision. In the instruction tuning stage, we continue training from the pre-trained checkpoint with a reduced learning rate of $2e-5$, keeping the batch size and sequence length unchanged. This staged approach allows the model to first acquire modality fluency without catastrophic forgetting, and then align with high-level interactive behaviors under a stabilized training regime.

Each input prompt is formatted with modality-aware delimiters. The model outputs a `<think>` block for symbolic reasoning before generating perceptual modalities. Modal tokens are aligned in a consistent autoregressive order to preserve temporal coherence and facilitate video rendering. We initialize text encoders from LLaMA2-7B, and retain the pretrained SpeechTokenizer and motion VQ-VAE modules as frozen during all LLM training. This modular design ensures representational stability while leveraging large-scale pretrained knowledge.

Diffusion-based Renderer. Our diffusion-based video renderer is built upon the WAN [18] backbone. To enable fine-grained control over multimodal synthesis, we introduce cross-attention layers into the video decoder that inject motion and audio features, allowing the model to produce synchronized body movement and speech directly from the autoregressively generated token sequence.

We train the model on 25fps video clips using $16 \times H100$ GPUs with the following settings: a constant learning rate of $1e-5$ with 100 warm-up steps, Adam optimizer with weight decay of 0.01, gradient clipping at 1.0, and bfloat16 mixed-precision training. We adopt optical flow-based filtering during preprocessing, using a minimum flow threshold of 0 and filtering frames based on percentile scores capped at 5. This ensures temporal coherence and training stability during synthesis learning.

Gaussian-based Renderer. Our Gaussian-based renderer is developed following the pipelines of [4, 22], leveraging 3D point cloud accumulation and SMPL-X-driven body control. The model is trained using a learning rate of $5e-4$ with a cosine annealing scheduler, Adam optimizer with a weight decay of 0.01, and runs on a single H100 GPU.

2.5. Inference Settings

During inference, the model operates in an autoregressive decoding mode with a unified token stream. The prompt first triggers internal `<think>` reasoning, followed by text, speech, and motion token generation. Generated acoustic tokens are converted to waveforms and re-encoded for video synthesis, while motion tokens are decoded into SMPL-X poses. Video rendering is subsequently performed by either a diffusion-based or Gaussian-based renderer, depending on the output mode.

2.6. Experimental Settings and Baselines

Data Splits. For all experiments, we adopt a 7:1 train-test split for each dataset. Multimodal dialogue and instruction-following evaluations are conducted on our curated test sets built from BEAT v2 and HumanML3D, respectively. For foundational tasks such as text-to-motion (T2M) and speech-to-motion (S2M), we use the similar test sets from HumanML3D [5] and BEAT v2 [13].

Baselines and Training Protocol. To ensure a rigorous and fair evaluation, we compare U-Mind against a diverse set of representative baselines. For methods whose pretrained weights are not publicly available, specifically SOLAMI [8], we retrain the model from scratch on our curated dataset following the official implementation and training protocols. For other baselines, including LOM [3], EMAGE [13], CaMN [12], and DisCo [11], we utilize their publicly released weights and codebases.

Furthermore, we extend our comparison to include three specialized categories of baselines: (1) **Cascaded Pipeline:** To benchmark against non-unified approaches, we construct a modular system integrating LLaMA2-Chat-7B [17] for dialogue, Orpheus-TTS [2] for speech synthesis, and LOM [3] for motion generation. (2) **Dialog-centric MLLMs:** We incorporate Next-GPT [19] to evaluate high-level multimodal conversation capabilities. (3) **Task-specific SOTA:** For the text-to-motion (T2M) task, we compare U-Mind against recent state-of-the-art models, including Atom [6] and PersonaBooth [9], ensuring a comprehensive assessment of motion fidelity and diversity.

Evaluation Metrics. Beyond standard metrics, we incorporate [7, 8, 10]: (1) **MPJPE/PA-MPJPE:** Average Euclidean distance between predicted and ground-truth 3D joints to assess structural fidelity. (2) **Accel:** Average joint acceleration to quantify motion smoothness and jitter. (3) **Angle Error (AE):** Mean Euclidean distance between predicted and ground-truth joint Euler angles. In our tables, the “**Dataset**” row denotes the performance of reference samples from the ground-truth test set, serving as the upper-bound gold standard.

3. Additional Qualitative Results

We provide extra cases in Fig 1 that demonstrate the model’s ability to interpret diverse user queries, perform intermediate reasoning, and generate coherent multimodal responses. These examples cover open-domain and context-dependent questions, showing how the system grounds its answers with appropriate gestures, synchronized speech, and internally reasoning traces. The responses reflect both semantic understanding and communicative clarity.

4. Additional Quantitative Evaluations

4.1. Multimodal Dialogue and User Study

To mitigate potential LLM evaluator bias, we conducted a user study with 30 participants across 20 randomized samples. As shown in Table 1, U-Mind significantly outperforms baselines in *Naturalness* and *Relevance*. These results are cross-validated using Gemma-2 as an independent automated judge, confirming that U-Mind’s superiority is consistent across both human perception and logical metrics.

Table 1. Updated Multimodal Dialogue Evaluation.

Method	Human Eval (1-5)		Gemma-2 Eval (1-10)	
	Rel. \uparrow	Nat. \uparrow	Rel. \uparrow	Nat. \uparrow
LLM+TTS+LOM	3.75	2.30	7.70	4.39
SOLAMI	1.71	2.40	0.89	5.28
Next-GPT	3.52	2.71	6.54	5.59
Ours	4.50	4.25	<u>7.33</u>	7.84

4.2. Expanded Text-to-Motion Comparison

We compare U-Mind with recent SOTA T2M methods in Table 2. While SOLAMI shows lower FGD, its poor Diversity score indicates mode collapse. U-Mind maintains high-fidelity dynamics (lowest MPJPE) and variance, essential for expressive agents.

Table 2. Expanded T2M Comparison.

Methods	FGD \downarrow	Div \uparrow	MPJPE \downarrow	PAMPJPE \downarrow	Accel \downarrow
LOM	14.22	10.42	316.94	225.05	236.19
SOLAMI	8.65	7.37	321.07	227.66	152.34
Atom	13.18	9.37	<u>193.32</u>	<u>112.84</u>	163.64
Ours	12.69	10.71	123.60	79.32	137.55

References

- [1] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222, 2020. 1
- [2] Canopy AI. Orpheus-tts: Towards human-sounding speech. <https://github.com/canopyai/OrpheusTTS>, 2025. Accessed: 2025-11-13. 1, 3
- [3] Changan Chen, Juze Zhang, Shrinidhi K Lakshmikanth, Yusu Fang, Ruizhi Shao, Gordon Wetzstein, Li Fei-Fei, and Ehsan Adeli. The language of motion: Unifying verbal and non-verbal language of 3d human motion. In *CVPR*, pages 6200–6211, 2025. 3
- [4] Jianchuan Chen, Jingchuan Hu, Gaige Wang, Zhonghua Jiang, Tiansong Zhou, Zhiwen Chen, and Chengfei Lv. Taoavatar: Real-time lifelike full-body talking avatars for augmented reality via 3d gaussian splatting. In *CVPR*, pages 10723–10734, 2025. 1, 3
- [5] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, pages 5152–5161, 2022. 1, 3
- [6] Haonan Han, Xiangzuo Wu, Huan Liao, Zunnan Xu, Zhongyuan Hu, Ronghui Li, Yachao Zhang, and Xiu Li. Atom: Aligning text-to-motion model at event-level with gpt-4vision reward. In *CVPR*, pages 22746–22755, 2025. 3
- [7] Fangzhou Hong, Vladimir Guzov, Hyo Jin Kim, Yuting Ye, Richard Newcombe, Ziwei Liu, and Lingni Ma. EgoIm: Multi-modal language model of egocentric motions. In *CVPR*, pages 5344–5354, 2025. 3
- [8] Jianping Jiang, Weiye Xiao, Zhengyu Lin, Huaizhong Zhang, Tianxiang Ren, Yang Gao, Zhiqian Lin, Zhongang Cai, Lei Yang, and Ziwei Liu. Solami: Social vision-language-action modeling for immersive interaction with 3d autonomous characters. In *CVPR*, pages 26887–26898, 2025. 3
- [9] Boeun Kim, Hea In Jeong, JungHoon Sung, Yihua Cheng, Jeongmin Lee, Ju Yong Chang, Sang-Il Choi, Younggeun Choi, Saim Shin, Jungho Kim, et al. Personaboost: Personalized text-to-motion generation. In *CVPR*, pages 22756–22765, 2025. 3
- [10] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. Towards a common framework for multimodal generation: The behavior markup language. In *IVA*, pages 205–217, 2006. 3
- [11] Haiyang Liu, Naoya Iwamoto, Zihao Zhu, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Disco: Disentangled implicit content and rhythm learning for diverse co-speech gestures synthesis. In *MM*, pages 3764–3773, 2022. 3
- [12] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *ECCV*, pages 612–630, 2022. 3
- [13] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *CVPR*, pages 1144–1154, 2024. 1, 3
- [14] Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*, 2023. 1
- [15] OpenAI. Whisper: A general-purpose speech recognition model. <https://github.com/openai/whisper>, 2022. Accessed: 2025-11-13. 1
- [16] Li Siyao, Tianpei Gu, Zhitao Yang, Zhengyu Lin, Ziwei Liu, Henghui Ding, Lei Yang, and Chen Change Loy.

- Duolando: Follower gpt with off-policy reinforcement learning for dance accompaniment. In *ICLR*, pages 1–11, 2024. [1](#)
- [17] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. [2](#), [3](#)
- [18] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. [1](#), [3](#)
- [19] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *ICML*, pages 1–11, 2024. [3](#)
- [20] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *ICCV*, pages 4210–4220, 2023. [1](#)
- [21] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. In *ACL*, pages 9637–9662, 2024. [2](#)
- [22] Youyi Zhan, Tianjia Shao, Yin Yang, and Kun Zhou. Real-time high-fidelity gaussian human avatars with position-based interpolation of spatially distributed mlps. In *CVPR*, pages 26297–26307, 2025. [3](#)
- [23] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Spechtokenizer: Unified speech tokenizer for speech language models. In *ICLR*, pages 1–11, 2023. [1](#)
- [24] Zeyi Zhang, Tenglong Ao, Yuyao Zhang, Qingzhe Gao, Chuan Lin, Baoquan Chen, and Libin Liu. Semantic gesticulator: Semantics-aware co-speech gesture synthesis. *TOG*, pages 1–17, 2024. [1](#)