Reasoning Gestures: LLM-Inferred Communicative Functions for Co-Speech Gesture Generation

Pinxin Liu[1], **Haiyang Liu**[2], **Yunlong Tang**[1], **Luchuan Song**[1]
[1]University of Rochester, [2]University of Tokyo

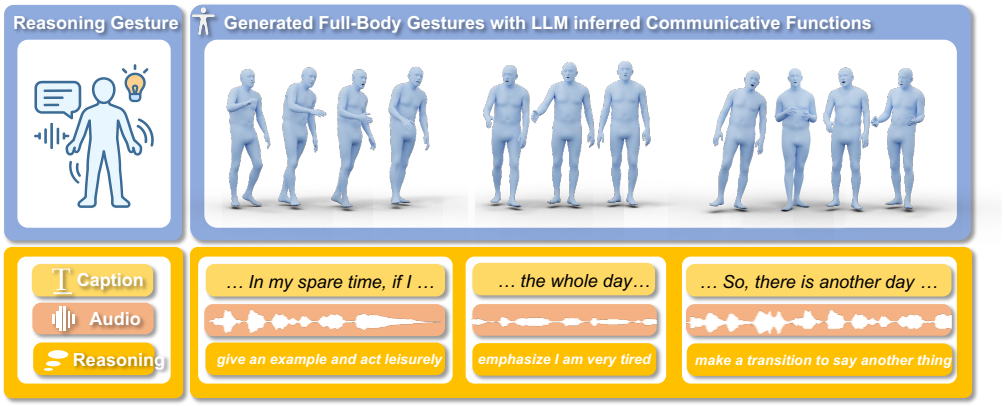


Figure 1: We present *Reasoning Gestures*, a reasoning-grounded framework for *communicative-function* controllable co-speech gesture generation. It infers high-level communicative functions via VLM reasoning and conditions motion generation on observer-centric function descriptions.

Abstract

When humans speak, gestures often realize communicative functions (*e.g.*, emphasizing key content, indicating referents, or depicting concepts). Most co-speech gesture generation methods, however, condition primarily on low-level cues (speech audio or transcripts) and do not explicitly represent these higher-level functions, leading to motions that synchronize with speech but remain under-specified in communicative meaning. To address this gap, we introduce **Reasoning Gestures**, a reasoning-grounded framework that conditions generation on interpretable communicative-function descriptions. We curate **InG: Interpreted Gestures** by augmenting BEAT-2 with sentence-level *VLM-inferred gesture-function annotations* that describe the discourse role of each gesture segment in context. These labels are *observer-centric interpretations* produced via constrained VLM reasoning rather than claims about speakers' private mental states. We further propose a **Reasoning Gesture Motion Tokenizer** that injects function signals into tokenized motion representations, enabling function-aware synthesis that is temporally aligned and communicatively grounded. Experiments on BEAT-2 show consistent gains in semantic consistency and overall quality, establishing new state-of-the-art. Project Page: `https://andypinxinliu.github.io/Reasoning-Gesture`

1 Introduction

Gestures accompany speech to convey ideas and facilitate comprehension De Ruiter et al. [2012], Song et al. [2023], forming an essential part of human communication. As AI advances, enabling virtual avatars to produce expressive gestures will be critical for digital avatar construction Tang et al. [2025], Huang et al. [2024], Song et al. [2024].

Recent works Liu et al. [2025b, 2023], Yi et al. [2023] have explored full-body gesture generation conditioned on speech audio or text transcripts. However, these approaches largely treat raw linguistic input as sufficient semantics, overlooking the deeper communicative functions (*e.g.*, emphasis, deixis, and affirmation), that implicitly govern gesture behavior. Recovering these latent functions is essential for generating semantically coherent non-verbal expressions.

Early systems Marsella et al. [2013], McNeill [1992] attempted to derive these communicative functions through rule-based templates, but such approaches lack scalability. Inspired by the linguistic–gesture link in psycholinguistics, we reframe gesture modeling as a reasoning task: the model first infers the communicative function underlying an utterance (as an observer-centric interpretation grounded in multimodal evidence), and then treats gestures as downstream realizations.

To achieve this, we augment the BEAT-2 dataset Liu et al. [2022a] with structured annotations of *LLM-inferred communicative functions* using Vision-Language Models, creating a large-scale **InG: Interpreted Gestures** dataset. This enriched corpus links speech, *reasoned function descriptions*, and corresponding gesture realizations, enabling function-aware gesture generation at scale.

Leveraging these LLM-inferred function annotations, we propose a CLIP-like gesture understanding model. It learns joint representations of speech rhythm, reasoned communicative functions, and motion dynamics via a hierarchical contrastive alignment framework, enabling it to capture both low-level rhythmic synchronization and high-level discourse roles.

Building upon this understanding model, we introduce the **Reasoning Gesture Tokenizer**, a novel quantization model that directly embeds communicative-function semantics into the motion representation space. Unlike prior works that discretize body parts independently without semantic grounding Mughal et al. [2025], Liu et al. [2025a], our tokenizer processes global body motion holistically and supervises latent representations using function-aware motion features. This design ensures that discrete motion tokens encode not only fine-grained motion patterns but also interpretable communicative meaning. With the proposed tokenizer, our approach generates gestures that are not only temporally synchronized but also semantically expressive and interpretable, advancing human-avatar communication. Our contributions can be summarized as follows:

- We formulate Gesture Generation as a *reasoning-grounded* task, and create **InG** Dataset with VLMs to obtain *inferred communicative function* annotations that augment the existing datasets.
- We introduce **Reasoning Gesture Tokenizer** to discretize global body motion while injecting communicative-function semantics into the latent space through semantic supervision.
- We demonstrate our method produces gestures that are temporally aligned with speech, communicatively grounded and semantically meaningful, improving realism and interpretability.

2 Related Works

Co-speech Gesture Generation. Existing works on co-speech gesture generation typically use skeleton- or joint-level pose representations. Several methods Liu et al. [2022b], Deichler et al. [2023], Xu et al. [2023], Liu et al. [2024a], Zhang et al. [2024a], Liu et al. [2025b] learn hierarchical semantics or apply contrastive learning to align audio and gesture embeddings. HA2G Liu et al. [2022b] constructs multi-level audio-motion embeddings, while TalkShow Yi et al. [2023], CaMN Liu et al. [2022a], and EMAGE Liu et al. [2023] introduce large-scale datasets for joint face-body modeling with GPT-style decoding. More recent models such as MambaTalk Xu et al. [2024], DiffSHEG Chen et al. [2024], and GestureLSM Liu et al. [2025a] focus on efficient flow matching Li et al. [2025], Zhu et al. [2024]. Semantic Gesticulator Zhang et al. [2024b] and RAG-Gesture Mughal et al. [2025] leverage LLMs to retrieve discourse-relevant gestures as references. In contrast, we explore the modeling of communicative functions as explicit semantics to control gesture generation.

Vision Tokenization for Generation. In visual generation, tokenization encodes raw pixels into compact representations Van Den Oord et al. [2017], Rombach et al. [2022]. Recent studies like REPA Yu et al. [2024] and Re.vs.Gen Yao et al. [2025] show that aligning generation representation with understanding improves synthesis quality. We further introduce the **Reasoning Gesture Tokenizer**, supervising motion tokenization with communicative-function-aligned semantics, and demonstrate this strategy enhances gesture generation quality.

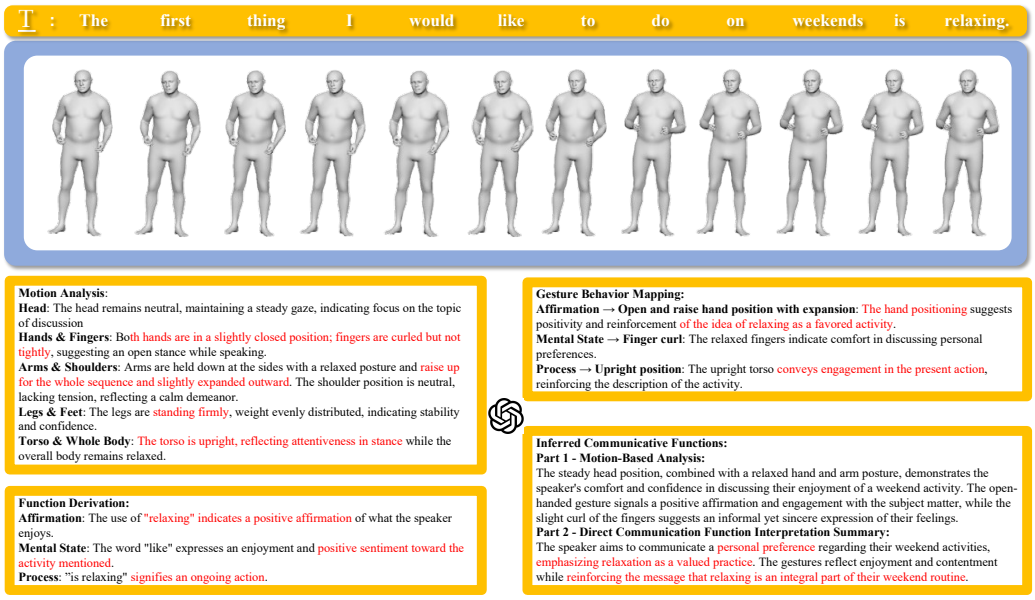


Figure 2: Overview of the annotation pipeline. Sentence-level segments and word-aligned keyframes, together with rule-based motion descriptors, are provided to the VLM, which produces (1) region-wise motion descriptions, (2) communicative function labels, (3) gesture behavior mappings, and (4) *reasoned discourse-level interpretations* of how the gesture supports the utterance. This structured annotation enables scalable creation of a communicatively grounded, visually aligned gesture dataset.

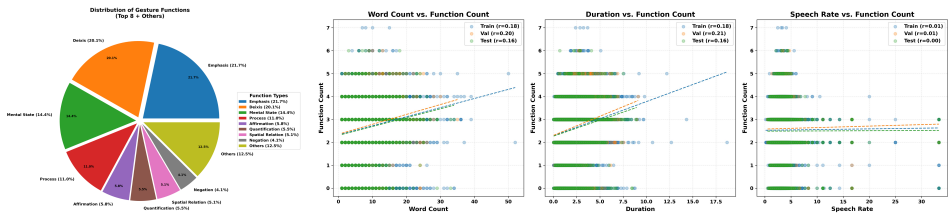
3 Interpreted Gesture Dataset (InG)

Gestures systematically reflect semantic, pragmatic, and rhetorical structures Marsella et al. [2013], McNeill [1992], Saund and Marsella [2021]; for example, affirmation often aligns with nodding and comparison with lateral gestures that follow spatial metaphors. Modeling such communicative functions is critical for generating expressive gestures. Marsella et al. [2013] introduced a rule-based system that derives communicative functions from syntactic and lexical patterns and maps them to gesture classes via hand-crafted rules. While effective, such systems rely on fixed rules and limited dictionaries, restricting scalability and domain coverage. Building on these insights, we propose a scalable pipeline using VLMs to automatically infer communicative functions and map them to gestures, enabling large-scale, data-driven modeling of communicatively grounded gestures. Importantly, we treat these VLM-inferred functions and discourse interpretations as *observer-centric, operational signals* for gesture generation rather than attempting to recover a single "true" speaker mental state; they are designed to be stable, interpretable, and controllable.

3.1 Dataset Construction

We introduce **InG** (Interpreted Gestures), a dataset that augments BEAT-2 Liu et al. [2022a] and Audio2PhotoReal Ng et al. [2024] with structured, reasoning-grounded annotations that link discourse-level communicative functions to gesture behaviors. Unlike prior datasets focused primarily on motion fidelity or prosody alignment, InG supports *function-controllable* co-speech gesture generation. We employ a two-stage VLM-based strategy, illustrated in Fig. 2.

Annotation Protocol. We design a unified annotation pipeline that integrates interpretable rule-based motion descriptors with structured VLM prompting to yield communicatively grounded gesture annotations. Rather than directly "asking GPT for a label", the pipeline first converts raw motion into constrained, symbolic descriptors, and then uses the VLM only within this structured schema. For training data, the VLM receives sentence context together with word-aligned keyframes and motion descriptors, enabling inference of communicative functions, motion descriptions, and a discourse-level rationale of how nonverbal behavior supports the utterance. For testing data, annotations are generated solely from transcripts to prevent leakage from observed gestures. Across all segments, we explicitly allow *multi-function* labels (e.g., Emphasis + Deixis + Mental State), reflecting the fact that gesture units often serve multiple discourse functions. Detailed prompting strategies, category definitions, and validation procedures are provided in the Appendix.



(a) (b) (c) (d)

Figure 3: (a) Distribution of communicative function types in InG dataset. (b-d) Correlations between function count and utterance-level features: word count, duration, and speech rate. Function count correlates positively with utterance length and duration.

Rule-Based Motion Description Generation. To provide interpretable motion descriptors aligned with linguistic units, we design a rule-based analysis procedure that converts raw pose sequences into structured gesture descriptions. The process operates in three stages: (1) *Temporal Windowing.* Each sentence is divided into sub-windows of 1–2 seconds based on word-level timestamps. Consecutive words within the same span are grouped together, while longer segments are recursively partitioned. This ensures that each window corresponds to a manageable co-speech gesture unit. (2) *Canonicalized Motion Representation.* All sequences are normalized to a canonical space relative to the body root. For elbows, wrists, head, and fingers, motions are represented by full 3D joint pose angle trajectories or positional change over the window, while for hands we track 3D positional displacements over time. (3) *Movement Characterization.* Within each window, trajectories are smoothed and segmented by direction and amplitude. Segments are categorized into qualitative tiers (*very slight*, *slight*, *moderate*, *significant*), where thresholds are defined relative to the standard deviation of motion distributions and refined through pilot annotation experiments and human evaluation. Detected patterns are then labeled as monotonic (e.g., "slightly forward"), bidirectional (e.g., "forward then backward"), or oscillatory (e.g., "repeated side-to-side"). Each description is anchored to visual evidence by presenting the first, last, and a representative keyframe from the corresponding word span, ensuring that textual descriptors remain verifiable against pose snapshots. We defer algorithm details to the Appendix.

Prompt-Based Gesture Function and Discourse Interpretation Annotation. The rule-based descriptors and reference frames form the grounding for higher-level annotation with GPT-4o-mini. Given segmented utterances and motion summaries, we design a structured prompt for multi-stage annotation: (1) *Motion Analysis* describing body poses across regions (head, arms, fingers, torso) using the rule-based descriptors together with reference frames; (2) *Function Derivation* identifying one or more communicative functions (e.g., emphasis, negation, deixis), explicitly allowing multi-label combinations; (3) *Gesture Behavior Mapping* linking communicative functions to gesture forms; and (4) *Discourse-Level Interpretation* summarizing, at the level of discourse pragmatics, how verbal and nonverbal cues jointly support the utterance (*e.g.*, what the gesture is doing communicatively in context), without claiming access to a speaker's private mental state. The model is required to output in a fixed motion to function to interpretation schema, which constrains free-form generation and reduces hallucinations. This layered approach ensures that annotations are both semantically interpretable and physically grounded, enabling function-controllable gesture generation.

Human in the Loop Validation. To ensure reliability of the VLM-generated annotations, we adopt a human-in-the-loop filtering stage. For each utterance, the LLM is instructed to produce five independent candidate annotations following the structured protocol described above. These candidates vary in phrasing and granularity, but all conform to the same motion to function to interpretation schema. Human labelers then evaluate the set of five responses and remove those that contain incorrect interpretations, inconsistent mappings, or implausible discourse-level rationales. Only the remaining high-quality annotations are retained in the dataset. In practice, this combination of rule-based preprocessing, schema-constrained prompting, and human filtering substantially mitigates typical VLM noise and bias while preserving scalability.

3.2 Dataset Statistics and Analysis

Our InG dataset includes 34,641, 3,598, and 9,674 annotated utterances for train/val/test splits, respectively, each enriched with motion-grounded or transcript-only *LLM-reasoned* communicative function annotations. We defer the specific statistics of the distribution of 16 function types, covering pragmatic categories such as Emphasis, Deixis, Mental State, and Process, to the Appendix.

As shown in Fig. 3 (a), Emphasis (21.7%) and Deixis (20.1%) are the most prevalent functions, followed by Mental State (14.4%) and Process (11.0%). We further analyze correlations between linguistic features and function density (Fig. 3, b–d). Function counts correlate positively with utterance length ($r = 0.18$–0.20) and duration ($r = 0.16$–0.21), but show minimal correlation with speech rate ($r \approx 0.00$–0.01). This suggests that communicative function density scales with informational content rather than speaking speed.

While function derivations follow a well-defined ontology, gesture behavior mappings and discourse-level interpretations exhibit open-ended variability across speakers and contexts. To ensure interpretability, our prompt design grounds mappings in established literature (McNeill McNeill [1992], Kendon Saund and Marsella [2021]). Additional statistics and analysis are provided in the Appendix.

Annotation Validation Study. We evaluate annotation quality through a pairwise human preference study on 100 randomly sampled utterances. Each utterance is annotated using: (1) our train-style VLM protocol (motion input), (2) our test-style VLM protocol (transcript only), and (3) a free-form human annotation. For each utterance, two comparisons are created: VLM (train-style) vs. human, and VLM (test-style) vs. human.

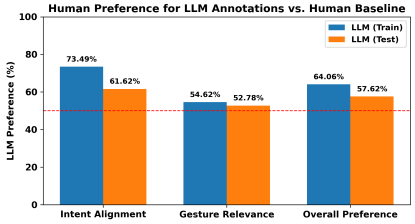


Figure 4: Pairwise preference for VLM vs. human annotations. Blue: train-style prompting; Orange: test-style (transcript only). Red dashed line indicates chance (50%).

The human baseline is produced by two expert annotators (non-authors) with prior exposure to gesture and communication literature. For each clip, they are given the speech transcript, rendered motion, and the same communicative function ontology as our model, and asked to write free-form descriptions of gesture functions and discourse roles without seeing any VLM outputs.

Three raters independently judge across: **E1: Function Alignment** (which better captures the communicative function/discourse role), **E2: Gesture Relevance** (which provides more plausible function–gesture mappings), and **E3: Overall Preference** (clarity and function–gesture alignment).

Final preferences are determined by majority vote. Shown in Fig. 4, both VLM protocols outperform the free-form human baseline. Inter-rater agreement averaged 0.76 Fleiss' κ, indicating substantial consistency. This comparison shows that our structured motion–function–interpretation pipeline yields consistent annotations and directly usable for modeling than free-form text.

4 Reasoning-Grounded Gesture Generation

4.1 Communicative Function Understanding

To leverage *LLM-reasoned communicative function descriptions* for co-speech gesture synthesis, we propose a CLIP-like understanding model, H-AuMoCLIP, to better encode discourse-level function features. We build this Audio–Motion CLIP on top of TANGO's Liu et al. [2024a] audio-motion alignment model (AuMoCLIP) by explicitly incorporating communicative-function signals.

Function Semantic Fusion. AuMoCLIP trains its joint embedding space using contrastive learning, where its audio tower combines acoustic features with BERT-encoded speech transcripts. Building upon this, we introduce *Function Semantic Fusion* to incorporate communicative-function semantics and create a richer, function-aware joint embedding space. Our fusion mechanism combines transcript and function embeddings in a linguistically grounded manner. Following TANGO, we align transcript tokens to the audio timeline using a CTC-based model, and encode both the transcript and the *LLM-inferred function descriptions* with separate BERT encoders. In addition to TANGO, the aligned transcript embeddings serve as queries in a cross-attention module, with function embeddings as keys and values. The output captures context-specific discourse cues, which are then concatenated with wav2vec2 audio features to form the final high-level representation. This composite embedding is used in contrastive training to jointly align audio, motion, and function semantics.

Functionality. As in Fig. 5, the resulting encoders improve generation by serving two roles: (1) the motion encoder provides semantic supervision for gesture tokenization (Sec. 4.2), and (2) the audio encoder conditions gesture generation with both rhythmic and function-level signals (Sec. 4.3).

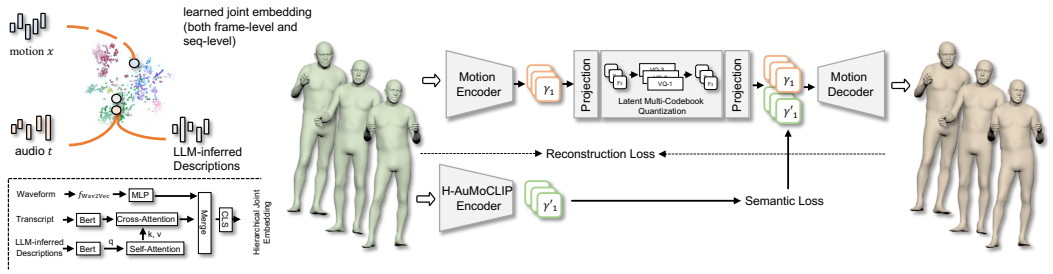


Figure 5: **Overview of our reasoning-grounded gesture generation framework.** Left: H-AuMoCLIP learns a hierarchical joint embedding of motion, audio, and communicative function. Transcript embeddings (BERT) aligned via CTC serve as queries in a cross-attention module with function embeddings as keys/values. The resulting semantic features are concatenated with wav2vec2 audio features for contrastive learning. Right: Motion is quantized via a multi-codebook VQ module and supervised by semantic features from H-AuMoCLIP, enabling expressive gesture generation.

4.2 Reasoning Gesture Tokenizer

Prior works Mughal et al. [2025], Liu et al. [2025a, 2023] model different body regions using separate encoders, decoders, and codebooks per body part. While this design encourages local disentanglement, it has three limitations: (1) high training and inference cost, (2) poor global body coherence, and (3) weak connection to utterance-level semantics. To address these issues, we propose a **Reasoning Gesture Tokenizer** with two core innovations: (1) learning disentangled latent factors within a unified global body representation through multi-codebook quantization, and (2) embedding explicit communicative-function semantics into the motion representation through semantic supervision.

Latent Multi-Codebook Quantization. Instead of separate body-specific quantizers, we discretize the global motion latent $f \in \mathbb{R}^d$ using a set of n independent codebooks $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_n\}$. The latent is partitioned into n chunks $\{f_1, f_2, \ldots, f_n\}$, each quantized separately against its own codebook:

$$\hat{f} = \text{Concat}\big(\mathcal{Q}(\mathcal{C}_1, f_1), \ldots, \mathcal{Q}(\mathcal{C}_n, f_n)\big), \tag{1}$$

where $\mathcal{Q}(\mathcal{C}_i, f_i)$ denotes vector quantization of chunk f_i against codebook $\mathcal{C}_i$, and $\hat{f} \in \mathbb{R}^d$ is the resulting (concatenated) quantized motion latent. This design maintains global body context while allowing structured disentanglement to emerge across codebooks during training. Unlike prior methods, our tokenizer processes the full body motion jointly, enabling better modeling of coordination patterns necessary for communicative gestures.

Communicative-Function Semantic Supervision. To embed context-dependent function signals into motion representations, we supervise the quantized latents using features from the pretrained H-AuMoCLIP motion encoder (Sec. 4.1). Let $\mathbf{m}_t \in \mathbb{R}^{d_m}$ denote the H-AuMoCLIP motion-encoder feature at timestep t, serving as the *semantic target*. A linear projection $g(\cdot)$ maps the quantized output $\hat{f}$ (Eq. 1) into the same space, yielding $\tilde{f}_t = g(\hat{f})_t \in \mathbb{R}^{d_m}$. We then minimize a temporal cosine-distance loss

$$\mathcal{L}_{\text{sem}} = \frac{1}{T} \sum_{t=1}^{T} \left(1 - \frac{\tilde{f}_t \cdot \mathbf{m}_t}{\|\tilde{f}_t\| \, \|\mathbf{m}_t\|} \right), \tag{2}$$

which is non-negative by construction since cosine similarity lies in $[-1, 1]$. This supervision enforces semantic alignment with utterance-level communicative-function cues derived from speech.

Training Objective. Our full training objective includes: (i) a reconstruction loss $\mathcal{L}_{\text{R}}$, (ii) a vector quantization loss $\mathcal{L}_{\text{VQ}}$, and (iii) the proposed semantic supervision loss $\mathcal{L}_{\text{sem}}$:

$$\mathcal{L} = \mathcal{L}_{\text{R}} + \lambda_{\text{VQ}} \mathcal{L}_{\text{VQ}} + \lambda_{\text{sem}} \mathcal{L}_{\text{sem}}. \tag{3}$$

We empirically set $\lambda_{\text{sem}} = 1$ and $\lambda_{\text{VQ}} = 0.25$ to ensure that quantized motion tokens preserve both fine-grained fidelity and semantic coherence across time.

4.3 Gesture Generation

We then leverage Sec. 4.1 and Sec. 4.2 to enhance gesture generation. We adapt GestureLSM Liu et al. [2025a] by replacing its original audio encoder and motion tokenizer. Specifically, we replace

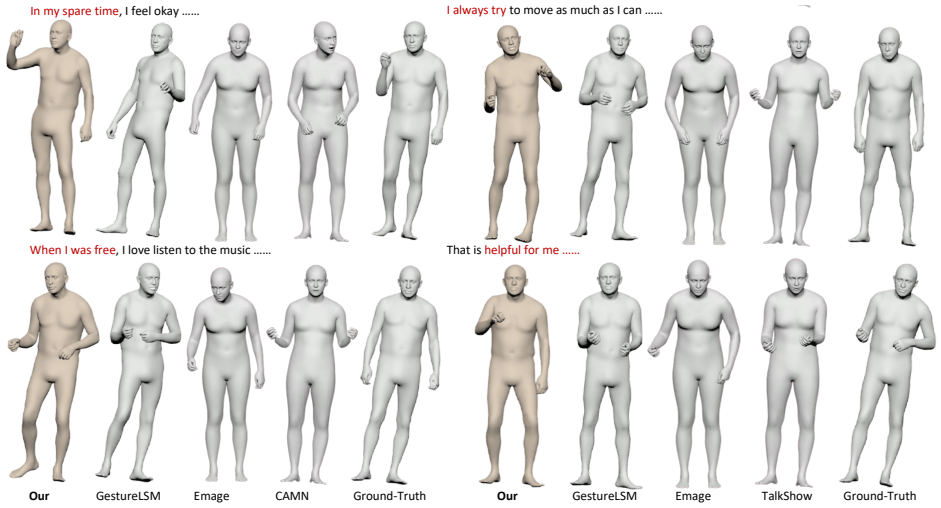


Figure 6: **The subjective comparisons.** Compared with other methods, Reasoning-Gesture presents more natural and coherent motion patterns to represent specific words or phrases (highlighted in red).

Table 1: **Comparison with methods trained on BEAT-2.** We demonstrate superior performance, especially when generalizing across multiple speaker identities.

	1 Speaker			All Speakers			Novel 5 speakers		
	FGD↓	BC→	Div.→	FGD↓	BC→	Div.→	FGD↓	BC→	Div.→
GT		0.703	11.97		0.477	7.29		0.671	8.93
CaMN Liu et al. [2022a]	0.604	0.676	9.97	0.512	0.200	5.58	0.812	0.563	6.71
EMAGE Liu et al. [2023]	0.570	0.793	11.41	0.692	0.284	6.06	0.936	0.643	7.47
Audio2Photoreal Ng et al. [2024]	1.02	0.550	**12.47**	0.849	0.326	6.24	1.012	0.464	5.97
RAG-Gesture Mughal et al. [2025]	0.879	0.730	12.62	0.447	**0.471**	9.03	-	-	-
GestureLSM Liu et al. [2025a]	0.408	**0.714**	13.24	0.446	0.525	9.23	0.664	0.621	10.45
Ours	**0.379**	0.690	11.00	**0.256**	0.534	**6.68**	**0.441**	**0.686**	**9.39**

GestureLSM's audio encoder with our function-aware audio encoder, which extracts hierarchical representations capturing both rhythmic audio features and communicative-function signals. Furthermore, we replace its RVQ-based tokenizer with our Reasoning Gesture Tokenizer. Additional architectural details of the original GestureLSM are deferred to the Appendix.

5 Experiments

We conduct main experiments on BEAT2 Liu et al. [2023], which comprises 60 hours of SMPL-based conversational or speech gesture data collected from 25 speakers. It contains 1,762 sequences, each with an average duration of 65.66 seconds, following the train-validation-test split protocol defined in EMAGE Liu et al. [2023]. In addition, we explore its application on Audio2Photoreal Ng et al. [2024], which provides about 8 hours of dyadic interactions between listening and speaking actions. We defer the results on Audio2PhotoReal to Appendix with video demos.

5.1 Quantitative Comparisons

Metrics. We evaluate Fréchet Gesture Distance (FGD) Yoon et al. [2020] for pose sequence angle distributional similarity, Diversity (Div.) Li et al. [2021a] as the average L1 distance across clips, and Beat Constancy (BC) Li et al. [2021b] for speech-motion synchronization.

Evaluation Results. We evaluate both quality and generalization in three settings: single-speaker, all-speaker, and zero-shot (training on 20 speakers and testing on 5 unseen speakers). As shown in Tab. 1, our method consistently outperforms baselines across all settings. We attribute these gains to our function-aware motion representation and communicative-function conditioning, which encourage more coherent and less artifact-prone gesture patterns. More detailed comparisons are provided in the Appendix. For runtime, communicative-function descriptions are generated from transcripts via an API prior to model execution, so inference incurs no additional online reasoning cost. Our method retains the efficiency of GestureLSM Liu et al. [2025a] while improving performance through a unified motion representation, achieving 30.4 FPS generation on an H100 GPU.

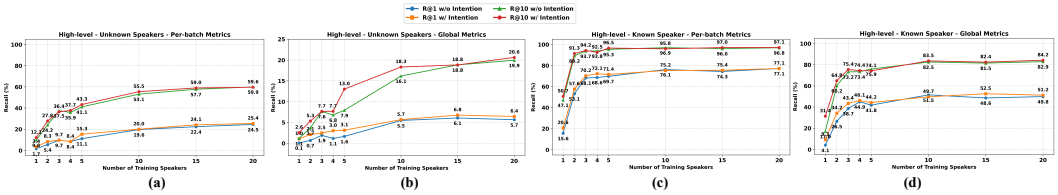


Figure 7: **Hierarchical Alignment Analysis.** Retrieval-based evaluation of H-AuMoCLIP across semantic feature settings. We report Recall@1/10 under per-batch and global retrieval for both seen and unseen speakers. Incorporating communicative-function semantics consistently improves retrieval, with the largest gains in low-data regimes (few training speakers). Increasing speaker diversity further enhances generalization and overall retrieval robustness.

5.2 Qualitative Comparisons

Evaluation Results. As depicted in Fig. 6, our approach generates gestures that exhibit improved rhythmic alignment and a more natural appearance compared to existing methods. For example, when conveying *"helpful for me"*, our method directs the subject to extend left hand forward, effectively representing the function of an explanation. In contrast, competing methods fail to capture this nuance, often generating static or unnatural poses where arms remain down.

User Study. We conducted a user study with **24** participants over a pool of 160 video samples (40 from each of GestureLSM Liu et al. [2025a], EMAGE Liu et al. [2023], CAMN Liu et al. [2022a], and Ours). Each participant rated **20 videos** drawn from the pool in randomized order, yielding 480 ratings in total ($\sim$120 per method, with each clip rated by multiple raters). Ratings were on a 1 (lowest) to 5 (highest) scale based on three criteria: (1) *realness*, (2) speech–gesture *synchrony*, and (3) *smoothness*. As shown in Tab. 2, our method outperforms other methods across all criteria, achieving higher Mean Opinion Scores (MOS). We follow a standard 1–5 Likert MOS scale with textual anchors (1 = very poor, 3 = acceptable, 5 = excellent), and compute MOS as the per-clip average rating averaged across raters. Detailed statistics are reported in Appendix.H.

Table 2: Subjective Evaluation from 24 participants shown as Mean Opinion Scores (MOS).

Methods	MOS_1	MOS_2	MOS_3
EMAGE	2.01	2.42	2.31
CAMN	1.34	2.23	2.14
GestureLSM	3.43	3.61	3.48
Ours	**3.76**	**4.11**	**3.92**

6 Ablation Studies and Analysis

We conduct ablations and analyses to evaluate (i) hierarchical audio–function and audio–motion alignment in H-AuMoCLIP, (ii) key design choices in the tokenizer and generator. We defer further analyses of (iii) the impact and reliability of our VLM-based annotations to the Appendix.

6.1 Evaluation of Hierarchical AuMoCLIP

We evaluate H-AuMoCLIP using **retrieval-based metrics** that quantify how well audio and communicative-function features align with corresponding motion features. We report Recall@1 and Recall@10 under two settings: **per-batch** (retrieval within a batch of 128) and **global** (retrieval across the entire test set). Retrieval is performed at the sequence level by mean-pooling audio and motion features for full-clip matching (following the protocol in TANGO Liu et al. [2024a]). We further evaluate generalization on both **seen** (in-domain) and **unseen** (out-of-domain) speakers to assess robustness across speaker identities.

Function-Aware Alignment Improves Performance. Communicative function semantics improves retrieval, with the largest gains observed in low-data regimes (1–5 speakers) (Fig. 7). Improvements are more pronounced for high-level retrieval than for low-level retrieval, suggesting that function-aware modeling better captures sequence-level semantics beyond rhythmic synchronization.

Speaker Diversity Enhances Generalization. Training on multiple speakers substantially improves generalization to unseen speakers. Even with comparable total training data, models trained on more speaker identities achieve higher high-level retrieval, indicating that speaker diversity encourages learning more robust audio–motion correspondences.

6.2 Evaluation of Reasoning Gesture Tokenizer

We evaluate our Reasoning Gesture Tokenizer using **reconstruction FGD (rFGD)**, **codebook utility**, and **L1 reconstruction error**; generation metrics are reported in Sec. 5.1.

Table 3: **Quantitative Ablation Results.** (Left) Tokenizer ablations comparing design variants and levels of communicative-function semantic supervision. (Right) Gesture generation ablations evaluating architectural modules, semantic conditioning signals, and sensitivity to noisy or mismatched function descriptions.

Model	rFGD↓	Utility↑	L1↓	FGD↓	BC→	Div.→
EMAGE (VQ)	0.0867	34.35	0.3643	0.512	0.623	7.76
RAG-Gesture (Contin.)	0.0423	-	0.448	0.423	0.625	8.64
ProbTalk (PQ)	0.0283	53.24	0.3283	0.451	0.656	7.96
GestureLSM (RVQ)	0.0012	27.56	**0.1557**	0.314	0.527	8.64
Ours (Multi-VQ)						
only transformer	0.0021	97.25	0.3043	0.332	0.474	8.01
only CNN	0.0013	98.43	0.2656	0.348	0.671	7.96
Optimal Quantizer Design	**0.0011**	**98.84**	0.2468	0.315	0.512	7.67
+Low-level Supervision	0.0014	96.72	0.2566	0.295	**0.477**	5.96
+High-level (seq)	0.0014	96.68	0.2512	0.286	0.614	6.21
+High-level (temporal)	0.0012	97.17	0.2488	**0.256**	0.534	**6.68**

Model	FGD↓	BC→	Div.→
GestureLSM Liu et al. [2025a]	0.446	0.525	9.23
+ DiT Peebles and Xie [2023]	0.354	0.523	6.63
+ Function Condition	0.339	0.545	6.44
+ H-AuMoCLIP	0.314	0.527	8.64
+ Reasoning-Tokenizer	**0.256**	**0.534**	**6.68**
Sensitivity Analysis			
Re-summarized intent	0.256	0.533	6.66
Unmatched intent	0.284	0.521	6.47
Window = 0.7	0.257	0.533	6.64
Window = 0.5	0.256	0.533	6.67
w/o intent text	0.281	0.523	6.54

Tokenizer Comparison. To isolate the effect of motion representation, we use the same generator architecture and swap in latent codes produced by different tokenizers. As shown in Tab. 3 (Left), our Reasoning Gesture Tokenizer achieves substantially lower rFGD and higher diversity than prior tokenizers Liu et al. [2023, 2024b], Mughal et al. [2025], Liu et al. [2025a], indicating improved reconstruction fidelity and richer discrete motion coverage.

Effect of Communicative-Function Semantic Supervision. We study three semantic supervision strategies using H-AuMoCLIP motion features: (a) CNN-layer supervision for low-level spatial detail, (b) transformer-layer supervision with mean pooling for clip-level semantics, and (c) supervision with frame-level temporal alignment. Tab. 3 (Left) shows that injecting communicative-function supervision into token learning consistently improves downstream generation quality while preserving strong reconstruction fidelity. Notably, frame-level temporal supervision performs best, suggesting that temporally aligned function–motion grounding is critical for learning expressive representations.

6.3 Evaluation of Gesture Motion Generation

Design Analysis. Replacing the GestureLSM generator with a DiT architecture improves generation performance. We further examine how different semantic conditioning signals affect gesture generation. Conditioning on sentence-level BERT embeddings provides only marginal gains (Tab. 3, Right), whereas conditioning on communicative-function embeddings derived from H-AuMoCLIP yields substantial improvements in FGD and diversity. Together, these results indicate that generation benefits from both function-aware alignment and well-structured discrete motion representations.

Sensitivity Analysis. For robustness evaluation, using LLM-rewritten paraphrases of function descriptions yields negligible changes in FGD and diversity, suggesting low sensitivity to surface-level wording. When using temporally mismatched or missing function text, FGD increases slightly and synchronization degrades marginally, but overall generation remains stable.

Further Analyses (in Appendix). We defer two additional studies to the supplementary material: (i) an ablation on annotation quality and a comparison of audio, motion-description, and communicative-function descriptions as control signals (App.F.1); and (ii) a semantic-consistency evaluation that probes function-aware retrieval and stratifies generation quality across neutral/beat-like, iconic, and metaphoric segments (App.F.2). The conclusions are consistent with the main results: schema-constrained, training-time function annotations yield the best overall trade-off, and function supervision delivers the largest gains on semantically loaded (iconic and metaphoric) segments.

7 Conclusion

We present **Reasoning Gestures**, a reasoning-grounded framework that formulates co-speech gesture generation as conditioning on high-level *communicative functions*. We curate **InG: Interpreted Gestures** by augmenting BEAT-2 with structured, *VLM-inferred* function annotations designed as observer-centric, controllable signals. We then introduce the **Reasoning Gesture Tokenizer**, which discretizes global motion while injecting communicative-function semantics into the representation space through targeted semantic supervision. By bridging discourse-level function descriptions and motion behaviors, our method generates gestures that are temporally synchronized and communicatively grounded. Experiments demonstrate consistent improvements in gesture quality, providing a controllable and modular foundation for expressive gesture generation in digital humans and embodied AI.

References

Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. DiffSHEG: A Diffusion-Based Approach for Real-Time Speech-driven Holistic 3D Expression and Gesture Generation, 2024. URL https://arxiv.org/abs/2401.04747.

Jan P De Ruiter, Adrian Bangerter, and Paula Dings. The Interplay Between Gesture and Speech in the Production of Referring Expressions: Investigating the Tradeoff Hypothesis. *Topics in cognitive science*, 4(2):232–248, 2012.

Anna Deichler, Shivam Mehta, Simon Alexanderson, and Jonas Beskow. Diffusion-based co-speech gesture generation using joint text and audio representation. In *International Conference on Multimodal Interaction*, ICMI '23. ACM, October 2023. doi: 10.1145/3577190.3616117. URL http://dx.doi.org/10.1145/3577190.3616117.

Chao Huang, Dejan Markovic, Chenliang Xu, and Alexander Richard. Modeling and driving human body soundfields through acoustic primitives, 2024. URL https://arxiv.org/abs/2407.13083.

Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11293–11302, 2021a.

Leyang Li, Shilin Lu, Yan Ren, and Adams Wai-Kin Kong. Set you straight: Auto-steering denoising trajectories to sidestep unwanted concepts. *arXiv preprint arXiv:2504.12782*, 2025.

Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. AI Choreographer: Music Conditioned 3D Dance Generation with AIST++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021b.

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. *arXiv preprint arXiv:2203.05297*, 2022a.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Naoya Iwamoto, Bo Zheng, and Michael J Black. EMAGE: Towards Unified Holistic Co-Speech Gesture Generation via Masked Audio Gesture Modeling. *arXiv preprint arXiv:2401.00374*, 2023.

Haiyang Liu, Xingchao Yang, Tomoya Akiyama, Yuantian Huang, Qiaoge Li, Shigeru Kuriyama, and Takafumi Taketomi. Tango: Co-speech gesture video reenactment with hierarchical audio motion embedding and diffusion interpolation. *arXiv preprint arXiv:2410.04221*, 2024a.

Pinxin Liu, Luchuan Song, Junhua Huang, Haiyang Liu, and Chenliang Xu. GestureLSM: Latent shortcut based co-speech gesture generation with spatial-temporal modeling, 2025a. URL https://arxiv.org/abs/2501.18898.

Pinxin Liu, Pengfei Zhang, Hyeongwoo Kim, Pablo Garrido, Ari Sharpio, and Kyle Olszewski. Contextual gesture: Co-speech gesture video generation through context-aware gesture representation, 2025b. URL https://arxiv.org/abs/2502.07239.

Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning Hierarchical Cross-Modal Association for Co-Speech Gesture Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10462–10472, 2022b.

Yifei Liu, Qiong Cao, Yandong Wen, Huaiguang Jiang, and Changxing Ding. Towards variable and coordinated holistic co-speech motion generation. *arXiv preprint arXiv:2404.00368*, 2024b.

Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '13, page 25–35, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450321327. doi: 10.1145/2485895.2485900. URL https://doi.org/10.1145/2485895.2485900.

David McNeill. Hand and Mind. *Advances in Visual Semiotics*, 351, 1992.

M. Hamza Mughal, Rishabh Dabral, Merel C. J. Scholman, Vera Demberg, and Christian Theobalt. Retrieving semantics from the deep: an rag solution for gesture synthesis, 2025. URL https://arxiv.org/abs/2412.06786.

Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. From audio to photoreal embodiment: Synthesizing humans in conversations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Carolyn Saund and Stacy Marsella. *Gesture Generation*, page 213–258. Association for Computing Machinery, New York, NY, USA, 1 edition, 2021. ISBN 9781450387200. URL https://doi.org/10.1145/3477322.3477330.

Luchuan Song, Guojun Yin, Zhenchao Jin, Xiaoyi Dong, and Chenliang Xu. Emotional listener portrait: Neural listener head generation with emotion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20839–20849, 2023.

Luchuan Song, Lele Chen, Celong Liu, Pinxin Liu, and Chenliang Xu. Texttoon: Real-time text toonify head avatar from single video. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024.

Yunlong Tang, Junjia Guo, Pinxin Liu, Zhiyuan Wang, Hang Hua, Jia-Xing Zhong, Yunzhong Xiao, Chao Huang, Luchuan Song, Susan Liang, et al. Generative ai for cel-animation: A survey. *arXiv preprint arXiv:2501.06250*, 2025.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Zunnan Xu, Yachao Zhang, Sicheng Yang, Ronghui Li, and Xiu Li. Chain of generation: Multi-modal gesture synthesis via cascaded conditional control, 2023. URL https://arxiv.org/abs/2312.15900.

Zunnan Xu, Yukang Lin, Haonan Han, Sicheng Yang, Ronghui Li, Yachao Zhang, and Xiu Li. Mambatalk: Efficient holistic gesture synthesis with selective state space models, 2024. URL https://arxiv.org/abs/2403.09471.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models, 2025. URL https://arxiv.org/abs/2501.01423.

Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating Holistic 3D Human Motion from Speech. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity. *ACM Transactions on Graphics*, 39(6), 2020.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Pengfei Zhang, Pinxin Liu, Hyeongwoo Kim, Pablo Garrido, and Bindita Chaudhuri. Kinmo: Kinematic-aware human motion understanding and generation, 2024a. URL https://arxiv.org/abs/2411.15472.

Zeyi Zhang, Tenglong Ao, Yuyao Zhang, Qingzhe Gao, Chuan Lin, Baoquan Chen, and Libin Liu. Semantic gesticulator: Semantics-aware co-speech gesture synthesis, 2024b. URL https://arxiv.org/abs/2405.09814.

Yuanzhi Zhu, Ruiqing Wang, Shilin Lu, Junnan Li, Hanshu Yan, and Kai Zhang. OFTSR: One-step flow for image super-resolution with tunable fidelity-realism trade-offs. *arXiv preprint arXiv:2412.09465*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Sec. 1 explicitly enumerate our three contributions—(i) the reasoning-grounded formulation of co-speech gesture generation, (ii) the InG dataset of VLM-inferred communicative-function annotations on BEAT-2, and (iii) the Reasoning Gesture Tokenizer—and the experimental claims (state-of-the-art FGD/BC/Diversity on BEAT-2 across single-speaker, all-speaker, and zero-shot settings) are supported by the results in Sec. 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated Limitations section is included in the appendix (App. **??**), discussing dependence on VLM annotation quality, residual error accumulation under sliding-window long-sequence inference, and scope of evaluation (English speech, BEAT-2-style data).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper presents an empirical contribution (a dataset and a generative model) and does not introduce formal theorems requiring proof.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We fully describe the architecture of H-AuMoCLIP (Sec. 4.1), the Reasoning Gesture Tokenizer (Sec. 4.2), and the generator (Sec. 4.3), and provide complete training details (optimizer, learning-rate schedule, batch size, number of epochs, hardware) together with the annotation prompting protocols and baselines in the appendix. Dataset splits follow the public EMAGE/BEAT-2 protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will publicly release the InG annotations, the annotation prompting code, model training/inference code, and pretrained checkpoints upon acceptance under a permissive research license. The supplementary material includes the full annotation protocol and configuration files needed to reproduce the main results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sec. 5 reports the dataset, splits (single-speaker, all-speaker, zero-shot novel-5), and evaluation metrics (FGD/BC/Diversity), and the appendix Implementation Details specify optimizer (Adam), learning rate (1×10^{-4}), schedule, number of epochs (800), and inference setup.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report mean ± standard deviation across multiple training seeds for our method and the strongest baselines (e.g., Tab. **??**). For the user study (Tab. 2) we report Mean Opinion Scores together with mean±std and 95% confidence intervals in App. **??**.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The main paper notes that inference runs at 30.4 FPS on a single NVIDIA H100 GPU (Sec. 5.1); the appendix Implementation Details specify the GPU type and number used for training, total training epochs, and approximate wall-clock time per stage (alignment / tokenizer / generator).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics. We use a publicly available dataset (BEAT-2) under its license, do not collect new personal data, anonymize all human-study participants, and discuss societal considerations in the appendix.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Ethical Statement in the appendix discusses positive impacts (more expressive digital avatars, accessibility, embodied AI) as well as potential negatives (misuse for deepfakes/avatar-based misinformation) along with mitigation directions such as gated checkpoint release and provenance tagging.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The released gesture-generation model produces only abstract body-pose sequences (no identity, voice, or photorealistic appearance) and the InG annotations are derived from the publicly licensed BEAT-2 corpus, so the misuse surface is limited. We will gate the model release behind a research-use license requesting non-deceptive use and forbidding identity impersonation, and document this in the model card.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets are credited at first use: the BEAT-2 dataset Liu et al. [2023], baseline implementations (EMAGE, GestureLSM, CaMN, DiffSHEG, MambaTalk, RAG-Gesture, etc.), and the VLM used for annotation are cited and used under their respective licenses for research purposes.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce two new assets: the InG annotations layered on BEAT-2, and the Reasoning Gestures model checkpoints. Both will be released with a README, annotation schema documentation, dataset card, model card, training/inference scripts, and license information. Anonymized versions are provided as part of the supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

Justification: The user study procedure, the instructions presented to raters, an interface screenshot, the rating criteria (realness, synchrony, smoothness on a 1–5 Likert scale), and the participant population (20 raters, English-proficient, 8 videos each) are all described in App. **??**. Participants were recruited under standard low-risk perception-study terms with informed consent.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: The user study is a low-risk subjective video-quality rating task with no sensitive content; participants gave informed consent and were free to withdraw at any time. The protocol was reviewed under the equivalent of an IRB-exempt category at the authors' institution. Institutional details are withheld to preserve anonymity and will be added to the camera-ready version.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: VLMs/LLMs are a core component of our pipeline: they produce the structured communicative-function annotations that constitute the InG dataset and condition the generator. The exact prompting protocols (training-time with motion frames, test-time transcript-only, and unstructured baseline), model identity, and the consistency / hallucination audits are documented in the appendix annotation-protocol section.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.