DuoGesture: Motion-Grounded Semantic Conditioning and Biomechanical Beat Priors for Co-Speech Gesture Generation

Ferdinand Paar[1,2,*], Lanmiao Liu[1,3,*], Aslı Özyürek[1,2], Serge Thill[2] and Esam Ghaleb[1]
[1]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[2]Radboud University, Nijmegen, The Netherlands
[3]Utrecht University, Utrecht, The Netherlands
{ferdinand.paar, asli.ozyurek, esam.ghaleb}@mpi.nl, l.liu03@uu.nl,
serge.thill@donders.ru.nl

arXiv:2605.26236v4 [cs.CV] 31 Oct 2026

***Abstract*— Co-speech gesture generation requires both semantic expressivity and biomechanically plausible rhythmic motion. Existing holistic gesture models mix lexically grounded semantic gestures with frequent prosody-aligned beat gestures. This limits semantic grounding, speech-motion alignment, and kinematic smoothness. We propose *DuoGesture*, a neuro-inspired and biomechanically informed approach that decomposes co-speech gesture synthesis into semantic and beat streams. The two streams are coordinated by a *Semantic Variational Information Bottleneck*, a stochastic frame-level gate that learns when semantic gestures should override rhythmic beat motion. The semantic stream is controlled by *Motion-Grounded Semantic Conditioning*, which replaces purely linguistic word embeddings with motion-language representations to provide motion-aligned semantic priors for long-tailed lexical triggers of gestures. The beat stream is further regularised by an *Inertial Beat Prior*, an anthropometry-weighted arm-chain module that reduces jitter and improves rhythmic consistency without constraining semantic frames. Objective evaluations and subjective experiments show that DuoGesture outperforms strong baselines, while component ablations confirm the complementary roles of semantic grounding, stochastic stream selection, and biomechanical regularisation. We encourage readers to view our demos and qualitative illustrations, which are available on the anonymized project page: https://duogesture.github.io/DuoGesture/.**

I. Introduction

Speech and co-speech gestures form an integrated communicative system, yet gesture types play different roles and are processed accordingly. Reliable co-speech gesture generation is therefore important for embodied agents, virtual avatars, and accessible speech-driven animation systems, where gestures affect perceived naturalness, communicative clarity, and user trust. Beat gestures align with prosody and rhythm, supporting turn-taking and the flow of interaction, while semantic gestures (deictic, iconic, metaphoric) are related to lexical content and occur more sparsely [20]. Cognitive neuroscience further indicates that manual gesture comprehension recruits partially dissociable but interacting networks: parietal–premotor circuits associated with visuomotor structure, and inferior-frontal/temporal circuits implicated in communicative meaning and gesture–speech integration [25], [3]. We adopt this view as a computational stance, where co-speech gesture should be modelled as coupled streams rather than as a single motion process, where recent work has pointed out the benefit of this dual stream approach [13], [32]. However, current gesture generation still has *three limitations*: (i) misalignment between gesture and speech timing, (ii) poor semantic expressivity, and (iii) jittery beat motion that lacks biomechanical principles.

We argue that these limitations stem from two main modelling drawbacks. First, semantic conditioning is usually derived from linguistic embeddings derived from off-the-shelf text or vision-language encoders such as BERT [8] or CLIP [24], whose representations are learned from linguistic or image-text supervision rather than from body-motion dynamics. This leaves a *linguistic-motion gap*, where embeddings encode *what was said* rather than *what motion should accompany it*. This is especially problematic for long-tailed semantic triggers that appear rarely, or never, during training. Second, beat gestures are learned without physical or biomechanical constraints. These choices create gaps: a *semantic gap*, where text embeddings provide weak priors for what a gesture should look like, and a *beat gap*, where rhythmically aligned gestures may achieve reasonable average objective metrics but remain prone to over-smoothing and acceleration artefacts.

To overcome these limitations and gaps, we propose *DuoGesture*, a model shown in Figure 1, that integrates lexically grounded semantic motion from prosody-aligned beat motion while allowing them to interact. Supported by our auxiliary analysis in Supplementary Materials Section 1, which shows distinct temporal and kinematic profiles for semantic- and beat gestures, DuoGesture uses stream-specific conditioning and regularisation. The semantic stream is conditioned by *Motion-Grounded Semantic Conditioning* (MGSC), which leverages a pretrained text-to-motion representation to provide motion-aligned cues for long-tailed lexical triggers of gestures. The beat stream is regularised by an *Inertial Beat Prior* (IBP), an anthropometry-weighted arm-chain loss that encourages smooth rhythmic motion while remaining inactive on semantic frames. The streams are weighted by a *Semantic Variational Information Bottleneck* (S-VIB), which learns when semantic motion should override beat motion. Our novel contributions are summarised as follows:

*Equal contribution.
Preprint. Under review.

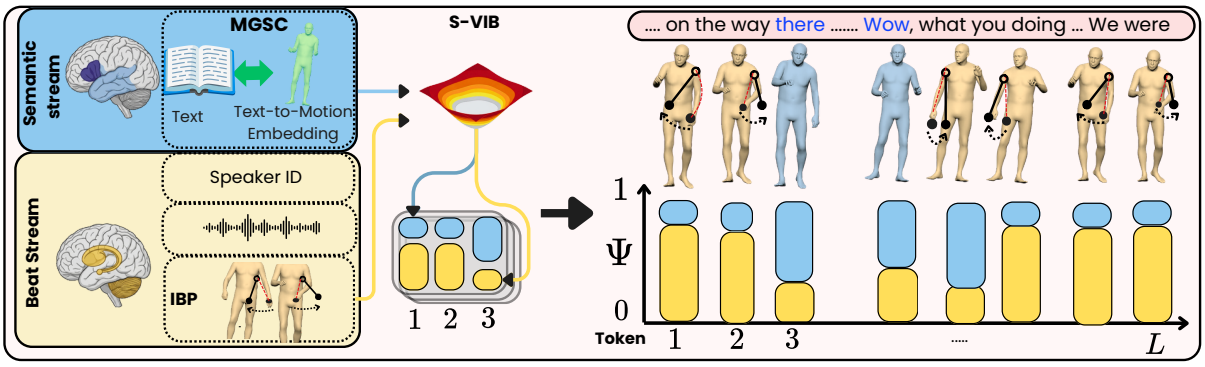


Fig. 1. DuoGesture integrates *motion-grounded semantic conditioning* and an *inertial prior* to ensure semantic expressivity and rhythmic smoothness.

- We introduce **MGSC**, a motion-grounded semantic conditioning module that, to our knowledge, is the first to drive a co-speech gesture generator from a motion-aligned text representation rather than a purely linguistic or image-text one.
- We design **IBP**, an anthropometry-weighted arm-chain regulariser that reduces beat-motion jitter and improves rhythmic consistency without constraining semantic frames.
- We present **S-VIB** that learns when semantic gestures should override beat motion using a stochastic frame-level gate.
- On the BEAT2 benchmark, DuoGesture improves distributional realism as measured by Fréchet Gesture Distance (FGD). We further report two metrics that test our contributions directly: a semantic motion score on semantically annotated frames, and a measure of motion smoothness. DuoGesture leads on all three, and a user study rates it above competitive baselines.

II. Related Work

a) Co-speech Gesture Generation.: Current work has been shaped by two-stage hierarchical quantisation paradigms. Early architectures, e.g., by TalkSHOW [28] and ProbTalk [19], established VAE-based stochastic priors and category-contingent decoding for multifaceted body dynamics. This trajectory toward granularity was extended by EMAGE [14], which proposed a spatially-decoupled tokenisation scheme across facial, manual, and corporal streams. PyraMotion [29] leveraged a multi-resolution Anchor-based Pyramid VQ-VAE to minimise distributional divergence, yielding a competitive FGD.

The aforementioned models assume **kinematic homogeneity** of co-speech gestures, treating it as a singular process. However, the difference between beat and semantic gestures is foundational in behavioural and cognitive sciences [25], [3], and researchers have attempted to exploit it. SEEG [13] encodes lexical and hand-crafted onset cues in two branches, but merges them in a single shared upper-body decoder. LivelySpeaker [34] used a sequential two-stage pipeline where its semantic-aware generator maps a whole-script CLIP embedding to motion, after which a separately trained audio-conditioned diffusion model adds noise and rhythmically refines that motion. Closest to ours, SemTalk [32] combines the streams via a deterministic frame-level gate. However, in our analysis, this gate tends to favour the dense beat pathway, and its semantic branch remains bottlenecked by shallow lexical conditioning and an absence of physical regularisation.

DuoGesture exploits the theoretical differences between beat and semantic gestures through how the two streams are conditioned, selected, fused, and regularised. Motion-Grounded Semantic Conditioning (MGSC) exploits time-aligned features from a pretrained text-to-motion space. A Semantic Variational Information Bottleneck (S-VIB) uses a stochastic gate so the two predictions are selectively combined rather than uniformly aggregated. Finally, an Inertial Beat Prior (IBP) imposes anthropometrically weighted regularisation on the proximal arm chain at beat frames only.

b) Cross-Modal Semantic Grounding Gesture Generation.: A major gap in co-speech gesture generation is the *linguistic-kinematic gap*, where word embeddings (e.g., BERT [8]) are insufficient to encapsulate the morphology of human motion. To solve this, Semantic Gesticulator [33] leverages LLM-based lexical parsing, yet the resulting representations remain inherently decoupled from the physical motion manifold. While GestureDiffuCLIP [2] employs CLIP latents to enforce semantic consistency, it operates within a text-centric latent space, often overlooking the fine-grained temporal dynamics of gesticulation. The RAG-Gesture [21] introduced non-parametric synthesis by retrieving explicit motion exemplars at inference time. Our work resolves this; instead of performing an explicit search, we *exploit* the structured knowledge of a motion-aligned encoder (TMR [23]) into our generative pipeline.

c) Physically-Consistent Motion Synthesis.: While physics-based constraints have been explored in character animation via differentiable simulators or post-hoc projections [22], [31], their integration as inductive biases within gesture models remains under-explored. In this work, we diverge from computationally expensive simulation-based approaches by incorporating *anthropometric priors* directly into the generator's rhythmic stream. Specifically, we leverage De Leva's segment mass distributions [7] to derive joint-

specific exponential moving average time constants. This formulation provides a computationally efficient, training-time-only regularisation that imposes no inference overhead. Concretely, IBP uses an auxiliary loss on arm-chain motion, with the De Leva weights setting each joint's contribution; unlike full physics simulation, it leaves the inference pipeline untouched.

III. Method

DuoGesture is a two-stage latent generator. Stage 1 (Sec. III-A) is a region-wise RVQ-VAE tokenizer. Our contributions are in Stage 2: (i) *Motion-Grounded Semantic Conditioning* (MGSC, Sec. III-B.1) grounds semantic conditioning in a text-to-motion latent space; (ii) *Inertial Beat Prior* (IBP, Sec. III-B.3.a) regularizes velocity consistency along the proximal arm chain at training time, injecting an anthropometric inductive bias into the beat stream and (iii) *Semantic Variational Information Bottleneck* (Sec. III-B.2) predicts per-frame weights Ψ under an information-bottleneck constraint.

A. Problem Formulation

A motion sequence of length L is split into four body regions $\mathcal{R} = \{\text{hand}, \text{upper}, \text{lower}, \text{face}\}$, $\mathbf{G} = \{\mathbf{G}^r\}_{r\in\mathcal{R}}$ with $\mathbf{G}^r \in \mathbb{R}^{L\times J_r}$. Stage 2 conditions on HuBERT audio features $e_a \in \mathbb{R}^{L\times 1024}$, a speaker identity embedding ID, motion-grounded semantic features $\mathbf{S}^m \in \mathbb{R}^{L\times 256}$ the per-frame output of the MGSC module (Sec. III-B.1) and a 4-frame seed pose $\tilde{\mathbf{p}}$. The target is the discrete latent code $\mathcal{Z}^q = \{\mathbf{Z}_r^q\}_{r\in\mathcal{R}}$ produced by the Stage 1 quantiser on the ground-truth motion.

a) Stage 1: Regional RVQ-VAE Tokeniser.: We adopt the regional RVQ-VAE of [14], [32], [16], [17]: one encoder–quantiser–decoder triple $(\mathcal{E}^r, \mathcal{Q}^r, \mathcal{D}^r)$ per region, trained with the standard reconstruction and codebook commitment losses. The stage-1 components are kept frozen once moving to Stage 2, decoupling (Stage 1) from temporal motion synthesis (Stage 2), and enabling the generator to operate in a structured, low-dimensional latent space.

B. DuoGesture

DuoGesture separates beat and semantic motion and adaptively fuses them via a dedicated module. The *beat backbone* f_b produces per-region latents from audio, speaker identity, and seed pose alone:

$$f_b : \big(e_a,\ \text{ID},\ \tilde{\mathbf{p}};\ \theta_b\big) \longrightarrow\ Z_r^{\text{b}},\quad r \in \mathcal{R}, \tag{1}$$

where Z_r^{b} is the beat codebook per region. This beat branch is modulated by a semantic branch via a dedicated weighting module, namely *S-VIB*. First, the *MGSC* module assembles per-frame lexical (e_s), motion-style (e_m), and emotion (e_ε) embeddings into a motion-grounded semantic feature, and *S-VIB* produces per-frame the following:

$$f_{s\text{-}vib} : \big(e_s, e_m, e_\varepsilon, e_a; \theta_{s\text{-}vib}\big) \longrightarrow \big(\mathbf{S}^m, \Psi\big). \tag{2}$$

Where $\mathbf{S}^m \in \mathbb{R}^{L\times 256}$ and $\Psi \in [0, 1]$. We then use the semantics features $\mathbf{S}^m$ to impose semantics on beat frames' output through the semantic branch:

$$f_s : \big(e_a,\ \text{ID}, \Psi,\ \mathbf{S}^m;\ \theta_s\big) \longrightarrow\ Z_r^{\text{s}},\quad r \in \mathcal{R}; \tag{3}$$

where the output of the semantic branch Z_r^{s} is then injected into Z_r^{b} via a dedicated module to produce the fused codebook $Z_r = f_{fusion}\big(Z_r^{\text{b}},\ Z_r^{\text{s}}, \Psi; \theta_{fusion}\big)$. The resulting codebook is then fed into the Stage-1 decoder $\mathcal{D}^r$, which maps the result to joint space: $\hat{\mathbf{G}}^r = \mathcal{D}^r\big(Z_r\big)$.

1) Motion-Grounded Semantic Conditioning: To bridge the gap between abstract semantic representations and their corresponding kinematic realisations, we propose the MGSC module. Existing gesture-generation methods typically use off-the-shelf text or vision-language encoders, for example FastText [5], BERT [8], or CLIP [24], whose representations are learned from linguistic or image-text supervision rather than from body-motion dynamics. In contrast, MGSC introduces a motion-grounded conditioning pathway that aligns semantic cues with their kinetic realisations. As illustrated in Fig. 2(a), MGSC produces a per-frame semantic feature $\mathbf{S}^m \in \mathbb{R}^{L\times 256}$ that conditions the semantic generator g_{sem} and feeds S-VIB. It assembles three streams, all projected to 256 dimensions: (i) $e_s \in \mathbb{R}^{256}$, a per-frame FastText [5] embedding of the word spoken at frame l, from BEAT2's forced-alignment timestamps (hence it gives word timing); (ii) e_m, an utterance-level motion-style embedding from Text-To-Motion [23]; and (iii) e_ε, an emotion embedding from Text-To-Motion, using emotion labels provided by BEAT2 [15]. e_m and e_ε are blended by a learned softmax gate into a fused memory $\mathbf{m}$:

$$\boldsymbol{\alpha} = \text{softmax}\left(\mathbf{W}_\alpha[e_m; e_\varepsilon]\right), \quad \mathbf{m} = \alpha_{(m)} e_m + \alpha_{(\varepsilon)} e_\varepsilon. \tag{4}$$

e_s then queries this memory via cross-attention to produce the final semantic feature: $\mathbf{S}^m = \text{MLP}\left(\text{CrossAttn}\left(Q = e_s,\ K = V = \mathbf{m}\right)\right) \in \mathbb{R}^{L\times 256}$.

2) Semantic Variational Information Bottleneck: We want a gate to select a semantic gesture at the right moment. For this, the gate needs to consider what and when to gesture semantically. At the same time, this bottleneck (gate) should not collapse into a single path: either all semantic or all beat. For this reason, we propose S-VIB. As illustrated in Fig. 2(b), S-VIB operates on two streams: the MGSC output $\mathbf{S}_m \in \mathbb{R}^{256}$ (*what* to gesture, derived from e_s) and a low-capacity HuBERT timing projection $\hat{e}_a \in \mathbb{R}^{64}$ (*when* to gesture). $\hat{e}_a$ is obtained by passing HuBERT features e_a through a convolutional encoder and bottlenecking to 64 dimensions; it is entirely independent of the FastText word embedding e_s.

The **Information-Theoretic Bottleneck Head** maps $\mathbf{S}^m$ and $\hat{e}_a$ to two 16-dimensional outputs via separate linear heads: a mean $\boldsymbol{\mu}_\phi \in \mathbb{R}^{16}$ and a log-variance $\log \boldsymbol{\sigma}_\phi^2 \in \mathbb{R}^{16}$. A stochastic sample is drawn via the reparametrisation trick, $\mathbf{z} = \boldsymbol{\mu}_\phi + \exp\big(\frac{1}{2}\log \boldsymbol{\sigma}_\phi^2\big) \odot \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$, and passed to the **Latent Sparsity Interpreter** κ (two-layer MLP, $\kappa : \mathbb{R}^{16} \to \mathbb{R}^2$), which outputs two-dimensional beat/semantic logits; the semantic probability is the per-frame gate: $\Psi =$

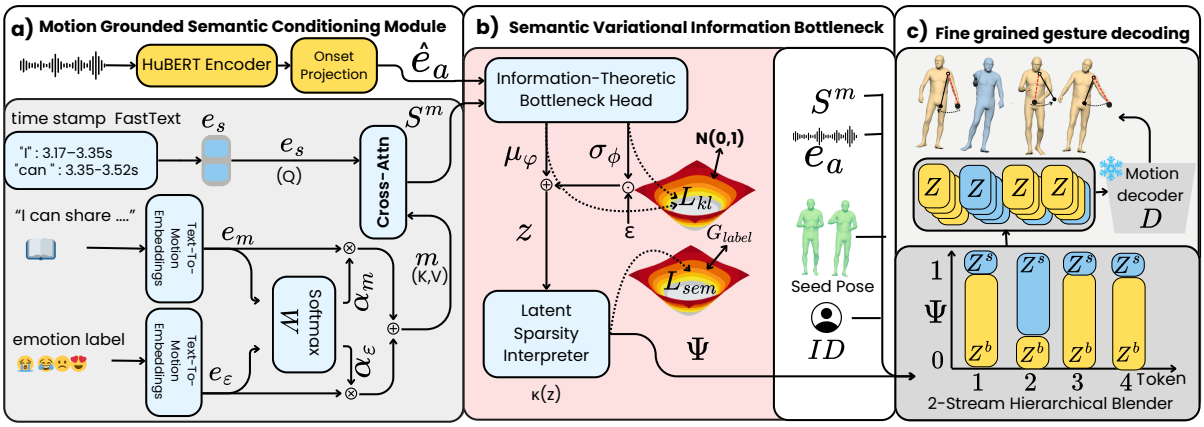


Fig. 2. DuoGesture pipeline. (a) MGSC fuses lexical semantics, motion-style, and emotion embeddings through cross-attention to produce the motion-grounded semantic representation $\mathbf{S}^m$. (b) S-VIB combines $\mathbf{S}^m$ with the HuBERT timing projection $\hat{e}_a$ to infer when semantic gestures should be activated and what semantic content they should express. The bottleneck samples $\mathbf{z}\tau \sim q\phi(\mathbf{z}\tau)$ under $\mathcal{L}$kl regularisation, and maps it to the semantic gate Ψ with semantic supervision. (c) Fine-grained decoding blends beat codebooks Z_r^{b} (yellow) and semantic codebooks Z_r^{s} (blue) using Ψ.

$\mathrm{softmax}\big(\kappa(\mathbf{z})\big)_{\mathrm{sem}} \in [0,1]$. The outputs $\boldsymbol{\mu}_\phi$ and $\log \boldsymbol{\sigma}_\phi^2$ are regularised against the standard Gaussian prior $\mathcal{N}(0, I)$ via a KL divergence [1], [9]:

$$\begin{aligned}\mathcal{L}_{\mathrm{kl}} &= D_{\mathrm{KL}}\big(\mathcal{N}(\boldsymbol{\mu}_\phi, \boldsymbol{\sigma}_\phi^2) \,\|\, \mathcal{N}(0, I)\big) \\ &= \frac{1}{2}\sum_{d=1}^{Z}\Big[\mu_{\phi,d}^2 + e^{\log \sigma_{\phi,d}^2} - \log \sigma_{\phi,d}^2 - 1\Big].\end{aligned} \tag{5}$$

In practice, we apply a per-dimension free-bits floor [9] ($\lambda_{\mathrm{fb}} = 0.50$ nats, $Z = 16$), whose per-dim KL falls below λ_{fb} receive no gradient, preventing the optimiser from over-regularising uninformative dimensions while preserving capacity for those that carry semantic signal. The gate Ψ is trained to predict the per-frame semantic annotations $s_\tau \in \{0, 1\}$ via a semantic loss $\mathcal{L}_{sem}$.

3) Hierarchical Blender: The hierarchical blender (Fig. 3) produces per-region beat residuals Z_r^{b} and semantic residuals Z_r^{s}, decoded by the frozen Stage-1 decoders. It has two parallel branches that differ in their conditioning and regularisation, reflecting the asymmetry between beat and semantic gesture production. **Beat stream**, (top, yellow in Fig. 3) takes as input the masked seed pose $\tilde{\mathbf{p}}$ embedding $\in \mathbb{R}^{T\times 768}$, where we add speaker identity ID and periodic positional encodings are applied before a self-attention pass. Speech onset projection ($\hat{e}_a$) is integrated with the resulting embeddings through a cross-attention layer. Three region-specific MLPs then project the refined beat region latents Z_u^{b}, Z_l^{b}, Z_h^{b}. The face latent Z_e^{b} is obtained from a separate face decoder conditioned on the HuBERT audio feature e_a.

a) Inertial Beat Prior (IBP).: On beat stream, we apply our proposal to constrain the upper-body and hand movements with biomechanical constraints. For this reason, the beat latents Z_u^{b}, Z_l^{b}, Z_h^{b} are decoded by the frozen Stage-1 decoder into raw (rot6d) poses $\mathbf{x}_{j,t}$ and an **Inertia Module** (fed by body-part masses m_j, gate $(1{-}\Psi)$, and VIB posterior variance σ_ϕ^2) produces per-joint weights τ_j that scale $\mathcal{L}_{acc}$. IBP is a training-time regulariser that teaches the beat-stream network θ_b to produce latents that decode into physically smooth poses.

IBP penalises deviation from constant-velocity motion. The frozen decoder maps the predicted tokens to 6D joint rotations $\mathbf{x}_{j,t}$. The loss compares the current frame with the speed from the last two frames. The loss is the mean squared error against the constant-velocity prediction $\hat{\mathbf{x}}_{j,t} = 2\mathbf{x}_{j,t-1} - \mathbf{x}_{j,t-2}$, weighted by $\tau_{j,t}$: $\mathcal{L}_{acc} = \mathbb{E}_{j,t}\Big[\tau_{j,t}\,\big\|\mathbf{x}_{j,t} - \hat{\mathbf{x}}_{j,t}\big\|^2\Big]$. Rather than penalising all joints equally, IBP weights each joint by a smoothing coefficient derived from De Leva's anthropometric masses [7] and the S-VIB gate:

$$\tau_{j,t} = \tau_{\mathrm{base}} \cdot \sqrt{\frac{m_j}{m_{\max}}} \cdot (1 - \Psi_t) \cdot (1 + \alpha\, \sigma_{\phi,t}^2), \tag{6}$$

where m_j is the De Leva body-segment mass fraction for joint j, $m_{\max}$ is the fraction of the heaviest segment (spine1/abdomen, $\approx$0.163 of body mass), so $\sqrt{m_j/m_{\max}} \in (0, 1]$ is a sqrt-compressed relative inertial weight; $\tau_{\mathrm{base}} = 0.5$ is the maximum smoothing applied to the heaviest arm-chain joint on a pure beat frame. Hence, a uniform IBP gives each joint the same weight. De Leva gives more weight to heavier body parts and less to the hands. The gate lowers the loss on meaning frames.

Hence, training with this regulariser encourages smoother latents without any explicit physics at inference time. The face latent Z_e^{b} does not pass through the IBP block since facial motion has distinct dynamics that should not be constrained by IBP.

Semantic stream (bottom, blue in Fig. 3) operates on MGSC output $\mathbf{S}^m$. We first add masked seed pose embeddings ($\tilde{\mathbf{p}}$) and ID, along with periodic positional encodings, before a self-attention pass. The resulting embeddings are then integrated with $\mathbf{S}^m$ through cross-attention. Then three region-specific MLPs project these into semantic region latents Z_u^{s}, Z_h^{s}, Z_l^{s}.

Inspired by [32], a *Hierarchical Cross-Attention (HCA)* block refines these latents by allowing each region to attend

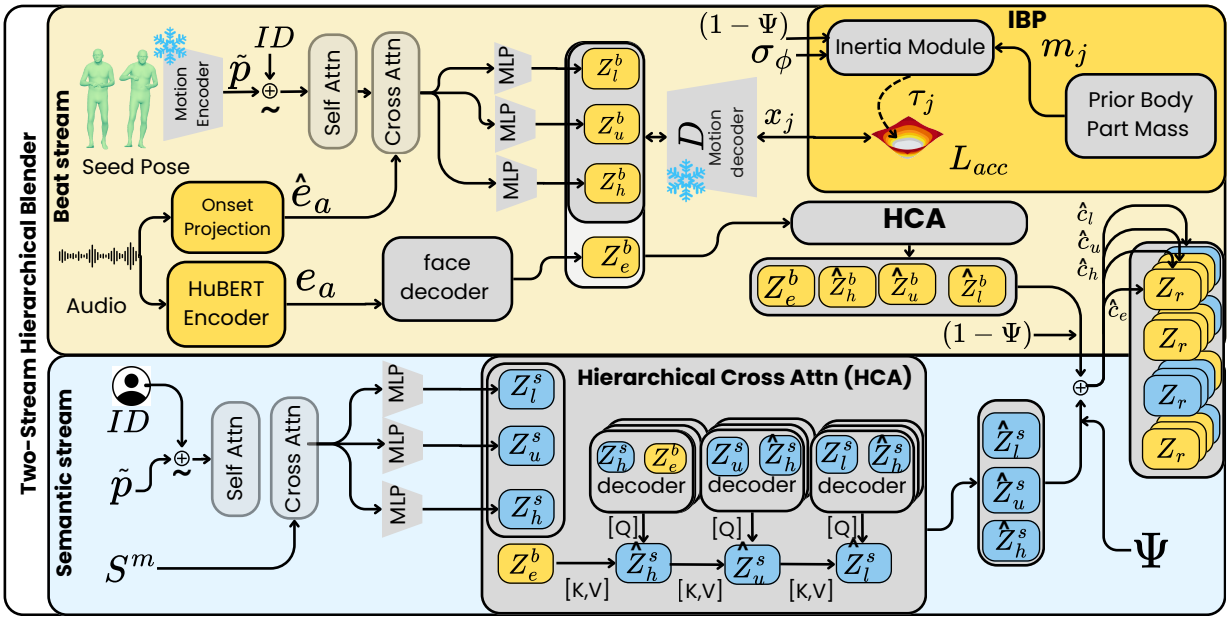


Fig. 3. Two-Stream Hierarchical Blender. The beat stream encodes the seed pose $\tilde{\mathbf{p}}$ and speaker ID to predict region-wise beat latents, with the face handled by a separate audio-conditioned decoder. Arm-chain beat latents are decoded during training and regularised by IBP through a τ_j-weighted smoothness loss. The semantic stream conditions parallel region-wise latents on the gated semantic feature. Hierarchical cross-attention refines both streams across body regions.

to its sibling semantic latents: the hands decoder attends to $Z_u^{\mathrm{s}} + Z_l^{\mathrm{s}}$ as [K, V], the upper decoder attends to $Z_h^{\mathrm{s}} + Z_l^{\mathrm{s}}$, and the lower decoder attends to $Z_u^{\mathrm{s}} + Z_h^{\mathrm{s}}$, producing final semantic latents $\hat{Z}_u^{\mathrm{s}}$, $\hat{Z}_h^{\mathrm{s}}$, $\hat{Z}_l^{\mathrm{s}}$. HCA is applied similarly on the beat codebooks. Finally, as illustrated in Fig. 2(c), for each of the three body regions $r \in \{h, u, l\}$, the beat residual Z_r^{b} (yellow token, $\Psi \approx 0$) and the semantic residual Z_r^{s} (blue token, $\Psi \approx 1$) are blended frame-by-frame via a fusion function. The fused and face latents are then quantised by nearest-neighbour lookup in the RVQ codebook:

$$Z_r = (1 - \Psi)\hat{Z}_r^{\mathrm{b}} + \Psi \hat{Z}_r^{\mathrm{s}},$$
$$\hat{c}_{r,t} = \arg \min_{k \in \{1,\ldots,K\}} \|e_k - Z_{r,t}\|_2^2 .$$

where e_k denotes the k-th codebook vector and $\hat{c}_{r,t}$ is the selected discrete token for region r at frame t. The selected tokens are decoded by the frozen Stage-1 decoder as $\hat{G}_r = D_r(\hat{c}_r)$. Finally, our **full training objective** is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{lat}} + \mathcal{L}_{\mathrm{cls}} + \mathcal{L}_{\mathrm{sem}} + \beta_{\mathrm{vib}}\,\mathcal{L}_{\mathrm{kl}} + \beta_{\mathrm{phys}}\,\mathcal{L}_{\mathrm{acc}}, \qquad (7)$$

where: $\mathcal{L}_{\mathrm{lat}}$ is the MSE between the predicted continuous latents and the Stage-1 VQ targets for all body regions; $\mathcal{L}_{\mathrm{cls}}$ is the cross-entropy over the four RVQ codebook levels for each region; $\mathcal{L}_{\mathrm{sem}}$ (Sec. III-B.2) is the S-VIB gate supervised against BEAT2 semantic flags; $\mathcal{L}_{\mathrm{kl}}$ (Sec. III-B.2) is the VIB bottleneck KL with a free-bits floor; and $\mathcal{L}_{\mathrm{acc}}$ (Sec. III-B.3.a) is the IBP inertia residual, active only on beat frames via $\tau_{j,t}$. The scalars β_{vib} and β_{phys} are warmup-scheduled weights.

IV. Experiments

a) Dataset.: We evaluate DuoGesture on BEAT2 [14], a standard benchmark for co-speech gesture generation. It contains approximately 76 hours of speech, motion, facial expression, and speaker identity annotations from 30 speakers, covering diverse expressive behaviours. Unlike other datasets, such as TalkSHOW [28], BEAT2 provides frame-level annotations of gesture type (including semantic and beat gestures) and eight emotion categories, making it suitable for our study. Following the standard BEAT2 protocol used in prior holistic gesture-generation work [14], [21], we report results in two settings: *(i) a single-speaker setting on Speaker 2* and *(ii) a multi-speaker setting over 25 speakers*. Unless otherwise stated, the data are split into training, validation, and test partitions at 85%/7.5%/7.5%. The **implementation details** are elaborated in Sections 2 and 3 of the Supplementary materials.

b) Evaluation Metrics.: We evaluate co-speech gestures along several dimensions: distributional realism, speech–motion synchrony, motion variation, and facial stability. We treat user studies and Fréchet Gesture Distance (FGD) [30] as the primary evaluations, since FGD compares generated and real motion distributions in a learned gesture-feature space, and is the standard metric with the strongest reported perceptual support, where FGD was the only objective metric found to correlate with subjective human-likeness ratings [10].

We report two additional metrics alongside FGD since they test the contributions we make. **Semantic-Relevant Gesture Recall** (SRGR) [15] weights gesture recall of annotated semantic gestures. It scores motion only where a gesture is expected to be semantic, which is what MGSC (Sec. III-B.1) targets. **Relative log dimensionless jerk** (ΔLDLJ) [4] measures smoothness as a signed difference from ground truth, $\Delta\mathrm{LDLJ} = \mathrm{LDLJ}_{\mathrm{gen}} - \mathrm{LDLJ}_{\mathrm{GT}}$, where positive values

TABLE I

COMPARISON ON BEAT2 (SINGLE SPEAKER, TOP; ALL SPEAKERS, BOTTOM). UNDER THE SAME PROTOCOL, ROWS MARKED * ARE TAKEN FROM PYRAMOTION [29] AND ALL OTHER ROWS ARE OUR OWN RUNS. SRGR AND RELATIVE LDLJ ARE THEREFORE UNAVAILABLE FOR * ROWS. BEST, SECOND, AND THIRD ARE MARKED PER COLUMN WITHIN EACH BLOCK.

Setting	Model	FGD $\times 10^{-1} \downarrow$	BA $\times 10^{-1} \uparrow$	Diversity $\uparrow$	MSE $\times 10^{-3} \downarrow$	LVD $\times 10^{-5} \downarrow$	SRGR $\uparrow$	Rel. LDLJ $\rightarrow 0$
One Speaker	TalkSHOW [28] (CVPR 2023)	6.209	6.947	13.47	7.791	7.771	0.248	1.6520
	AMUSE* [6] (CVPR 2024)	12.11	**8.318**	**14.93**	–	–	–	–
	ProbTalk* [19] (CVPR 2024)	5.040	7.711	13.27	8.617	–	–	–
	EMAGE [14] (CVPR 2024)	5.512	7.724	13.06	7.680	7.556	0.314	0.0470
	MambaTalk* [27] (NeurIPS 2024)	5.366	7.812	13.95	**6.289**	**6.897**	–	–
	RAG-Gesture* [21] (CVPR 2025)	8.080	7.340	11.97	–	–	–	–
	GestureLSM [18] (ICCV 2025)	4.247	7.290	13.76	7.139	7.469	0.320	1.5138
	SemTalk [32] (ICCV 2025)	4.278	7.770	12.91	7.153	6.938	0.315	-0.5910
	PyraMotion* [29] (NeurIPS 2025)	4.612	7.420	13.42	7.176	7.270	–	–
	Ours (DuoGesture)	**4.101**	7.557	12.34	7.103	7.646	**0.343**	**-0.0304**
All Speakers	TalkSHOW [28] (CVPR 2023)	6.145	6.863	**13.12**	7.953	7.824	0.257	1.7360
	EMAGE [14] (CVPR 2024)	5.643	**7.707**	12.92	7.694	7.593	0.326	0.0530
	GestureLSM [18] (ICCV 2025)	4.268	5.250	11.20	7.582	7.685	0.305	1.5386
	SemTalk [32] (ICCV 2025)	5.214	7.689	12.74	7.612	**7.498**	0.302	-0.6020
	Ours (DuoGesture)	**4.081**	7.699	12.83	**7.502**	7.658	**0.348**	**-0.0390**

indicate over-smoothing and negative values indicate excess jitter. Relative LDLJ tests the effectiveness of IBP (Sec. III-B.3.a). In addition, we include Beat Alignment (BA) [12], pairwise L1 Diversity [11], facial MSE, and L1 Vertex Difference (LVD) [26] as secondary diagnostic metrics. These metrics are informative but not sufficient indicators of perceptual quality. High BA can reflect exaggerated beat-like motion, and high Diversity can reflect large motion variance rather than plausible or semantically appropriate gestures.

c) Comparison Methods.: We benchmark DuoGesture against a broad set of representative holistic co-speech gesture generators, covering early speech-to-gesture models, hierarchical and diffusion-based methods, semantic-based approaches, and recent BEAT2 state-of-the-art systems. The comparison includes RAG-Gesture [21], ProbTalk [19], MambaTalk [27], TalkSHOW [28], EMAGE [14], SemTalk [32], GestureLSM [18], and PyraMotion [29].

A. Overall Quantitative Comparison

Table I shows the BEAT2 comparison with state-of-the-art methods. DuoGesture achieves the lowest FGD in both the single-speaker and all-speaker settings (4.101 and 4.081, respectively), outperforming GestureLSM and SemTalk. It also leads on both contribution-specific metrics: SRGR and $|\Delta\text{LDLJ}|$. This shows that gestures are semantically better recalled and closer to the dynamics of real motion smoothness than the presented baseline. On the secondary diagnostics, DuoGesture remains competitive. For BA and Diversity, it sits within a small margin of the best baselines. It also achieves the best MSE in the all-speaker setting. However, methods with higher BA or Diversity, such as AMUSE, achieve this with a worse FGD. Since FGD is the objective metric most consistently linked to perceived human-likeness [10], we interpret this as a favourable overall trade-off.

TABLE II

MGSC INPUT ABLATION (ALL-SPEAKER SETTINGS).

MGSC condition	FGD $\downarrow$	SRGR $\uparrow$
Full MGSC (TMR + emotion)	**4.081**	**0.348**
Without emotion	4.110	0.335
CLIP instead of TMR	4.436	0.318
CLIP, without emotion	4.332	0.305

B. Ablation Studies

a) Motion-grounded semantic conditioning.: Table II tests how MGSC improves semantic gesture recall by varying only its inputs. Replacing TMR [23] with CLIP [24] substantially lowers performance, raising FGD from 4.081 to 4.436 and lowering SRGR from 0.348 to 0.318, whereas removing the emotion embedding has a minimal impact (4.110, 0.335). Performance in terms of SRGR degrades monotonically as motion-grounded information is removed. Interestingly, FGD improves when CLIP is used without emotion embedding. In general, motion grounding of the representation is the source of the performance gain.

TABLE III

IBP ABLATION (ALL-SPEAKER SETTINGS).

IBP variant	FGD $\downarrow$	BA $\uparrow$	Diversity $\uparrow$	ΔLDLJ $\rightarrow 0$
No IBP	4.178	7.446	12.77	-0.4904
Uniform IBP	4.278	7.450	12.80	1.2790
De Leva IBP	**4.081**	**7.699**	**12.83**	$\mathbf{-0.0390}$

b) Inertial beat prior.: The benefit of IBP could come from smoothing alone or specifically from the anthropometric weighting in Eq. 6. Table III separates the two by comparing no prior, a uniform variant applying the same τ to every arm-chain joint, and the full De Leva weighting. Uniform smoothing is less effective, raising FGD from 4.178 to 4.278, worse than applying no prior at all. Most importantly, without

IBP, ΔLDLJ is (-0.490). Uniform smoothing results in over-smoothing with ΔLDLJ of ($+1.279$). However, the anthropometric weighting gives the best results, near ground-truth smooth motion (-0.039), while also producing the largest beat-alignment gain (7.699).

TABLE IV
COMPONENT-WISE ABLATION OF DUOGESTURE (ALL-SPEAKERS).

Variant	MGSC	S-VIB	IBP	FGD ↓	BA ↑	Diversity ↑
(a) w/o MGSC (S-VIB + IBP only)	–	✓	✓	4.803	7.531	12.61
(b) MGSC only (linear σ-gate)	✓	–	–	4.306	7.551	12.52
(c) MGSC + S-VIB (no IBP)	✓	✓	–	4.178	7.446	12.77
(d) MGSC + IBP (linear σ-gate)	✓	–	✓	4.137	7.557	12.65
(e) Full DuoGesture	✓	✓	✓	**4.081**	**7.699**	**12.83**

c) Component-wise ablation.: Table IV checks the contribution of each main module in the all-speaker setting: each row removes or replaces one component while holding the others fixed. MGSC dominates the FGD gain. This involves removing all three MGSC components: emotion embeddings and timing from FastText — and replacing motion-grounded semantic representations with standard semantic representations. This raises the FGD from 4.081 to 4.803. The effect of this is larger than that of removing S-VIB (+0.056) or IBP (+0.097). This is consistent with our hypothesis that motion-grounded semantic conditioning closes the linguistic–kinematic gap left open by purely lexical embeddings.

IBP improves beat alignment, conditional on the stochastic gate. With S-VIB, adding IBP raises BA from 7.446 to 7.699, the largest BA change in the table. Under the deterministic gate, BA changes minimally (7.551 to 7.557). The two components are therefore complementary, mirroring Table III. Replacing S-VIB with the deterministic σ gate (variants b and d) lowers diversity to 12.52 and 12.65 versus 12.83 for the full model.

Hence, the ablation supports the design claim that semantic grounding, stochastic stream selection, and biomechanical beat regularisation are useful individually and most effective when combined.

C. Subjective and Qualitative Results

1) User study.: We conduct a perceptual study using 35-second clips from the BEAT2 test set. The study includes six narrated topics and compares Ground Truth and representative state-of-the-art methods, namely, EMAGE, SemTalk, and DuoGesture. Thirty native English-speaking participants from the UK, balanced by self-reported gender (male: female = 1:1; mean age 38.6), evaluated 24 randomly ordered videos using a five-point Likert scale. Participants rated each video along three axes: naturalness, motion diversity, and alignment with speech content and timing. The order of methods and clips was randomised for each participant.

2) User Studies Interface and Details: Figure 4 presents the evaluation interface used in our user study, which was developed in Qualtrics. During the evaluation, participants were required to first watch a co-speech gesture animation video and subsequently assess the quality of the generated gestures using a five-point Likert scale. The evaluation focused on three aspects: *Naturalness*, *Diversity*, and *Alignment with Speech Content and Timing*. In total, the questionnaire consisted of 24 video samples for evaluation. To minimise potential ordering effects and subjective bias, all videos were displayed in a randomised order for each participant. Participant recruitment and study administration were conducted through the Prolific platform, targeting participants from English-speaking countries.

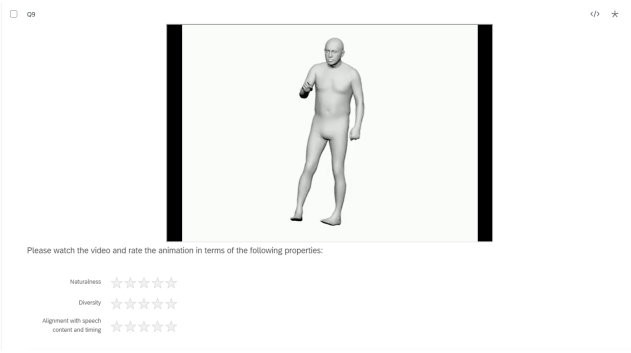


Fig. 4. User study interface developed using Qualtrics. Participants were instructed to watch a co-speech gesture animation video and to evaluate the generated motions using a five-point Likert scale across three aspects: naturalness, diversity, and alignment with speech content and timing.

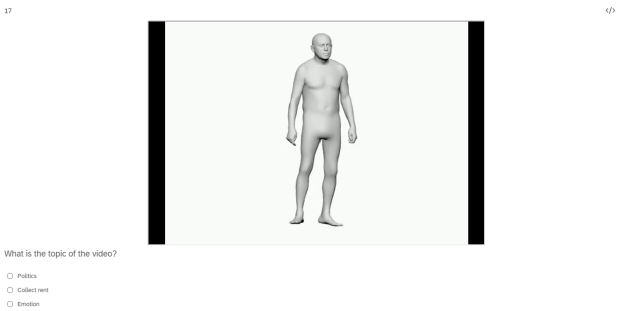


Fig. 5. Attention-check interface employed during the user study to verify participant engagement and comprehension of the narrated content.

To further ensure the reliability and validity of the collected responses, we incorporated additional attention-check questions throughout the evaluation process, as illustrated in Figure 5. In these questions, participants were asked to identify the topic of the narrated speech from several candidate options. Responses from participants who failed the attention-check questions were excluded from the final analysis. These quality-control measures were introduced to verify that participants carefully watched the videos and adequately understood the accompanying speech content during the evaluation.

3) Results.: As shown in Fig. 7, Ground Truth receives the highest scores, as expected. Among the generated visualisations, DuoGesture obtains the strongest overall perceptual ratings, with higher judgments than SemTalk and EMAGE across the evaluated axes. This result is consistent with the quantitative analysis: DuoGesture gives a better perceptual trade-off between realism, synchrony, and expressiveness. The user study, therefore, provides independent evidence that

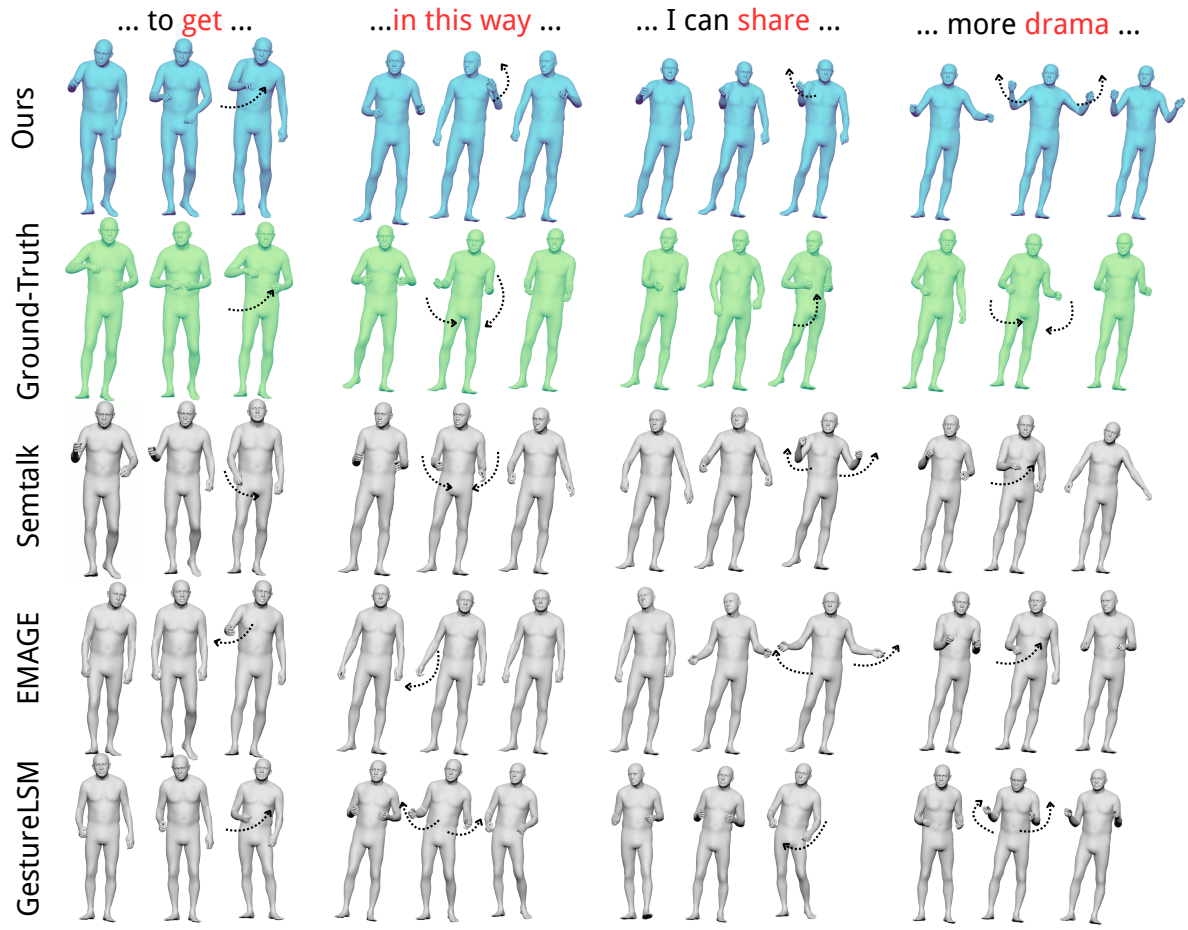


Fig. 6. Qualitative comparison of co-speech gesture generation across semantic and beat-dominant speech contexts. We visualise motion sequences conditioned on representative phrases, including “to get”, “in this way”, “I can share”, and “more drama”.

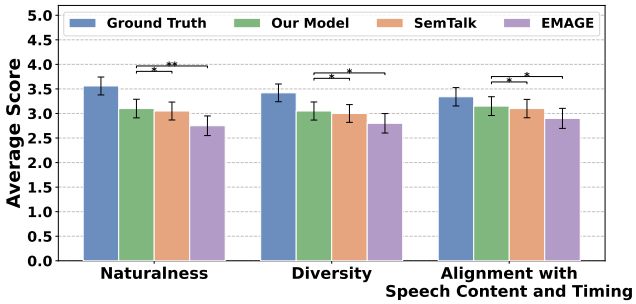


Fig. 7. User study results comparing Ground Truth, DuoGesture, SemTalk, and EMAGE. Stars show significant differences.

the observed results in Table I correspond to motions that users judge as more natural and better aligned with speech.

D. Qualitative Results

a) Qualitative comparison.: Figure 6 illustrates a comparison of the generated motion sequences for speech segments containing semantic and beat phrases. For semantic phrases such as ’to get’, ’I can share’, and ’more drama’, DuoGesture produces gestures with clearer arm trajectories and a more visible, phrase-dependent structure than the other methods. SemTalk and GestureLSM tend to generate weaker or more ambiguous movements, whereas EMAGE often produces more plausible, albeit less semantically differentiated, gestures. For the beat-dominant phrase ’in this way’, the generated motion remains temporally coherent with the speech rhythm, yet avoids excessive smoothing.

V. Conclusion

We presented DuoGesture, a co-speech gesture generation approach that uses a semantic stream to generate lexically grounded gestures and a beat stream to generate rhythmic motion aligned to prosody. A stochastic frame-level gate coordinates the interaction between these streams. DuoGesture achieves the lowest FGD on BEAT2 in both the single-speaker and all-speaker settings by combining motion-grounded semantic conditioning, stochastic stream selection, and an inertial beat prior, while remaining competitive on alignment, diversity, and facial-motion fidelity. It also outperforms other approaches on SRGR and $|\Delta\text{LDLJ}|$ in both settings, the two metrics that test our two main contributions: MGSC and IBP. Ablation studies and user testing confirm that these components provide complementary benefits, sup-

porting the idea of an explicit design process for co-speech gesture generation.

REFERENCES

[1] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

[2] T. Ao, Z. Zhang, and L. Liu. Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Transactions on Graphics (TOG)*, 42(4):1–18, 2023.

[3] M. A. Arbib. *How the brain got language: The mirror system hypothesis*, volume 16. Oxford University Press, 2012.

[4] S. Balasubramanian, A. Melendez-Calderon, A. Roby-Brami, and E. Burdet. On the analysis of movement smoothness. *Journal of neuroengineering and rehabilitation*, 12(1):112, 2015.

[5] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[6] K. Chhatre, R. Danecek, N. Athanasiou, G. Becherini, C. Peters, M. J. Black, and T. Bolkart. Emotional speech-driven 3d body animation via disentangled latent diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1942–1953, 2024.

[7] P. De Leva. Adjustments to zatsiorsky-seluyanov's segment inertia parameters. *Journal of biomechanics*, 29(9):1223–1230, 1996.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[9] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29, 2016.

[10] T. Kucherenko*, P. Wolfert*, Y. Yoon*, C. Viegas, T. Nikolov, M. Tsakov, and G. E. Henter. Evaluating gesture generation in a large-scale open challenge: The genea challenge 2022. *ACM Transactions on Graphics*, 43(3):1–28, 2024.

[11] J. Li, D. Kang, W. Pei, X. Zhe, Y. Zhang, Z. He, and L. 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, 2021.

[12] R. Li, S. Yang, D. A. Ross, and A. Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13401–13412, 2021.

[13] Y. Liang, Q. Feng, L. Zhu, L. Hu, P. Pan, and Y. Yang. Seeg: Semantic energized co-speech gesture generation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10463–10472. IEEE, 2022.

[14] H. Liu, Z. Zhu, G. Becherini, Y. Peng, M. Su, Y. Zhou, X. Zhe, N. Iwamoto, B. Zheng, and M. J. Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1144–1154, 2024.

[15] H. Liu, Z. Zhu, N. Iwamoto, Y. Peng, Z. Li, Y. Zhou, E. Bozkurt, and B. Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European conference on computer vision*, pages 612–630. Springer, 2022.

[16] L. Liu, E. Ghaleb, A. Özyurek, and Z. Yumak. Semges: Semantics-aware co-speech gesture generation using semantic coherence and relevance learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13963–13973, 2025.

[17] L. Liu, E. Ghaleb, A. Özyürek, and Z. Yumak. Semconflow: Semantic grounding of holistic co-speech gesture generation with contrastive flow-matching. *arXiv preprint arXiv:2603.26553*, 2026.

[18] P. Liu, L. Song, J. Huang, H. Liu, and C. Xu. Gesturelsm: Latent shortcut based co-speech gesture generation with spatial-temporal modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10929–10939, 2025.

[19] Y. Liu, Q. Cao, Y. Wen, H. Jiang, and C. Ding. Towards variable and coordinated holistic co-speech motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1566–1576, 2024.

[20] D. McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.

[21] H. Mughal, R. Dabral, M. C. Scholman, V. Demberg, and C. Theobalt. Retrieving semantics from the deep: an rag solution for gesture synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16578–16588, 2025.

[22] X. B. Peng, P. Abbeel, S. Levine, and M. Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018.

[23] M. Petrovich, M. J. Black, and G. Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9488–9497, 2023.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, 2021.

[25] C. M. Tipper, G. Signorini, and S. T. Grafton. Body language in the brain: constructing meaning from expressive movement. *Frontiers in Human Neuroscience*, Volume 9 - 2015, 2015.

[26] J. Xing, M. Xia, Y. Zhang, X. Cun, J. Wang, and T.-T. Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790, 2023.

[27] Z. Xu, Y. Lin, H. Han, S. Yang, R. Li, Y. Zhang, and X. Li. Mambatalk: Efficient holistic gesture synthesis with selective state space models. *Advances in Neural Information Processing Systems*, 37:20055–20080, 2024.

[28] H. Yi, H. Liang, Y. Liu, Q. Cao, Y. Wen, T. Bolkart, D. Tao, and M. J. Black. Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 469–480, 2023.

[29] Z. Yin, Y. H. Tsui, and P. Hui. Pyramotion: Attentional pyramid-structured motion integration for co-speech 3d gesture synthesis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[30] Y. Yoon, B. Cha, J.-H. Lee, M. Jang, J. Lee, J. Kim, and G. Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020.

[31] Y. Yuan, J. Song, U. Iqbal, A. Vahdat, and J. Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16010–16021, 2023.

[32] X. Zhang, J. Li, J. Zhang, Z. Dang, J. Ren, L. Bo, and Z. Tu. Semtalk: Holistic co-speech motion generation with frame-level semantic emphasis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13761–13771, 2025.

[33] Z. Zhang, T. Ao, Y. Zhang, Q. Gao, C. Lin, B. Chen, and L. Liu. Semantic gesticulator: Semantics-aware co-speech gesture synthesis. *ACM Transactions on Graphics (TOG)*, 43(4):1–17, 2024.

[34] Y. Zhi, X. Cun, X. Chen, X. Shen, W. Guo, S. Huang, and S. Gao. Livelyspeaker: Towards semantic-aware co-speech gesture generation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20750–20760. IEEE, 2023.

Supplementary Material

S1. Motion Analysis: Beat vs. Semantic Motion on BEAT2

a) Setup and controlled sampling.: We analyse BEAT2 [15], [14] test-split motion ($\geq$ 15-frame windows, all 25 speakers) using the dataset's per-frame semantic annotation flags to segment sequences into contiguous *beat windows* (flag $= 0$) and *semantic windows* (flag $= 1$). To ensure a fair, balanced comparison, we draw a matched sample: for each speaker we randomly draw the same number of beat and semantic windows (capped at the smaller set), yielding $\approx 1{,}870$ windows per class from 265 clips. All metrics are duration-weighted and accompanied by 95% CIs from hierarchical bootstrap (speaker $\rightarrow$ clip $\rightarrow$ window). The matched windows are representative of typical gesture production for each class (short rhythmic repetitions for beat, and clear lexical stroke events for semantic), and the class balance removes any confound from the natural 12% semantic sparsity.

b) Results.: The two classes show clearly distinct kinematics (Table S1 and Fig. S1). Beat motion peaks at 1.12 Hz with moderate inter-joint coupling (PLV $= 0.31$), indicating that shoulder and forearm oscillate at partially independent rhythms, consistent with a damped-pendulum model. A constant-velocity (inertial) oscillator explains $R^2 = 0.41$ of beat variance, and this improves by $\Delta R^2 = +0.84$ when the mass grouping covers arm + core joints, but *degrades* by $\Delta R^2 = -0.95$ when all 55 joints are included, directly motivating IBP's arm-chain-only mask. Semantic motion peaks at 1.69 Hz with strongly higher inter-joint coupling (PLV $= 0.53$, ΔPLV $= +0.22$), indicating holistic co-activation of the arm during a lexical stroke, a pattern better captured by shape-conditioned generation than by a mass-weighted smoother. Arm-swing PSD (Fig. S1) further shows that beat spectra have a narrow, peaked structure (half-BW 0.46 Hz, prominence $4.6\times$) while semantic spectra are $1.9\times$ broader and less tonal, confirming that beat motion is rhythmically regular and semantic motion is not.

S2. Technical Implementation

a) Data.: BEAT2 [14], 85%/7.5%/7.5% split, 64-frame clips at 30 fps (stride 20), 4-frame seed pose.

b) Training.: Adam (lr $= 10^{-4}$, no weight decay), step LR decay ($\gamma = 0.3$), 200 epochs. Stage 2 trained on 4 GPUs A100 via PyTorch DDP, effective batch size 256.

c) Representations.: Motion: 55 SMPL-X joints in rot6d (330-dim). Audio: HuBERT (1024-dim). Text: FastText [5] (300-dim). All streams projected to 256-dim.

d) Stage 1.: Regional RVQ-VAE [14], codebook $C = 256$, 4 RVQ levels, frozen during Stage 2.

e) Stage 2.: Transformer backbone: hidden size 768, 1 self-attention layer; 3 decoder layers and 1 codebook head per region. MGSC embeddings (e_m, e_ε) from TMR [23] (256-dim). S-VIB: bottleneck $Z = 16$, timing projection 64-dim, KL warmup epochs 20–100 ($\beta_{\text{target}} = 0.01$), free-bits $\lambda_{\text{fb}} = 0.5$ nats, semantic boost $\rho = 3.0$. IBP: physics warmup epochs 30–80 ($\lambda = 0.01$), $\tau_{\text{base}} = 0.5$, $\alpha = 1.0$, arm-chain mask only.

S3. Architecture Details

Table S2 summarises the key dimensions of each Stage-2 component.

Fusion. For $r \in \{h, u, l\}$: $Z_r = (1-\Psi)\hat{Z}^b_r + \Psi\hat{Z}^s_r$, quantised as $\hat{c}_r = \arg\min_k \|e_k - Z_r\|_2$. The face region uses Z^b_e directly (no semantic stream). All regions decoded by frozen $\mathcal{D}^r$: $\hat{\mathbf{G}}^r = \mathcal{D}^r(\hat{c}_r)$.

TABLE S1

KINEMATIC STATISTICS FOR MATCHED BEAT AND SEMANTIC WINDOWS (BEAT2 TEST SPLIT, 25 SPEAKERS; HIERARCHICAL BOOTSTRAP 95% CI IN BRACKETS).

Metric	Beat	Semantic
Shoulder peak (Hz)	1.12 [1.02, 1.23]	1.69 [1.56, 1.86]
PLV (shoulder ↔ forearm)	0.308 [0.268, 0.349]	0.525 [0.487, 0.557]
Oscillator R^2	0.406 [0.388, 0.427]	0.509 [0.493, 0.524]
ΔR^2: arm+core vs. shoulder	+0.84	—
ΔR^2: all joints vs. shoulder	−0.95	—
Arm-swing half-BW (Hz)	0.46	0.89

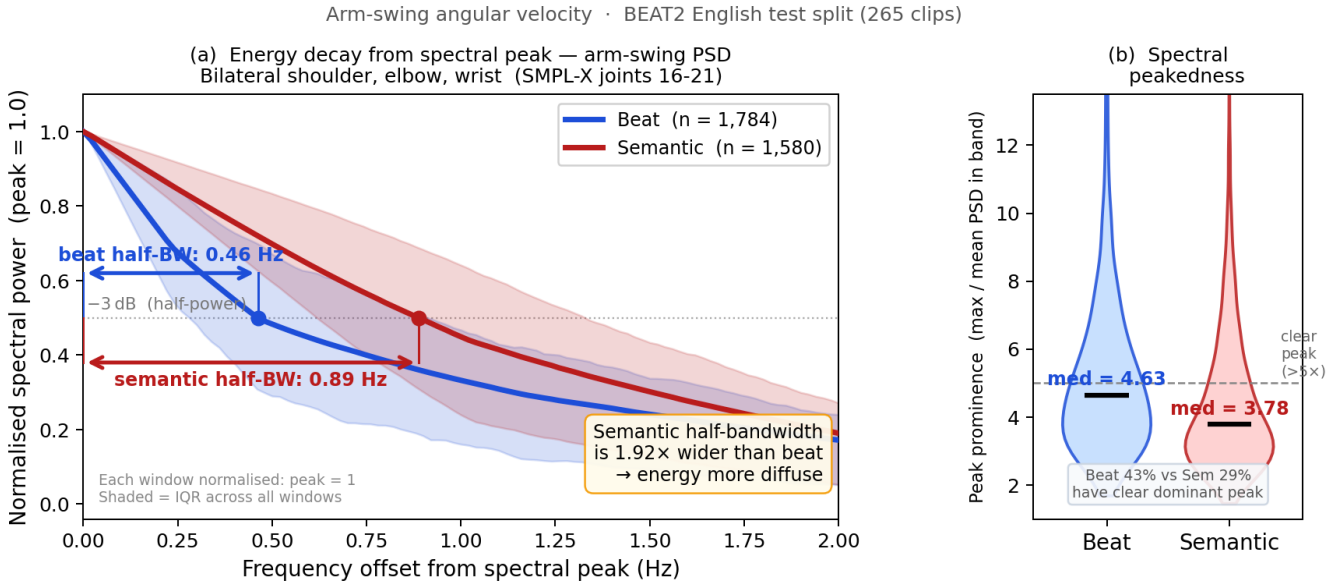


Fig. S1. Arm-swing spectral analysis for bilateral shoulder/elbow/wrist joints (SMPL-X joints 16–21), matched beat and semantic windows (BEAT2 test split, ≈ 1,870 windows each, 25 speakers). **(a)** Mean normalised PSD decay from the spectral peak (shaded = IQR): beat spectra (blue) have a half-bandwidth of 0.46 Hz; semantic spectra (red) are 1.9× broader (0.89 Hz), indicating diffuse rather than tonal energy. **(b)** Peak prominence (max/mean PSD in band): 43% of beat windows have a clear dominant peak ($> 5\times$) versus only 29% of semantic windows, confirming that rhythmic regularity is a beat-specific property.

TABLE S2

STAGE-2 ARCHITECTURE SUMMARY. ALL TRANSFORMER LAYERS USE $H = 4$ HEADS, DROPOUT 0.1, AND BATCH-FIRST LAYOUT. STAGE-1 WEIGHTS ARE FROZEN.

Component	**Details**	**Dims**
Inputs		
HuBERT encoder	2× Conv1d (k=3, p=1), BN, GELU	$1024 \to 256$
Onset/amplitude	2-layer MLP	$3 \to 256$
FastText	Linear	$300 \to 256$
Seed pose	VQEncoderV6 (3 layers, frozen)	$337 \to 768$
Speaker ID	Embedding table (25 entries)	$\to 768$
Shared backbone (both streams)		
Self-attention	TransformerEncoder, 1 layer, d_{ff}=1536	$d = 768$
Audio cross-attention	TransformerDecoder, 8 layers, d_{ff}=1536	$d = 768$
MGSC		
TMR projections [23]	Linear (e_m, e_ε)	$256 \to 256$
Softmax memory gate	2× Linear	$256 \to 1$
Semantic cross-attention	TransformerDecoder, 1 layer, d_{ff}=512	$d = 256$
Output MLP	2-layer	$256 \to 256$
S-VIB		
Timing projection	Linear + LN + GELU	$256 \to 64$
Bottleneck heads	2× Linear (mean, log-var)	$320 \to 16$
Latent Sparsity Interpreter	2-layer MLP + softmax	$16 \to 2$
Region decoders (beat & semantic, per region $r \in \{h, u, l\}$)		
Region MLP	2-layer	$768 \to 256$
Temporal downsample	Conv1d, k=2, s=2	$L \to L/2$
HCA (3 decoders)	TransformerDecoder, 1 layer, d_{ff}=512	$d = 256$
Face (beat only)		
Face decoder	TransformerDecoder, 4 layers, d_{ff}=1536	$d = 768$
Downsample	Linear + 2× Conv1d (k=2,s=2) + MLP	$768 \to 256$
Codebook prediction heads (per region, 5 levels)		
Autoregressive decoder	TransformerDecoder, 3 layers, $H = 8$, d_{ff}=1024	$d = 256$
Level classifier	2-layer MLP	$256 \to 256$
Total Stage-2 params		$\approx$ 122M