STROKEFORM: SEMANTIC CO-SPEECH GESTURE GENERATION WITH KINEMATICALLY GROUNDED MOTION TOKENS

Anonymous authors
Paper under double-blind review

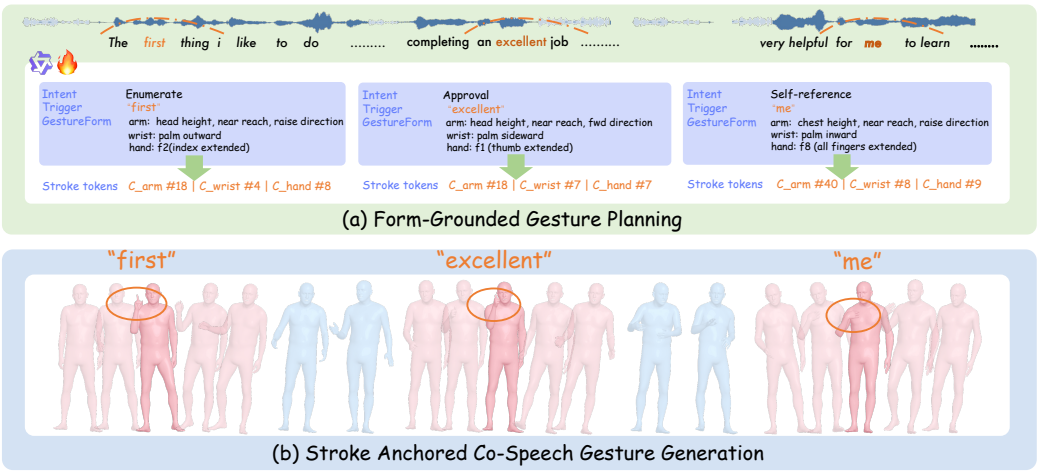


Figure 1: **Overview of StrokeForm.** (a) The planner links utterance meaning to semantic gesture forms and linguistic triggers, compiling supported forms into fixed stroke-token addresses. (b) Speech-conditioned generation preserves these anchors, while phase-conditioned decoding guides realization near the triggers. Pink denotes semantic gesture events, blue surrounding co-speech motion, and orange circles highlight selected semantic gestures.

ABSTRACT

Semantic co-speech gestures express linguistic meaning through body and hand movements. Existing speech-conditioned methods generate natural, rhythmically synchronized motion but remain limited in generating contextually appropriate semantic gestures. Learning such gestures from speech–motion training data is challenging due to their sparse, long-tailed distribution and diverse realizations of the same meaning. Large language models can support context-aware gesture planning, but their plans require an explicit interface to motion generation. We present StrokeForm, a framework that connects language-based gesture planning to motion generation through kinematically grounded motion tokens. An adapted language model plans semantic gestures by specifying their arm configurations, palm orientations, and handshapes, together with contextually relevant linguistic triggers. A grounded tokenizer maps these kinematic attributes to addressable stroke tokens. Anchored near their linguistic triggers, these tokens guide a speech-conditioned masked infiller to generate complete gesture events, while a separate phase condition guides temporal realization. Experiments on BEAT2 show that StrokeForm obtains the highest mean human-rated semantic appropriateness among the evaluated generation methods and supports controllable realization of supported gesture forms, with motion-quality trade-offs.

1 INTRODUCTION

Co-speech gesture generation aims to synthesize body and hand movements that naturally accompany speech. Recent methods have improved motion realism, rhythmic synchrony, and holistic body–hand coordination (Liu et al., 2024; Qi et al., 2025; Liu et al., 2025c). However, natural movement does not necessarily convey the meaning of an utterance. Human evaluations show that motion

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

naturalness and appropriateness for the accompanying speech can diverge substantially (Nagy et al., 2025). This distinction is particularly important for semantic co-speech gestures, whose communicative role depends on their relation to linguistic content.

Semantic gestures convey meaning through body and hand movements coordinated with speech (McNeill, 1992; Kendon, 2004). An extended index finger may express enumeration near "first," while a body-directed hand configuration may express self-reference near "me." Generating such expressions requires interpreting linguistic context and realizing an appropriate gesture at the relevant moment. Neither step has a unique solution. An utterance may admit several gesture expressions, and each expression can be realized through different joint configurations and trajectories. We focus on gestures with salient arm or hand configurations around a single stroke peak, allowing the requested configuration to be specified without prescribing the complete movement.

Existing methods introduce semantic guidance through linguistic representations, communicative intents, or retrieved motion examples (Zhang et al., 2024; 2025a; Mughal et al., 2025; Liu et al., 2025b; Suresh et al., 2026). These approaches establish the value of linguistic and exemplar-based guidance. Our interface instead specifies measurable kinematic attributes at linguistic anchors without prescribing a complete trajectory. Shared geometric readouts supervise and evaluate specification adherence; human evaluation separately assesses semantic appropriateness, following the distinction emphasized by Nagy et al. (2025).

We present StrokeForm, connecting planning and generation through kinematically grounded stroke tokens. An adapted language model predicts arm configurations, palm orientations, handshapes, and linguistic trigger spans. A grounded tokenizer maps supported specifications to fixed stroke-token addresses while learning their motion realizations. A speech-conditioned masked infiller completes surrounding motion, while separate phase conditions guide realization at scheduled stroke peaks. This constrains requested attributes while leaving surrounding co-speech motion generative.

On BEAT2, StrokeForm achieves the highest mean human-rated semantic appropriateness among evaluated methods. Planner conditioning increases mean semantic appropriateness and speech–gesture alignment over audio-only generation, with motion-quality trade-offs.

Our contributions are summarized as follows.

- **Stroke-peak event formulation.** We formulate a class of semantic co-speech gestures as stroke-peak-centered events, linking measurable form specifications to linguistic triggers while leaving surrounding motion generative.
- **Kinematically grounded motion tokens.** We introduce a tokenizer compiling supported specifications into fixed, partwise stroke-token addresses, plus a stroke-anchored masked infiller and phase-conditioned decoder for gesture-event realization.
- **Language-guided gesture realization.** We connect language-guided planning to the grounded motion interface. Experiments show higher mean human-rated semantic appropriateness with planner conditioning and responsiveness to specification edits, while quantifying execution accuracy and motion-quality trade-offs.

2 Related Work

Semantic co-speech gesture generation. Speech-driven gesture generation has advanced motion realism, rhythmic synchrony, and holistic body–hand coordination (Liu et al., 2024; 2025c; Qi et al., 2025; Cheng & Huang, 2025; Zhang et al., 2026). Text-aware approaches incorporate linguistic context through learned representations, token priors, or frame-level semantic emphasis (Liu et al., 2025d; Zhang et al., 2025a; Liu et al., 2025a; 2026). Other methods introduce explicit semantic guidance through motion retrieval (Mughal et al., 2025; Cheng et al., 2024; Yang et al., 2025), semantic categories or communicative intents (Zhang et al., 2024; Liu et al., 2025b), and natural-language descriptions paired with motion examples (Suresh et al., 2026; Chen et al., 2025). Our interface specifies measurable arm configurations, palm orientations, and handshapes at linguistic anchors, rather than a complete motion exemplar. Supported specifications map to stroke-token addresses; geometric readouts measure execution, while human evaluation assesses semantic appropriateness.

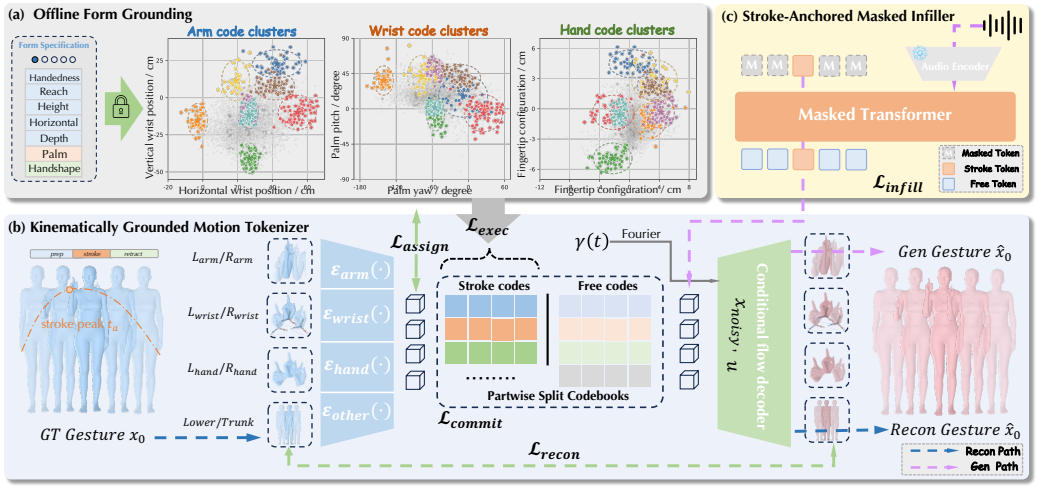


Figure 2: **StrokeForm architecture.** (a) Offline grounding groups stroke-peak arm, wrist, and hand readouts into form regions with fixed specification-to-token mappings. (b) Part-wise encoders produce stroke and free-motion codes; a phase-conditioned flow decoder reconstructs motion under stroke-peak attribute supervision. (c) The speech-conditioned infiller preserves stroke anchors and predicts surrounding free-motion tokens, decoded into gesture events by the shared decoder.

Motion representations and controllable generation. Discrete motion tokenizers support generative modeling through part-wise and residual vector quantization (Liu et al., 2024; Guo et al., 2024; Nagano et al., 2026). Recent work further explores diffusion-decoded and language-guided codebooks (Gu et al., 2026; Yan et al., 2026), discrete–continuous factorization (Wang et al., 2026), and codebook regions associated with semantic categories (Zhao et al., 2026). Complementary approaches incorporate explicit motion constraints, including keyframes and spatial targets, through masked token modeling (Pinyoanuntapong et al., 2024; 2025), diffusion-based control (Dai et al., 2024; Zhao et al., 2025), and sparse-keyframe interpolation (Zhan et al., 2026). StrokeForm combines structured tokenization and sparse conditioning through fixed addresses grounded in measurable arm configurations, palm orientations, and handshapes, rather than semantic categories alone. Embeddings and decoded realizations remain learned, allowing supported forms to be requested without complete poses or exemplars. Execution supervision promotes joint attribute realization at the scheduled peak, while iterative masked generation (Chang et al., 2022; Guo et al., 2024) preserves anchors and completes surrounding motion.

3 Method

StrokeForm connects language-guided gesture planning to speech-conditioned motion generation through stroke-peak event specifications (Section 3.1). A kinematically grounded tokenizer maps these specifications to addressable stroke tokens and learns their motion realization (Section 3.2). A masked infiller generates the surrounding co-speech motion while preserving the stroke anchors (Section 3.3). The gesture planner predicts specifications and linguistic triggers, which are scheduled and compiled into anchors (Section 3.4). Figure 2 illustrates the motion pipeline.

3.1 Gesture Events and Form Specifications

Gesture events. Semantic co-speech gestures convey meaning through body and hand movements coordinated with linguistic content. An extended index finger may express enumeration near "first," while a body-directed hand configuration may express self-reference near "me." We therefore represent a semantic gesture as an event linking a form specification to a linguistic trigger span and an event schedule, rather than as an isolated pose.

Following the account of the stroke as the principal meaning-bearing phase of a gesture (McNeill, 1992), we focus on expressions with a salient arm or hand configuration near a stroke peak. The peak provides a reference moment for realizing the requested form, while the surrounding movement

completes the gesture event. Each event is represented as

$$E_m = (\ell_m, h_m, \kappa_m, \mathbf{t}_m), \qquad \mathbf{t}_m = (t_{0,m}, t_{a,m}, t_{1,m}), \tag{1}$$

where ℓ_m is the linguistic trigger span, $h_m \in \{\text{left}, \text{right}, \text{both}\}$ identifies participating hands, and κ_m is the form specification. The timestamps $t_{0,m} < t_{a,m} < t_{1,m}$ denote event onset, stroke peak, and offset. The trigger associates the gesture with linguistic content, while the motion interval may extend beyond that span. Figure 2(b,c) illustrates how stroke tokens anchor the requested form while surrounding motion remains generative. This formulation does not assume that meaning is determined by a single pose. It uses one shared stroke peak per event and excludes expressions that depend primarily on extended trajectories, multiple strokes, or asynchronous bimanual structure.

Form specifications. We describe gesture form through complementary kinematic factors. Arm configuration specifies body-relative hand placement, reach, and elbow flexion, together with inter-hand relations for bimanual events. Palm orientation specifies facing direction, and handshape specifies finger configuration. These factors capture placement, orientation, and shape cues relevant to semantic expression without prescribing a complete pose or trajectory. Their interpretation remains context-dependent. The schema contains 9 field types with 67 field-specific values. For each participating side s, the specification is

$$\kappa_m^s = (\kappa_{m,1}^s, \ldots, \kappa_{m,9}^s) \in \prod_{j=1}^{9} \big(\mathcal{V}_j \cup \{\varnothing\}\big), \qquad \sum_{j=1}^{9} |\mathcal{V}_j| = 67, \tag{2}$$

where $\mathcal{V}_j$ is field type j's value set and $\varnothing$ denotes an inactive or unavailable attribute. The event specification κ_m collects the participating side specifications, which share handedness h_m and the bimanual relation. For motion-derived labels, the seven per-side geometric fields satisfy

$$\kappa_{m,j}^s = Q_j\big(r_j^s(\mathbf{x}, t_{a,m})\big), \qquad j \in \mathcal{J}_{\text{geom}}, \tag{3}$$

where s denotes a participating side, $\mathcal{J}_{\text{geom}}$ indexes the seven geometric fields, r_j^s computes the corresponding forward-kinematic readout, and Q_j discretizes it.

Discretization reflects training-data support and empirical decoding resolution. A pilot decoder conditioned on continuous stroke-peak readouts guides continuous-attribute granularity, while handshape categories follow finger geometry. The schema is fixed before planner adaptation and tokenizer training. Motion-derived labels supervise the planner and define the partwise keys used to construct addressable stroke-token vocabularies (Section 3.2; Fig. 2(a)). Readout definitions, attribute values, and calibration rules are detailed in the appendix.

3.2 Kinematically Grounded Motion Tokenizer

Semantic gesture planning requires explicit kinematic constraints, yet reconstruction-oriented tokenizers do not inherently map requested attributes to codebook indices. We assign supported attribute combinations (Section 3.1) fixed, partwise stroke-token addresses, providing a stable interface for deterministic compilation into anchors that the masked infiller preserves while generating surrounding speech-conditioned motion tokens. Continuous realization of the requested attributes remains learned and probabilistic, supported by stroke-peak execution supervision and phase-conditioned decoding.

Grounding specifications to stroke tokens. We partition SMPL-X rotational motion into eight groups comprising the left and right arms, wrists, and hands, together with the trunk and lower body. Each part-wise encoder produces latents $\mathbf{z}_p = E_p(\mathbf{x}_p)$ at one quarter of the motion frame rate. Arm, wrist, and hand partitions use split codebooks $\mathcal{C}_p = [\mathcal{C}_{\text{stroke},p} \mid \mathcal{C}_{\text{free},p}]$ with disjoint index ranges. The trunk and lower body use only free-motion codes.

Within each controllable partition, training stroke-peak readouts are grouped by form keys to construct a supported vocabulary $\mathcal{K}_p$. The mapping g_p assigns each key a unique stroke-token address (Figure 2(a)). The vocabulary and mapping are fixed before tokenizer training, while the embeddings and decoded motion realizations are learned. Each address identifies a supported form class rather than a single pose or motion exemplar. Vocabulary construction is detailed in the appendix.

Let $\mathcal{P}_m$ contain the arm, wrist, and hand partitions of participating sides, and $\kappa_{m,p}$ their corresponding form keys. An event can be retained only if every required key is complete and belongs to its vocabulary, $\kappa_{m,p} \in \mathcal{K}_p$ for all $p \in \mathcal{P}_m$. For a retained event, let $n_{a,m}$ denote the token position containing its stroke peak. Quantization follows

$$q_{n,p} = \begin{cases} g_p(\kappa_{m,p}), & n = n_{a,m},\ p \in \mathcal{P}_m, \\ \mathrm{NN}_{\mathcal{C}_{\mathrm{free},p}}\big(\mathbf{z}_{n,p}\big), & \text{otherwise}, \end{cases} \tag{4}$$

where $\mathbf{z}_{n,p}$ is the encoder latent at position n for partition p, and NN returns the nearest free-code index. Motion from events that are not retained remains in training and uses free-motion codes at positions without retained anchors. The corresponding embeddings are concatenated across partitions and decoded jointly to produce coordinated body and hand motion (Figure 2(b)).

Phase-conditioned decoding. A stroke token spans multiple frames, so a separate frame-level phase condition locates the requested configuration within its interval.

Within each retained event, signed progress τ_t is measured relative to $t_{a,m}$ and normalized by the pre-peak or post-peak duration. An additional offset $\delta_t = (t - t_{a,m})/s_{\mathrm{time}}$, with fixed s_{time}, preserves absolute time scale. The condition is

$$\gamma_t = \mathrm{PhaseEnc}\big(\tau_t, \delta_t, e_t\big), \tag{5}$$

where e_t indicates retained-event membership; frames outside these events receive a null condition.

Token embeddings are repeated to frame rate to obtain $\mathbf{z}_{\mathrm{frame}}$. A conditional flow-matching decoder with clean-sample parameterization predicts SMPL-X joint rotations,

$$\hat{\mathbf{x}} = D\big(\mathbf{z}_{\mathrm{frame}}, \gamma, \mathbf{x}_{\mathrm{noisy}}, u\big), \tag{6}$$

where $\mathbf{x}_{\mathrm{noisy}}$ is the interpolated noise–motion input and u flow time. Phase conditioning modulates decoder blocks through FiLM; inference integrates the flow from noise.

Learning attribute execution. Assignment supervision aligns encoder latents with prescribed stroke codes, whose embeddings are updated toward assigned latents. Anchors use embeddings addressed by g_p. We supervise stroke-peak attributes on the decoder's clean-motion estimate,

$$\mathcal{L}_{\mathrm{exec}} = \frac{1}{|\mathcal{I}|} \sum_{(m,s,j)\in\mathcal{I}} w(u) \frac{d_j\big(r_j^s(\hat{\mathbf{x}}, t_{a,m}), \Omega_j(\kappa_{m,j}^s)\big)}{s_j}, \tag{7}$$

where $\mathcal{I}$ indexes retained events, participating sides, and their specified geometric fields. The readout r_j^s is defined in Eq. (3), and Ω_j is the admissible region defined by the fixed discretization in Section 3.1. The distance d_j, normalized by measurement scale s_j, is zero inside Ω_j and penalizes deviations outside. The weight $w(u)$ emphasizes lower flow times with less target-motion information; the loss is zero if $\mathcal{I}$ is empty.

Differentiable forward kinematics evaluates all supervised geometric attributes on the same decoded peak pose, not over the token interval. Together with joint decoding, this encourages simultaneous adherence without prescribing a unique pose. Training combines reconstruction, quantization, assignment, and execution losses; the appendix details the full objective, including auxiliary temporal and elbow losses.

An additional anchor-free training view uses only free-motion codes and null phase conditioning, without assignment or execution losses, supporting audio-only generation when no event is supplied or retained. The encoders, codebooks, and decoder are frozen before infiller training.

3.3 Stroke-Anchored Token Infilling

Following the masked generative modeling framework of MoMask (Guo et al., 2024), we generate speech-conditioned motion tokens while preserving compiled stroke anchors $\mathcal{A}$ (Fig. 2(c)). All remaining part-wise positions $\mathcal{F}$ are predicted exclusively from free-motion codebooks. A shared transformer with partition-specific heads conditions on visible tokens, speech features pooled to token rate, and token-level phase ϕ relative to scheduled stroke peaks, complementing the decoder's frame-level γ.

Masked training. Given targets $\mathbf{q}^\star$ from the frozen tokenizer, we mask a nonempty subset $\mathcal{M} \subseteq \mathcal{F}$ while keeping anchors visible:

$$\mathcal{L}_{\text{infill}} = -\frac{1}{|\mathcal{M}|} \sum_{(n,p)\in\mathcal{M}} \log P_\theta \left(q^\star_{n,p} \mid \tilde{\mathbf{q}}, \phi, \text{speech} \right), \tag{8}$$

where $\tilde{\mathbf{q}}$ contains anchors, visible free tokens, and masks; anchors are never prediction targets. Anchor-free training uses $\mathcal{A} = \varnothing$ and null phase conditioning to support audio-only generation.

Anchor-centric generation. Starting with all free positions masked, each round predicts masked tokens and retains a scheduled fraction ranked by

$$R_{n,p} = w_{\text{phase}} R_{\text{phase},n} + w_{\text{conf}} R_{\text{conf},n,p}, \tag{9}$$

where $R_{\text{phase},n}$ decreases with distance to a stroke peak and $R_{\text{conf},n,p}$ measures prediction confidence. Lower-scoring predictions are remasked, favoring context completion near anchors. Within each round, updates proceed through arm, wrist, and hand partitions in sequence; anchors remain fixed throughout. Without events, remasking uses confidence alone. Completed tokens are decoded under γ; specification edits replace anchors and regenerate free motion under unchanged speech and timing. Phase encodings and update schedules are detailed in the appendix.

3.4 Form-Grounded Gesture Planner

Supervision and adaptation. The planner maps linguistic context to gesture intents and forms. We construct 13K training-event quadruples $(\text{context}, \ell, \text{intent}, \kappa)$ by pairing model-assisted trigger and intent annotations, audited on a sampled subset, with motion-extracted form labels. These describe observed events, not exhaustive plans. We adapt Qwen3-8B with LoRA, serializing trigger, intent, and form. With teacher-forced triggers, supervision covers only intent and form tokens:

$$\mathcal{L}_{\text{adapt}} = -\mathbb{E}_{(c,y)} \sum_{k\in\mathcal{S}_{\text{IF}}(y)} \log p_{\theta,\Delta}(y_k \mid c, y_{<k}), \tag{10}$$

where c is linguistic context, y the serialized event, θ frozen model parameters, Δ LoRA parameters, and $\mathcal{S}_{\text{IF}}$ the intent/form token positions. Trigger selection is elicited by prompting and is not directly supervised during adaptation.

Prediction and compilation. In a single call, the planner receives the transcript, word timestamps, and interface schema. Prompting elicits a sparse, possibly empty event list; constrained decoding enforces legal fields and values. A training-calibrated scheduler places stroke peaks within aligned trigger spans and assigns positive pre- and post-peak durations.

Compilation requires valid trigger timing and complete, sufficiently confident keys in $\mathcal{K}_p$ for all participating partitions. Events are processed in peak order, rejecting overlapping intervals or occupied anchor-token positions. Retained events set $q_{n_{a,m},p} = g_p(\kappa_{m,p})$ for $p \in \mathcal{P}_m$ and supply ϕ and γ; rejected events supply neither anchors nor phase conditions. User plans undergo the same checks without confidence filtering. Annotation, scheduling, and compilation details are in the appendix.

4 Experiments

4.1 Settings

Dataset and implementation. We use BEAT2 English with the official clip-level split: `train+additional` for training, `val` for model selection, and `test` for evaluation. Filtering for annotated gesture events and available features yields 1,144/92/225 clips (43.8/1.9/4.4 hours). The 25 speakers overlap across splits, so evaluation concerns held-out clips rather than unseen-speaker generalization. Motion uses 6D rotations over 52 SMPL-X joints at 30 fps, with 120-frame windows and temporal downsampling by four. The tokenizer is frozen before infiller training, and the planner uses LoRA-adapted Qwen3-8B. Each module is trained on a single NVIDIA H20. Annotation procedures and implementation details are provided in the appendix.

Table 1: **Quantitative comparison with SOTA on BEAT2.** Semantic appropriateness and speech–gesture alignment are mean matching scores: +1 for correct matches, -1 for incorrect matches, and 0 for tie/both/neither responses. Motion realism uses Elo ratings with ground truth fixed at 1000. → denotes closer to ground truth;Fig. 3 shows 95% confidence intervals.

	Single speaker			All 25 speakers			Human evaluation		
Method	FGD ↓	BC →	DIV →	FGD ↓	BC →	DIV →	Sem. ↑	Speech ↑	Realism ↑
Ground truth	–	0.703	11.97	–	0.477	7.29	0.41	0.45	1000
EMAGE (Liu et al., 2024)	0.551	0.772	13.06	0.692	0.284	6.06	-0.05	0.05	934
SemTalk (Zhang et al., 2025a)	**0.428**	0.777	12.91	**0.356**	0.841	**8.41**	0.23	0.07	942
EchoMask (Zhang et al., 2025b)	**0.462**	0.774	13.37	0.566	**0.495**	9.30	-0.04	-0.10	935
RAG-Gesture (Mughal et al., 2025)	0.808	**0.734**	**11.92**	0.487	**0.514**	9.94	0.04	-0.09	**947**
Ours (audio-only)	0.497	**0.736**	13.18	**0.421**	0.784	9.23	**0.25**	**0.16**	**967**
Ours (with planner)	0.604	0.798	**12.68**	0.588	0.828	**8.26**	**0.40**	**0.25**	934

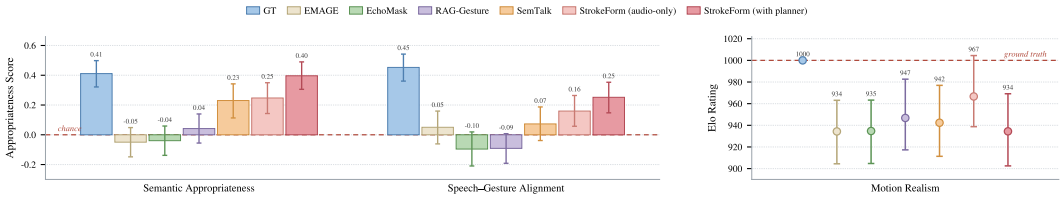


Figure 3: **Human evaluation on BEAT2.** Plots show mean matching scores for semantic appropriateness and speech–gesture alignment, and Elo ratings for realism. Error bars indicate participant-bootstrap 95% confidence intervals. Zero denotes no net preference for the matched alternative.

Generation evaluation. We report Fréchet Gesture Distance (FGD), beat consistency (BC), and motion diversity (DIV), complemented by disentangled human evaluation following GENEA protocols (Nagy et al., 2026). Human evaluation separately assesses realism, speech–gesture alignment, and semantic appropriateness, the latter targeting utterance-specific meaning. We recruit 120 participants for realism and 80 for each appropriateness task. Table 1 defines the scores; Fig. 3 reports participant-bootstrap 95% confidence intervals. Task designs are described in Appendix D.

Form-execution evaluation. We assess specification adherence and responsiveness to edits under fixed speech and event schedules, independently of planner prediction. Execution metrics are computed on compilable held-out events. Slot agreement measures attribute-wise accuracy, and joint event success requires all evaluated per-side geometric attributes to be satisfied simultaneously at the requested stroke peak. Handedness and bimanual relations are excluded. These metrics use fixed geometric readouts shared with training and assess operational adherence rather than independent motion validity. Reference-event coverage is reported in Appendix B.2.

4.2 Results

Quantitative results. Table 1 and Fig. 3 show that StrokeForm obtains the highest mean semantic appropriateness (0.40) and speech–gesture alignment (0.25) among evaluated generation methods. Planner conditioning raises these means from 0.25 and 0.16. The realism point estimate decreases ($\delta = -33$ Elo, 95% CI $[-71, +9]$); the interval includes zero but does not establish preserved realism. Relative to audio-only generation, diversity approaches ground truth, but FGD and beat consistency worsen.

	Slot agreement ↑			Event
Request	Arm	Wrist	Hand	success ↑
Audio-only	30.7	38.4	42.9	3.1
Prior-sampled	33.1	41.2	45.7	3.8
Reference κ	**79.8**	**85.1**	**86.9**	**44.2**
Original κ	22.6	29.4	32.1	1.2
Edited κ'	**71.6**	**77.9**	**80.8**	**33.8**

Table 2: **Form execution (%).** Scores are evaluated against κ (upper) and κ' (lower).

Of 3,599 annotated test events, 68% pass reference-request compilation; this is not the acceptance rate of predicted plans. Rejections mainly reflect

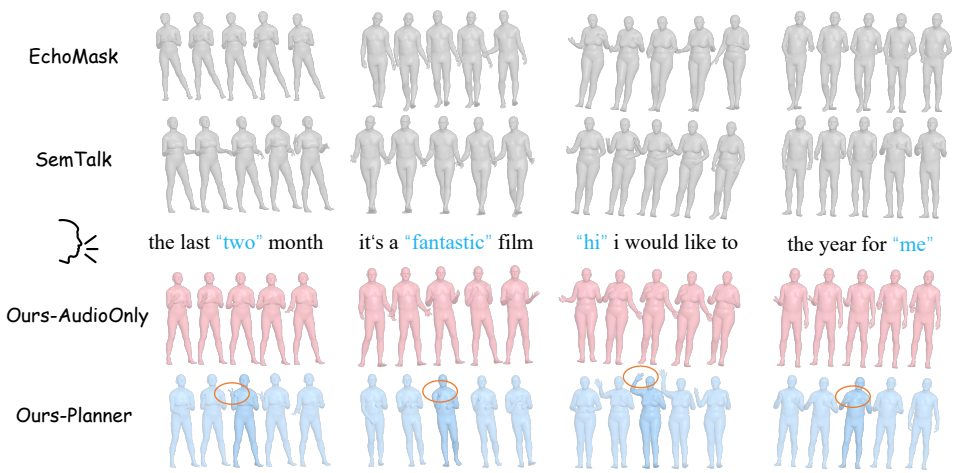


Figure 4: **Qualitative comparison on BEAT2.** Each column compares chronological snapshots for the same speech. Blue text marks target words; orange circles highlight our planner-conditioned gestures. Compared with EchoMask, SemTalk, and our audio-only variant, planner conditioning yields more explicit gestures for enumeration, approval, greeting, and self-reference.

Table 3: **Ablation studies.** Planner variants share linguistic inputs and annotated trigger spans; motion variants share reference specifications and schedules.

(A) Planner adaptation

Variant	Intent (%)↑	Arm (%)↑
LoRA (ours)	**71.5**	**43.8**
Schema-only	41.2	18.6
Retrieval	–	26.3

(B) Motion mechanisms

Variant	Success (%)↑	Peak (ms)↓	FGD↓
Full	**44.2**	**140**	**0.571**
w/o $\mathcal{L}_{\text{exec}}$	24.5	149	0.579
w/o phase conditioning	11.8	518	0.643
Anchor-agnostic remasking	29.7	176	0.689

insufficient training support for rare partwise attribute combinations. Table 2 evaluates execution on compilable held-out events independently of planner prediction. Slot agreement measures attribute-wise accuracy; joint success requires all evaluated geometric attributes to hold at the requested stroke peak. Reference specifications yield 79.8%, 85.1%, and 86.9% arm, wrist, and hand agreement, respectively.

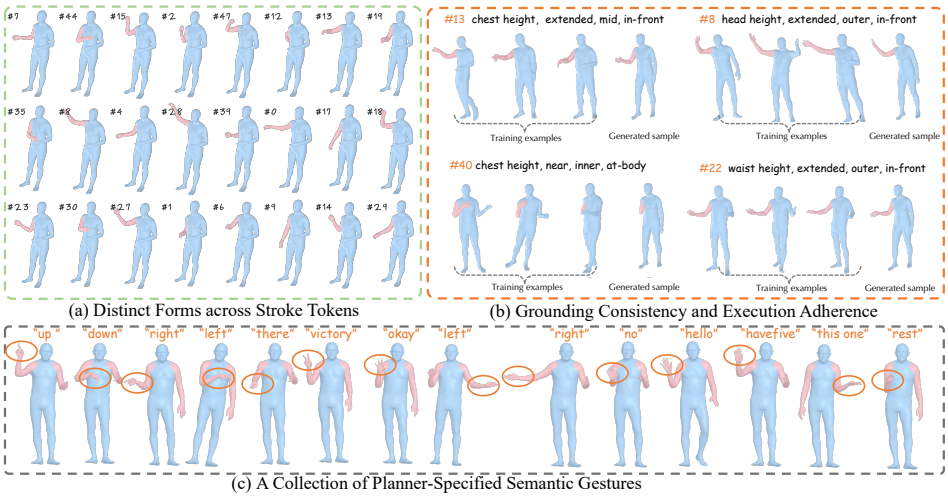
Joint success reaches 44.2%, indicating incomplete simultaneous execution even for supported requests. Audio-only generation and prior-sampled requests achieve 3.1% and 3.8%, respectively, against the same reference targets. The lower block compares requests κ and κ', both evaluated against κ' under unchanged speech and timing. Editing the request raises joint success from 1.2% to 33.8%, demonstrating responsiveness to specification edits.

Qualitative results. Figure 4 shows more explicit enumeration, approval, greeting, and self-reference gestures with planner conditioning than with EchoMask, SemTalk, or our audio-only variant. Recognizable handshapes and arm configurations appear near their linguistic triggers. We strongly encourage readers to view the supplementary videos for the semantic expressiveness of our generated gestures.

Figure 5 links right-arm stroke tokens to real and generated motions sharing the requested attributes, illustrating form consistency across varied realizations. Planner-specified examples further demonstrate semantic expression while leaving surrounding trajectories generative.

4.3 Ablation Studies

We isolate conditional planning from motion execution: planner variants share annotated triggers, while motion variants share reference specifications and schedules.



(a) Distinct Forms across Stroke Tokens

(b) Grounding Consistency and Execution Adherence

(c) A Collection of Planner-Specified Semantic Gestures

Figure 5: **Kinematic grounding and semantic gesture realization.** **(a)** 24 of the 48 right-arm stroke tokens under a fixed body. **(b)** For four selected stroke tokens, three assigned BEAT2 examples and a form-specified decoding exhibit the same specified kinematic attributes, illustrating attribute consistency across these examples. **(c)** Semantic gesture examples generated from planner-specified forms, with orange circles highlighting the relevant arm and hand configurations.

Planner adaptation. LoRA raises intent accuracy from 41.2% to 71.5% and arm-slot agreement from 18.6% to 43.8%, compared with 26.3% for retrieval (Table 3(A)). This supports interface-specific supervision for intent/form prediction. Annotated triggers exclude event-selection errors, and agreement with one reference does not capture all valid alternatives.

Execution supervision. Removing $\mathcal{L}_{\text{exec}}$ reduces event success from 44.2% to 24.5%, while peak error changes from 140 to 149 ms and FGD from 0.571 to 0.579 (Table 3(B)). The larger adherence loss despite small timing and FGD changes shows that distributional motion quality alone does not ensure attribute execution.

Phase conditioning. This variant removes the infiller's ϕ and the decoder's γ, while retaining anchor-centric remasking. Event success drops to 11.8% and peak error rises to 518 ms, supporting the contribution of phase information in this architecture. Success is measured at the requested frame without frame search, so this comparison evaluates timely adherence rather than timing-independent form accuracy.

Anchor-centric remasking. Under identical decoding and update budgets, anchor-agnostic remasking reduces success from 44.2% to 29.7%, increases peak error from 140 to 176 ms, and raises FGD from 0.571 to 0.689. Thus, surrounding token generation affects execution even with fixed anchors. Prioritizing context near anchors improves adherence, timing, and distributional motion quality on this subset.

Peak error is the median absolute error between detected and requested stroke-peak times. FGD uses the reference-event subset and is not directly comparable to Table 1.

5 Conclusion

We presented StrokeForm, connecting language-guided gesture planning to motion generation through kinematically grounded stroke tokens. Fixed form-to-token addresses and phase-conditioned generation support explicit attribute requests while leaving surrounding motion generative. On BEAT2, planner conditioning improves mean human-rated semantic appropriateness, and controlled evaluations demonstrate adherence and responsiveness to form edits. Remaining limitations include incomplete joint execution, motion-quality trade-offs, and restriction to single-peak events within the supported vocabulary. Future work will extend coverage to trajectory-dependent and multi-stroke gestures.

REFERENCES

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

Bohong Chen, Yumeng Li, Youyi Zheng, Yao-Xiang Ding, and Kun Zhou. Motion-example-controlled co-speech gesture generation leveraging large language models. In *SIGGRAPH 2025 Conference Papers*, pp. 15, 2025. doi: 10.1145/3721238.3730611.

Qingrong Cheng, Xu Li, Xinghui Fu, Fei Xia, and Zhongqian Sun. Siggesture: Generalized co-speech gesture synthesis via semantic injection with large-scale pre-training diffusion models. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 13, 2024. doi: 10.1145/3680528.3687677.

Yongkang Cheng and Shaoli Huang. Hologest: Decoupled diffusion and motion priors for generating holistically expressive co-speech gestures. In *International Conference on 3D Vision (3DV)*, 2025.

Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. In *Computer Vision – ECCV 2024*, 2024. doi: 10.1007/978-3-031-72640-8_22.

Chenyang Gu, Mingyuan Zhang, Haozhe Xie, Zhongang Cai, Lei Yang, and Ziwei Liu. Bridging semantic and kinematic conditions with diffusion-based discrete motion tokenizer. *arXiv preprint arXiv:2603.19227*, 2026. doi: 10.48550/arXiv.2603.19227.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1900–1910, 2024. doi: 10.1109/CVPR52733.2024.00186.

Adam Kendon. *Gesture: Visible Action as Utterance*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511807572.

Haiyang Liu, Zihao Zhu, et al. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *CVPR*, pp. 1144–1154, 2024.

Lanmiao Liu, Esam Ghaleb, Aslı Özyürek, and Zerrin 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*, 2025a.

Lanmiao Liu, Esam Ghaleb, Aslı Özyürek, and Zerrin Yumak. Semconflow: Semantic grounding of holistic co-speech gesture generation with contrastive flow-matching, 2026.

Pinxin Liu, Haiyang Liu, Luchuan Song, Jason J. Corso, and Chenliang Xu. Intentional gesture: Deliver your intentions with gestures for speech, 2025b.

Pinxin Liu, Luchuan Song, Junhua Huang, and Chenliang Xu. GestureLSM: Latent Shortcut based Co-Speech Gesture Generation with Spatial-Temporal Modeling. In *IEEE/CVF International Conference on Computer Vision*, 2025c.

Pinxin Liu, Pengfei Zhang, Hyeongwoo Kim, Pablo Garrido, Ari Shapiro, and Kyle Olszewski. Contextual gesture: Co-speech gesture video generation through context-aware gesture representation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025d. doi: 10.1145/3746027.3755140.

David McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, 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. In *Computer Vision and Pattern Recognition (CVPR)*, 2025.

Koki Nagano, Hongyu Liu, Seonwook Park, Tianye Li, Amrita Mazumdar, Christian Jacobsen, Shengze Wang, Michael Stengel, Rajarshi Roy, Ka Chun Cheung, Simon See, and Shalini De Mello. Dyaplex: Full-duplex speech-motion model for dyadic interaction, 2026.

Rajmund Nagy, Hendric Voss, Thanh Hoang-Minh, Mihail Tsakov, Teodor Nikolov, Zeyi Zhang, Tenglong Ao, Sicheng Yang, Shaoli Huang, Yongkang Cheng, M. Hamza Mughal, Rishabh Dabral, Kiran Chhatre, Christian Theobalt, Libin Liu, Stefan Kopp, Rachel McDonnell, Michael Neff, Taras Kucherenko, Youngwoo Yoon, and Gustav Eje Henter. Towards reliable human evaluations in gesture generation: Insights from a community-driven state-of-the-art benchmark. *arXiv preprint arXiv:2511.01233*, 2025.

Rajmund Nagy, Silvia Arellano García, Hendric Voss, Mihail Tsakov, Taras Kucherenko, Youngwoo Yoon, and Gustav Eje Henter. The genea challenge 2026: A large-scale disentangled evaluation of speech-driven gesture generation on the seamless interaction dataset, 2026. URL `https://arxiv.org/abs/2608.10839`.

Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1546–1555, 2024.

Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Korrawe Karunratanakul, Pu Wang, Hongfei Xue, Chen Chen, Chuan Guo, Junli Cao, Jian Ren, and Sergey Tulyakov. Maskcontrol: Spatio-temporal control for masked motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Xingqun Qi, Yatian Wang, Hengyuan Zhang, Jiahao Pan, Wei Xue, Shanghang Zhang, Wenhan Luo, Qifeng Liu, and Yike Guo. Co3gesture: Towards coherent concurrent co-speech 3d gesture generation with interactive diffusion. In *International Conference on Learning Representations*, 2025.

Varsha Suresh, Mohammad Mahdi Abootorabi, Mohamed Salman, M. Hamza Mughal, Christian Theobalt, Ashwin Ram, Jürgen Steimle, and Vera Demberg. Semantic motion anchors: Bridging motion and meaning in co-speech gestures, 2026.

Hequan Wang, Xue'an Chen, Jiaxu Zhang, Zhengbo Zhang, and Zhigang Tu. Dc-motion: Decoupling structure and details via discrete-continuous tokens for human motion generation. *arXiv preprint arXiv:2606.14721*, 2026. doi: 10.48550/arXiv.2606.14721.

Sheng Yan, Yong Wang, Xin Du, Junsong Yuan, and Mengyuan Liu. Language-guided transformer tokenizer for human motion generation. In *Computer Vision – ECCV 2026*, Lecture Notes in Computer Science, pp. 312–332. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-37041-9_17.

Quanwei Yang, Luying Huang, Kaisiyuan Wang, Jiazhi Guan, Shengyi He, Fengguo Li, Hang Zhou, Lingyun Yu, Yingying Li, Haocheng Feng, and Hongtao Xie. Gesturehydra: Semantic co-speech gesture synthesis via hybrid modality diffusion transformer and cascaded-synchronized retrieval-augmented generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Xiaoyu Zhan, Xinyu Fu, Chenghao Yang, Xiaohong Zhang, Dongjie Fu, Pengcheng Fang, Tengjiao Sun, Xiaohao Cai, Hansung Kim, Yuanqi Li, Jie Guo, and Yanwen Guo. Umo: Unified sparse motion modeling for real-time co-speech avatars, 2026.

Xiangyue Zhang, Jianfang Li, Jiaxu Zhang, Ziqiang Dang, Jianqiang Ren, Liefeng Bo, and Zhigang Tu. Semtalk: Holistic co-speech motion generation with frame-level semantic emphasis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13761–13771, 2025a.

Xiangyue Zhang, Jianfang Li, Jiaxu Zhang, Jianqiang Ren, Liefeng Bo, and Zhigang Tu. Echomask: Speech-queried attention-based mask modeling for holistic co-speech motion generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10827–10836, 2025b.

Xiangyue Zhang, Jianfang Li, Jianqiang Ren, and Jiaxu Zhang. Mitigating error accumulation in co-speech motion generation via global rotation diffusion and multi-level constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Zeyi Zhang, Tenglong Ao, Yuyao Zhang, et al. Semantic gesticulator: Semantics-aware co-speech gesture synthesis. *ACM Transactions on Graphics*, 43(4), 2024.

Junchuan Zhao, Qifan Liang, and Ye Wang. Personagest: Personalized co-speech gesture generation with semantic-guided hierarchical motion representation, 2026.

Kaifeng Zhao, Gen Li, and Siyu Tang. Dartcontrol: A diffusion-based autoregressive motion model for real-time text-driven motion control. In *International Conference on Learning Representations*, 2025.

Appendix

Table of Contents

A Use of Large Language Models **14**

B Gesture Annotation and Form Grounding **14**

B.1 Event Annotation and Planner Supervision 14

B.2 Kinematic Form Schema and Vocabulary 15

C Implementation Details **16**

C.1 Tokenizer and Motion Decoder 16

C.2 Planner and Event Compilation 17

C.3 Phase-Conditioned Masked Infilling 18

D Human Evaluation Protocols **18**

A USE OF LARGE LANGUAGE MODELS

We used large language models to assist with drafting and polishing sections of the paper, including grammar, wording, and clarity. We also used GPT-5.6 to generate candidate semantic annotations for planner supervision, as described in Appendix B.1; physical form labels were extracted from motion rather than generated by the language model. The authors reviewed and revised the manuscript and take responsibility for its scientific content, analyses, and conclusions.

B GESTURE ANNOTATION AND FORM GROUNDING

This section describes the event annotations used in Sections 3.1 and 3.4 and the form-to-token interface in Section 3.2. Motion-based detection and geometric labeling are independent of semantic interpretation: the language model selects triggers and intents but does not generate or revise the physical targets.

B.1 EVENT ANNOTATION AND PLANNER SUPERVISION

Motion-based candidates. We extract body-relative wrist trajectories from BEAT2 SMPL-X motion at 30 fps. Wrist height and speed are low-pass filtered at 2 Hz and 4 Hz, respectively. Each side's rest baseline is its 10th-percentile height; the reference range is the pooled 90th–10th percentile difference across both wrists. A hysteresis unit opens at 0.35 reference ranges above baseline and closes below 0.15 ranges. Units shorter than six frames are removed.

Candidates are split at stationary runs lasting at least 0.4 s with both wrist speeds below 12 cm/s, and at valleys below 0.4 times the lower adjacent speed peak. Qualifying peaks reach 25 cm/s and are separated by at least eight frames. Retained intervals require dominant-wrist travel of at least 25 cm and peak speed of at least 25 cm/s. Left/right intervals merge when temporal IoU exceeds 0.5 or their intersection exceeds 0.8 of the shorter interval. Detection is wrist-motion-based; finger configurations are characterized within detected events.

Event timing. For interval $[f_0, f_1)$, let a_f be the maximum of the two filtered wrist speeds. Within the first 75% of the interval, we select the first local peak f_p reaching 0.55 of the interval maximum, with an argmax fallback. The apex is the subsequent arrival frame:

$$f_a = \min\{f : f_p \le f < f_1,\ a_f < 0.30 a_{f_p}\}. \tag{11}$$

Candidates with no frame satisfying this condition are discarded. This arrival frame serves as our operational stroke-peak annotation. Short, lower-speed preparation intervals and qualifying downward retraction phases merge with adjacent strokes. Overlapping duplicates and isolated return-to-rest intervals are removed, and affected apices are recomputed. Retained events satisfy $f_0 < f_a < f_1$, with $(t_0, t_a, t_1) = (f_0, f_a, f_1)/30$. These observed timestamps differ from the inference-time schedules constructed from linguistic triggers in Section 3.4.

Semantic annotation and audit. For each candidate, GPT-5.6 receives the local transcript, word timestamps, and a geometric event summary. It proposes a trigger, a fine-grained intent from a fixed 32-class taxonomy, a realization mechanism, a coarse class, a confidence label, and a short justification. Figure 6(b) shows the semantic distribution. The fine-grained intent is the planner's intent target; the other labels support annotation analysis. Triggers must occur in the aligned transcript, with multiword spans extending from the first word onset to the last word offset.

Sampled manual audits identify systematic errors and guide pipeline refinement before freezing. Supervision retains high-confidence annotations with valid semantic and geometric checks, aligned triggers, and complete supported keys. Ambiguous records and expressions requiring extended trajectories, repeated strokes, or asynchronous bimanual phases are excluded from the addressable set.

The resulting set contains 13,216 quadruples $(\text{context}, \ell_m, \text{intent}_m, \kappa_m)$ (Table 4), with contributions from 25 speakers shown in Figure 6(c). Unannotated intervals are not confirmed negative examples. Planner adaptation teacher-forces triggers and supervises only intent and form tokens, as specified in Section 3.4.

Table 4: Planner-supervision construction. Rows are cumulative; percentages use all detected candidates as the denominator.

Stage	Events	Retained (%)
Detected motion candidates	30,637	100.0
Complete, supported form keys	14,902	48.6
Semantic, geometric, and trigger checks passed	14,772	48.2
High-confidence planner supervision	13,216	43.1

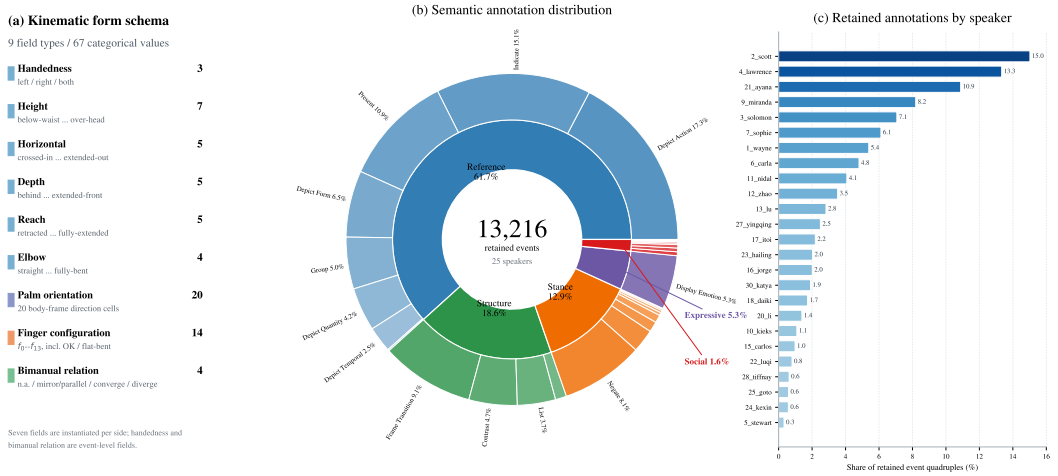


Figure 6: Form schema and planner-supervision statistics. (a) Nine categorical field types define 67 field-specific values; these counts describe the schema, not observed frequencies or stroke-code counts. (b) Coarse-class and fine-grained intent distributions among 13,216 retained training events. Inner and outer rings represent coarse classes and intents, respectively; selected intent labels are shown. Percentages are relative to all retained events. (c) Speaker shares of the same retained set. All distributional statistics use train-plus-additional data.

B.2 Kinematic Form Schema and Vocabulary

Schema and readouts. Figure 6(a) summarizes the nine categorical field types and their 67 field-specific values. Seven fields are instantiated for each participating side; handedness and bimanual relation are shared:

$$\kappa_m^q = (h_m, z_m^q, x_m^q, d_m^q, r_m^q, e_m^q, o_m^q, f_m^q, b_m). \tag{12}$$

The complete event specification collects the participating sides, allowing different left/right forms.

Placement uses a shoulder-centered body frame, with the shoulder line defining the lateral axis and the orthogonalized pelvis-to-neck direction defining the vertical axis. Coordinates are normalized by rest-pose shoulder width; lateral coordinates are mirrored across sides. Height bins use anatomical landmarks. Reach is shoulder-to-wrist distance divided by total arm length. Elbow bins use the shoulder–elbow–wrist angle, with boundaries at $70°$, $120°$, and $150°$.

Palm orientation is computed from the palm-facing direction derived from the global wrist rotation, expressed in the body frame, and discretized using a fixed 26-to-20 directional-cell mapping. Finger configurations use training-calibrated extension thresholds, supplemented by opposition and proximity tests for pinch and OK, and coordinated partial flexion for flat-bent. Handedness follows event-level wrist activity. For bilateral events, inter-wrist distance changes above 12 cm are labeled diverging when the minimum precedes the maximum, and converging otherwise; mirror/parallel uses forearm directions near the apex. This bimanual readout describes event-level motion, not a single peak pose.

Calibration and addressing. A pilot decoder conditioned on continuous readouts guides continuous-attribute bin resolution. The schema and quantizers are frozen before tokenizer training and planner adaptation.

Each participating side supplies three keys:

$$\kappa^q_{m,\text{arm}} = (h_m, z^q_m, x^q_m, d^q_m, r^q_m, e^q_m, b_m), \tag{13}$$

$$\kappa^q_{m,\text{wrist}} = (o^q_m), \qquad \kappa^q_{m,\text{hand}} = (f^q_m). \tag{14}$$

Complete training keys meeting the fixed support criteria form $\mathcal{K}_p$. Each receives one immutable address through g_p; embeddings and decoded realizations remain learned. Handedness selects participating partitions and distinguishes unilateral from bilateral arm contexts. Trigger timing determines anchor placement, not token identity.

Every required key must be complete and supported. Otherwise, the entire event supplies neither anchors nor phase conditions. Unsupported keys are not mapped to neighboring forms; their motion remains available for free-motion training. Inference additionally applies the timing and conflict checks in Section 3.4.

Data isolation and coverage. Calibration, vocabulary construction, and planner supervision use train-plus-additional data. Validation is used for model selection; test labels are evaluated with frozen readouts and vocabularies. Of 3,599 annotated test events, 68% pass the reference-event compilation checks reported in Section 4.2. This held-out coverage differs from the training-supervision retention rate in Table 4.

C Implementation Details

We provide implementation details for the tokenizer, planner, event compiler, and masked infiller. Event annotation and supported-key construction are described in Appendices B.1 and B.2, respectively.

C.1 Tokenizer and Motion Decoder

Architecture. The eight partitions comprise left/right arms, wrists, and hands, the trunk, and the lower body, using 3, 1, 15, 5, and 9 joints per corresponding partition. Rotations use a 6D representation; wrist inputs are expressed in the body frame. Each arm partition contains 48 stroke codes and 384 free-motion codes; each wrist contains 16 stroke codes and 128 free-motion codes; each hand contains 10 stroke codes and 512 free-motion codes. The trunk and lower body use only free-motion codes, numbering 128 and 256, respectively. The partition-specific vocabularies therefore contain 148 stroke entries and 2,432 free-motion entries, totaling 2,580. Stroke-code counts describe supported partwise keys, not the number of categorical values in the annotation schema.

Inputs contain 120 frames at 30 fps. Two stride-2 temporal stages produce 30 token positions, with hidden widths $256 \to 512 \to 512$, two residual blocks per stage, kernel size 4, and dropout 0.1. Code embeddings have dimension 512. The shared flow decoder uses eight blocks of width 512 and phase FiLM with zero-initialized scale/shift heads.

Training objective. For $u \sim \mathcal{U}(0,1)$ and $\epsilon \sim \mathcal{N}(0, I)$, the decoder predicts clean motion:

$$\mathbf{x}_u = (1-u)\epsilon + u\mathbf{x}, \qquad \hat{\mathbf{x}} = D(\mathbf{z}_{\text{frame}}, \gamma, \mathbf{x}_u, u). \tag{15}$$

We group rotation reconstruction and temporal/elbow auxiliary terms into

$$\mathcal{L}_{\text{flow}} = \mathcal{L}_{\text{rot}} + 0.5\mathcal{L}_{\text{vel}} + 0.25\mathcal{L}_{\text{acc}} + 0.1\mathcal{L}_{\text{elbow}}. \tag{16}$$

The complete objective is

$$\begin{aligned}\mathcal{L}_{\text{tok}} = \mathcal{L}_{\text{flow}} + \mathcal{L}_{\text{book}} + 0.25\mathcal{L}_{\text{commit}} \\ + a\left[\mathcal{L}_{\text{assign}} + 2\mathcal{L}_{\text{exec}}\right].\end{aligned} \tag{17}$$

where $a = 0$ for anchor-free samples and $a = 1$ otherwise. Assignment uses fixed-address cross-entropy with temperature 4.

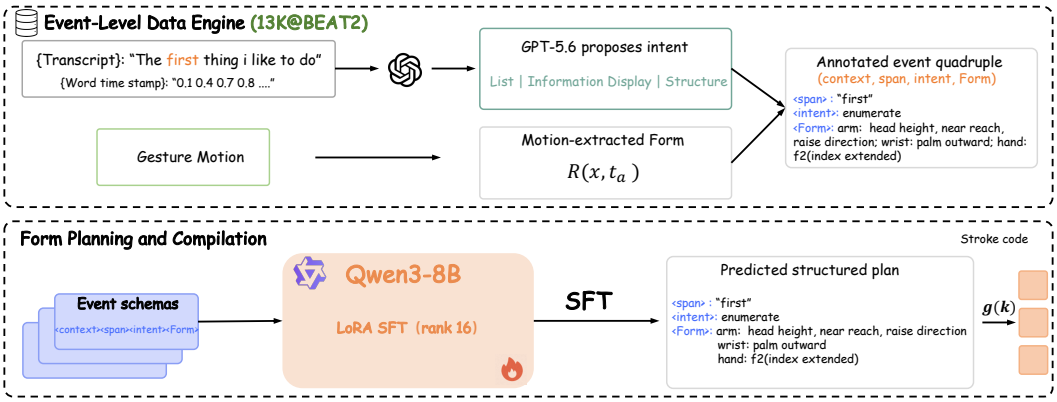


Figure 7: Planner supervision and compilation. Semantic annotations and motion-extracted forms produce 13,216 retained event quadruples. Qwen3-8B is adapted with rank-16 LoRA to predict structured plans. The compiler maps supported partwise keys to fixed stroke addresses and uses trigger-derived schedules to place the anchors.

The execution loss follows Eq. 7, with $w(u) = 1 - u + 0.1$, which reweights execution without changing the uniform flow-time distribution. Execution constraints are applied to the same $\hat{\mathbf{x}}$.

The seven supervised geometric fields are height, horizontal position, depth, reach, elbow configuration, palm orientation, and finger configuration. Handedness selects participating sides, and the dynamic bimanual label provides addressing context; neither enters execution supervision or agreement metrics. Bilateral success concerns both sides at a shared peak, not dynamic coordination.

Inference uses 20 Euler steps with velocity $(\hat{\mathbf{x}} - \mathbf{x}_u)/(1-u)$, evaluated before $u = 1$. Tokenizer training uses AdamW, learning rate 2×10^{-4}, batch size 128, and 20k steps, with 1k warmup steps followed by cosine decay.

C.2 Planner and Event Compilation

Adaptation. Figure 7 separates semantic annotation from motion-derived physical targets. We adapt Qwen3-8B in non-thinking instruction mode using LoRA with rank 16, alpha 32, and dropout 0.05, applied to attention and MLP projection layers. The maximum sequence length is 512. Triggers are teacher-forced during adaptation, and supervision covers only intent and form tokens, including handedness. Trigger selection is elicited by prompting and is not directly supervised during adaptation. Training uses AdamW at 2×10^{-4}, effective batch size 64, three epochs, and 3% warmup followed by cosine decay. Inference uses greedy, schema-constrained decoding.

Scheduling and compilation. For an aligned trigger $[s_0, s_1]$, requested timing is

$$t_a = s_0 + \rho_{\text{intent}}(s_1 - s_0), \qquad t_0 = t_a - \widehat{D}_{\text{pre}}, \quad t_1 = t_a + \widehat{D}_{\text{post}}. \tag{18}$$

The scheduler uses intent-conditioned training medians: ρ_{intent} is the median normalized observed apex position, clipped to $[0.2, 0.8]$, and $\widehat{D}_{\text{pre/post}}$ are median phase durations. Compilation requires valid timing with $t_0 < t_a < t_1$ after grid projection and complete, supported keys for every participating partition.

Field confidence is the geometric mean of its token probabilities; event confidence is the minimum over required form fields:

$$c_f = \exp\left(\frac{1}{|T_f|}\sum_{k\in T_f}\log p(y_k \mid y_{<k}, c)\right), \qquad c_m = \min_{f\in F_m^{\text{req}}} c_f, \tag{19}$$

where c denotes the linguistic input. Events passing support and confidence checks are processed in peak-time order. An event is rejected if its interval overlaps an accepted event or its anchor-token position is already occupied. Rejected events supply neither anchors nor phase conditions. User-specified events undergo the same compilation checks without confidence filtering.

C.3 PHASE-CONDITIONED MASKED INFILLING

Phase conditioning. Non-overlapping compiled intervals uniquely determine event ownership. Within an event, let $D_{\text{pre}} = t_a - t_0$ and $D_{\text{post}} = t_1 - t_a$. Using times in seconds,

$$\tau_t = \begin{cases} (t - t_a)/D_{\text{pre}}, & t < t_a, \\ (t - t_a)/D_{\text{post}}, & t \geq t_a, \end{cases} \qquad \delta_t = (t - t_a)/(1\,\text{s}). \tag{20}$$

Fourier features with four frequencies are projected to a 128D frame-level phase condition. Token-level conditioning combines an embedding of the clipped apex offset, $\text{clip}(n - n_a, -15, 15)$, a normalized relative-time feature, and event membership. Event-phase conditioning is disabled outside valid events and for anchor-free samples.

Masked generation. The infiller uses eight Transformer layers, width 512, eight attention heads, FFN width 2048, and dropout 0.1, with partition-specific free-code heads. Training samples a mask ratio $\cos(\pi r/2)$ for $r \sim \mathcal{U}(0, 1)$, masking at least one free-motion position; stroke anchors are never prediction targets. For anchor-free training, targets are re-quantized using only free-motion codebooks, with null phase conditioning.

Inference uses four outer rounds with remaining-mask ratios $m_r = \cos(\pi r/8)$, $r = 1, \ldots, 4$, relative to the initial number of generatable positions. Within each round, updates proceed arm, wrist, then hand; trunk and lower-body updates accompany the arm step. Anchors remain fixed. Predictions are retained using $R_{n,p} = 0.6R_{\text{phase}}(n) + 0.4R_{\text{conf}}(n, p)$, where confidence is the maximum predicted probability and

$$R_{\text{phase}}(n) = \max_{m \in \mathcal{E}_{\text{ret}}} \left[1 - \min\left(\frac{|n - n_{a,m}|}{15}, 1 \right) \right], \tag{21}$$

with $\mathcal{E}_{\text{ret}}$ denoting retained events. Lower-scoring predictions are remasked, and the final round completes all remaining positions. Audio-only generation uses confidence-only retention.

Infiller training uses AdamW at $5{\times}10^{-5}$, batch size 128, and 6k steps, with 300 warmup steps followed by cosine decay. All modules use weight decay 0.01, gradient clipping 1.0, and bf16 precision.

D HUMAN EVALUATION PROTOCOLS

We separately evaluate motion realism, speech–gesture alignment, and semantic appropriateness, following disentangled human-evaluation designs (Nagy et al., 2025; 2026). Realism uses silent cross-method comparisons, whereas the two matching tasks hold motion fixed and vary its paired audio or text.

Motion realism. We use 10 unique segments containing complete utterances, with valid motion and rendering and no severe tracking artifacts. All seven motion sources, including ground truth, use the same 10 speech segments. Participants compare two silent videos from different sources for the same segment and select the motion that appears more human-like, considering naturalness, smoothness, and plausibility. Removing audio separates this judgment from speech matching. Preferences are converted to Elo ratings, with strong and weak preferences assigned weights of 2 and 1, respectively. Ground-truth motion is fixed at an Elo rating of 1000.

Speech–gesture alignment. We use 20 unique segments. Each trial presents the same motion paired with its original audio and with audio from another segment of the same speaker. Mismatched audio is selected to have a similar duration. For each segment, the mismatched audio source is fixed across all evaluated methods, and the left/right positions of matched and mismatched alternatives are balanced. Participants select the pairing that fits better. This assesses correspondence in timing, emphasis, and speech content while keeping visual motion identical across alternatives.

Semantic appropriateness. We use 20 unique segments, each containing at least one annotated semantic event. Participants view a muted motion clip with two candidate sentences: its original transcript sentence and a semantically different sentence of similar length. The same sentence pair is used across all evaluated methods, and candidate positions are balanced. Text is displayed as

time-aligned subtitles. Participants select the sentence that better matches the meaning expressed by the gestures. This task evaluates semantic matching on event-containing segments, rather than an unfiltered sample of the test set.

Scoring and analysis. Both matching tasks use equal-weight scores: $+1$ for selecting the matched alternative, -1 for selecting the mismatched alternative, and 0 for a neutral response. The reported score is the mean over retained responses, including neutral responses in the denominator. Neither matching task uses preference-strength weighting. Zero indicates no net preference for the matched alternative.

Attention-check trials are used for quality control and excluded from scoring. We report participant-bootstrap 95% confidence intervals, resampling participants together with their associated responses and recomputing the matching scores or Elo ratings.

These intervals quantify participant-level uncertainty conditional on the selected stimuli.