REACTMOTION: GENERATING REACTIVE LISTENER MOTIONS FROM SPEAKER UTTERANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

When a friend says *I finally got the job*, we lean in, clap, or raise a hand; when the news is bad, we respond differently. A listener's body answers what is said, and we study how to generate this answer from the speaker's audio, transcript, and emotion. Two obstacles stand in the way: a recording shows only one of the many reactions an utterance can elicit, and a generic nod looks acceptable next to almost any utterance, so evaluation must measure whether a motion depends on its input. We address both. **ReactMotionNet** builds one-to-many supervision at scale: it curates an inventory of listener reactions, writes utterances that would elicit each one, cross-pairs conditions with candidate motions, and grades every pair with three language-model scorers calibrated on human judgments, yielding 85,142 condition–motion pairs, 71,286 of them supportive or socially appropriate. **ReactMotion** is a multimodal language model that predicts body-and-hand motion tokens directly from the speaker condition, without an intermediate motion caption, trained on the supportive pairs and refined with Direct Preference Optimization. **Gain** scores how much better a motion fits its own utterance than unrelated ones, so an input-blind generator scores zero. On 358 held-out utterances, ReactMotion reaches 0.3392 Gain after supervised fine-tuning and 0.3790 after preference optimization, against at most 0.1157 for twelve caption-mediated pipelines, and human raters prefer it to retrieval and cascade baselines.

1 INTRODUCTION

Conversation is a two-person performance. While one person speaks, the other listens with the whole body: a nod of agreement, a bow of thanks, hands raised in celebration, a step back in surprise. Figure 1 shows how open this choice is. Told *I'm so grateful for everything you've done*, one listener bows and another forms a heart with both hands, and both are right. An avatar or social robot that is to hold up its side of a conversation must therefore do more than move fluently: it must choose, among many acceptable responses, one that fits what the speaker said and how it was said (Ng et al., 2022; Ghosh et al., 2025; Ko et al., 2021).

This task differs from established settings. Text-to-motion models translate an explicit action description, such as *raise both hands*, into motion (Guo et al., 2022; Tevet et al., 2023; Zhang et al., 2023); co-speech gesture models animate the person who is speaking (Alexanderson et al., 2023; Ao et al., 2022; Liu et al., 2024). We instead generate the silent listener's body-and-hand reaction from the speaker's audio, transcript, and optional emotion label, with neither a description of the target motion nor the speaker's motion track.

The first obstacle is supervision. A conversation records one response per utterance, not the other reactions that would also have been valid. Recording many listeners for the same utterance is expensive, and treating the single observation as the only correct target turns valid alternatives into false negatives.

The second obstacle is evaluation. A generic nod looks natural next to almost any utterance while carrying no information about it. The benchmark must keep several valid responses per utterance and measure whether the generated motion depends on the intended condition.

We address both obstacles with a dataset built to expose many reactions per utterance. As Figure 2 shows, **ReactMotionNet** starts from a curated inventory of conversational reactions, writes speaker

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.

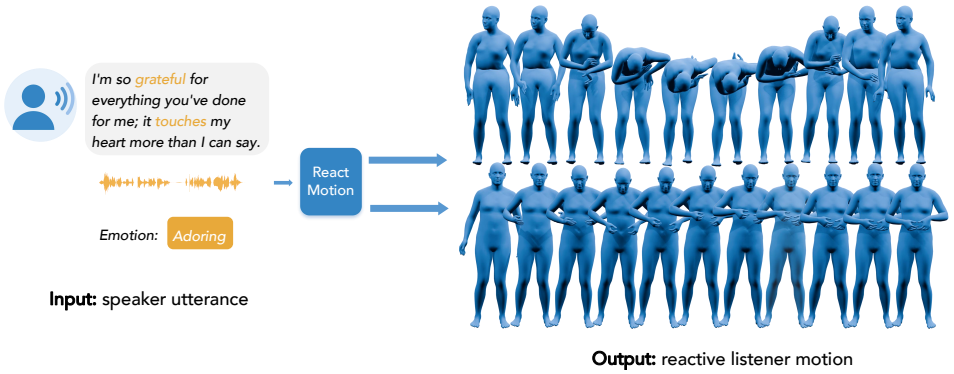

Figure 1: **One utterance can support multiple valid reactions.** The same expression of gratitude may elicit a bow (top) or a two-handed heart gesture (bottom).

utterances that would elicit each one, cross-pairs every utterance with several candidate motions, and grades each pair. A small human study fixes the score thresholds for supportive, socially appropriate, and incorrect pairs and a reliable preference margin; three language models then apply these thresholds to the whole corpus. The result is 85,142 condition–motion pairs, 71,286 of them graded supportive or socially appropriate, with lower-graded candidates as same-condition contrasts.

The grades feed directly into learning and evaluation. **ReactMotion** extends a multimodal language model with motion tokens and maps the speaker condition straight to body-and-hand motion. Same-condition score orderings give preference pairs for Direct Preference Optimization (DPO). A learned compatibility judge and the distractor-subtracted **Gain** metric measure whether a generated motion fits its own utterance better than unrelated ones.

Our contributions are: (1) ReactMotionNet, to our knowledge the first benchmark with graded one-to-many supervision for utterance-conditioned listener body motion, built with a human-calibrated grading pipeline that scales a few thousand human judgments to the full corpus; (2) ReactMotion, a direct utterance-to-motion generator with a unified motion vocabulary and preference refinement, which outperforms twelve caption-mediated pipelines by a wide margin and is preferred by human raters; and (3) an evaluation protocol for utterance-specific listener motion, built around Gain, that is validated against human labels and input-independent controls and complemented by retrieval, coverage, and human preference measures.

2 Related Work

From describing motion to eliciting a reaction. Text-conditioned motion generation learns mappings from explicit movement descriptions to motion, using discrete autoregressive models (Zhang et al., 2023; Jiang et al., 2023; Zhang et al., 2024; Guo et al., 2024) or diffusion models (Tevet et al., 2023; Chen et al., 2023). We build on discrete motion modeling, but change the meaning of the input: a speaker utterance is a cause for a possible reaction, not a description of that reaction. Caption-mediated baselines connect these two tasks explicitly, which lets us compare direct learning with this modular alternative.

Conversational and interactive motion. Co-speech gesture methods synthesize movement associated with the speaker's own speech (Alexanderson et al., 2023; Liu et al., 2024; Mughal et al., 2024). Listener modeling instead concerns the partner's response, including facial reactions (Ng et al., 2022; Luo et al., 2024; 2025) and listening behavior (Ghosh et al., 2025). Whole-body interaction methods generate paired motion from descriptions (Liang et al., 2024; Wang et al., 2024) or react to a partner's observed movement (Ghosh et al., 2024; Xu et al., 2024). Closest to our input, Audio2Photoreal (Ng et al., 2024) synthesizes face and body motion for both participants of a dyadic conversation from audio, and the Seamless Interaction dataset (Agrawal et al., 2025) provides large-scale recorded dyadic audiovisual data. Both model recorded dyads, where each utterance has one observed response. Our setting excludes the partner's motion, targets body-and-hand responses to an utterance, and

constructs graded alternative reactions per utterance; the constructed corpus therefore complements recorded interaction benchmarks.

Evaluation and preference learning. Motion benchmarks commonly measure distributional similarity, text–motion retrieval, diversity, and multimodality (Guo et al., 2022; Tevet et al., 2023). Retrieval evaluates input matching but depends on the candidate pool and learned embedding. The REACT challenges (Song et al., 2023) score multiple appropriate facial reactions against a set of appropriate references. We retain these views and add a distractor-subtracted compatibility score whose operating points—thresholds for supportive, socially appropriate, and incorrect reactions, and preference gaps—are selected from human studies and then distilled into a compact judge. We use DPO (Rafailov et al., 2023) to rank one-to-many alternatives rather than claiming the algorithm as a contribution, and because learned rewards can favor their own shortcuts (Christiano et al., 2017; Stiennon et al., 2020), we validate the judge against human labels and adversarial controls.

3 REACTMOTIONNET: GRADED SUPERVISION FOR LISTENER REACTIONS

Task and scope. Let $c = (A, T, E)$ denote the speaker condition, where A is the audio waveform, T is its transcript, and E is an optional categorical emotion label. The output $m \in \mathbb{R}^{L\times 623}$ is a listener motion of L frames, represented by 623 features over 52 body-and-hand joints at 20 fps. We learn the conditional distribution $p_\theta(m \mid c)$ because several different motions may be appropriate for the same condition. ReactMotionNet stores triples (c, m, s), where $s \in [0, 1]$ measures how well motion m responds to condition c. Multiple high-scoring motions provide alternative targets, while low-scoring motions provide informative contrasts. We model turn-level reactions; frame-accurate response onset, facial animation, and multi-turn dialogue are outside the present scope.

3.1 FROM MOTION ASSETS TO GRADED REACTION SUPERVISION

ReactMotionNet constructs the alternatives that recordings miss by separating two questions: which motions can function as listener reactions, and which speaker conditions can elicit each reaction. Figure 2 shows the four-stage pipeline; its output is a graded set of candidate motions under each condition, which supports one-to-many generation and same-condition preference learning.

1. Build a reusable reaction inventory. We standardize 52,074 captured and animated clips from sources including BONES-SEED (Bones Studio, 2026), Kimodo (Rempe et al., 2026), Motion-X and its IDEA400 subset (Lin et al., 2023), and HAA500 (Chung et al., 2021). Human annotators assess proposed action groups, and each decision propagates to the group's clips: 4,772 group-level annotations yield verdicts for 22,430 clips and retain 10,795 source clips. Low-motion filtering and caption revision leave motions with observable body-and-hand descriptions.

2. Construct eliciting speaker conditions. Given a reaction description, GPT-4o (Hurst et al., 2024) proposes an utterance and emotion that could elicit that reaction. Naturalness filtering reduces 47,689 proposals to 27,392, after which a commonness filter removes unlikely situations. Speech synthesis completes the speaker condition $c = (A, T, E)$.

3. Expand each condition into candidate reactions. Two retrieval channels cross-pair conditions and motions. The *elicitation* channel retrieves conditions through a hypothetical trigger for the reaction; the *topical* channel retrieves them directly from the motion caption. Their different semantics broaden the candidate set, so several motions are compared under one condition and one motion under several conditions.

4. Convert compatibility scores into graded supervision. Gemma 3 12B (Gemma Team et al., 2025), Qwen3-30B-A3B (Yang et al., 2025), and Qwen3.6-27B (Qwen Team, 2026) independently score each candidate pair from the utterance, its emotion, and the normalized body-and-hand caption of the motion; their mean defines $s \in [0, 1]$ (Appendix A explains why scoring reads the caption rather than rendered frames). Human judgments select the operating points (Section 3.2): supportive (gold) if $s \geq 0.67$, socially appropriate (silver) if $0.18 \leq s < 0.67$, and incorrect (negative) if $s < 0.18$. Gold pairs are generation targets, silver pairs retain valid alternatives, and negative pairs

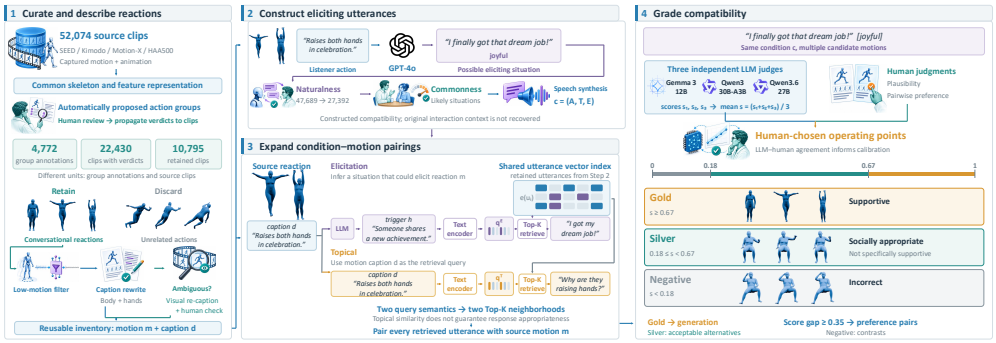

Figure 2: **From motion assets to graded one-to-many supervision.** The four stages build a reusable reaction inventory, construct eliciting speaker conditions, expand condition–motion candidates through two retrieval channels, and convert three-model compatibility scores into human-calibrated tiers. The tiers provide generation targets, acceptable alternatives, and same-condition preference contrasts.

provide contrasts. A score gap of at least 0.35 defines a preference pair under the same condition. Human calibration therefore specifies the supervision, while automatic scoring applies it at corpus scale.

Table 1: **ReactMotionNet statistics.** Pairings = gold (supportive) + silver (socially appropriate) + recorded (captured dyadic reactions without grades, excluded from all training; Appendix E.6). Negative candidates are not counted. "Gr. mot." counts unique graded-motion identities, with Total their union across splits; "Asset-h" sums clip durations over pairings, with assets reused.

Split	Utter.	Pairings	Gold	Silver	Recorded	Gr. mot.	Asset-h	Words	Voices	Emot.
Train	13,787	81,618	14,648	53,114	13,856	657	123.1	15.3	11	48
Validation	361	1,467	304	1,163	–	558	2.2	16.1	2	41
Test	358	2,057	513	1,544	–	593	3.1	15.6	2	40
Total	14,506	85,142	15,465	55,821	13,856	657	128.4	15.3	15	48

Splits and evaluation targets. Table 1 summarizes 85,142 split pairings. Supportive (gold) and socially appropriate (silver) examples appear in every split. The 13,856 recorded reactions carry no compatibility grade, are excluded from all training, and serve only the evaluation of Appendix E.6; supervised fine-tuning (SFT) uses the 14,648 supportive gold pairs. Incorrect candidates ($s < 0.18$) provide lower-ranked alternatives for preference and judge training. SFT, DPO, and judge fitting use only the training partition; validation supports model selection, and the test partition is used only for final evaluation. The main evaluation uses one paired reference for each of 358 held-out utterances. Utterances are disjoint across splits, while the reaction inventory is shared: all 593 test and 558 validation motion identities also occur among the 657 training identities, so the test asks whether a model produces the right reaction for a new utterance. Section 5.6 verifies that this does not depend on having seen a reaction as an SFT target.

The inventory covers 48 emotion categories and 15 named synthesis-speaker identities, which are disjoint across splits (Appendix A).

3.2 Human Priors for Automatic Scoring

The human study defines how automatic scores become supervision. Eleven annotators provide 5,996 judgments on 935 sampled pairs. Repeated quality-control comparisons, presented in randomized order, measure each annotator's self-consistency, and calibration uses the 3,024 decisive judgments of the four annotators who exceed 95% consistency; their individual curves agree with one another (Figure 3, center). Among these judgments, the fraction rated plausible increases from 38.7% in the incorrect tier to 66.2% in the socially appropriate tier and 87.1% in the supportive tier, supporting thresholds at 0.18 and 0.67.

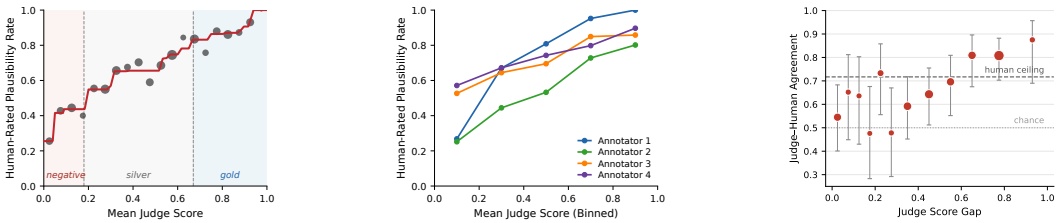

Figure 3: **Human calibration of automatic compatibility scores.** Left: the fraction judged plausible increases with score; dashed lines show the selected tier thresholds. Center: the same monotonic relationship appears for each retained annotator. Right: agreement with the automatic ordering increases as the score difference grows, supporting the 0.35 minimum gap used for preference pairs.

For pairwise choices, agreement between the automatic ordering and human preference grows with the score gap, from 0.55 for gaps below 0.05 to 0.88 for gaps above 0.85 (0.679 over all 458 comparisons); we therefore form preference pairs only when the gap is at least 0.35. People define the categories and the preference margin on a manageable sample; the three-model score applies them to the full candidate set. Figure 3 shows the pooled and per-annotator evidence, and Appendix A.3 reports leave-one-annotator-out agreement, which bounds the agreement any automatic scorer can attain on these labels. Section 5.4 separately evaluates generated motions with human raters.

4 REACTMOTION: DIRECT UTTERANCE-TO-MOTION LEARNING

ReactMotion learns directly from the graded condition–motion pairs; it never translates the utterance into a motion caption. We build on Qwen2.5-Omni-3B (Xu et al., 2025) (Omni-3B), which jointly represents audio and text, and extend its autoregressive vocabulary with body, hand, duration, and control tokens, so one transformer reads the speaker condition and predicts the listener motion. Direct Preference Optimization (DPO) (Rafailov et al., 2023) then uses same-condition graded pairs to raise the relative probability of the better-matched reaction.

4.1 CONDITIONING AND MOTION REPRESENTATION

The pretrained audio encoder maps waveform A to continuous acoustic features. The transcript T, emotion label E, and task instruction are tokenized as text, and the multimodal transformer jointly processes these representations. Motion symbols are added as new token classes in the same autoregressive output space. We apply low-rank adaptation (LoRA) (Hu et al., 2022) to the transformer projections and optimize the complete input-embedding matrix and output projection, while the remaining backbone and audio encoder stay fixed. The resulting model has 716.7M trainable parameters; Appendix B reports the component-wise parameter counts.

Following the body–hand hierarchy introduced by H2VQ (Lu et al., 2024), our tokenizer represents the body and hands separately and extends each component with three residual codebooks of 512 entries. At quantized time step n, the six discrete indices are

$$x_n = [b_{0,n}, b_{1,n}, b_{2,n}, h_{0,n}, h_{1,n}, h_{2,n}]. \tag{1}$$

Here $b_{r,n}$ and $h_{r,n}$ are the body and hand code indices, respectively; $r \in \{0, 1, 2\}$ identifies the residual-codebook level, and $n \in \{1, \ldots, N\}$ identifies the quantized time step. Each level encodes the residual left by earlier levels, and distinct token namespaces preserve component identity (body or hand) and residual level.

The target sequence contains a duration token, a start marker, the six codes from x_1 through x_N in temporal order, an end marker, and an end-of-sequence token. The duration specifies N quantized steps rather than the number of decoded 20-fps frames. Let $y = (y_1, \ldots, y_M)$ denote this complete serialized sequence of M tokens. The model assigns its probability autoregressively:

$$p_\theta(y \mid c) = \prod_{t=1}^{M} p_\theta(y_t \mid y_{<t}, c). \tag{2}$$

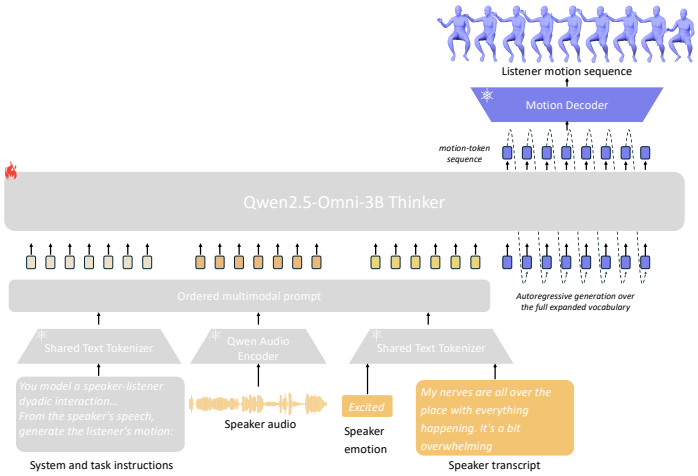

Figure 4: **ReactMotion architecture.** A pretrained audio encoder represents the speaker waveform; transcript, emotion, and task instructions are represented as text tokens. The adapted multimodal transformer predicts motion tokens autoregressively, and the frozen H2VQ decoder (Lu et al., 2024) reconstructs the listener motion. The flame identifies the trainable LoRA parameters, token embeddings, and output projection.

Here y_t is the token at sequence position t, $y_{<t} = (y_1, \ldots, y_{t-1})$ is its preceding prefix, c is the speaker condition, and θ denotes the trainable generator parameters. A frozen H2VQ decoder maps the predicted codes back to continuous motion, with the hand decoder additionally conditioned on the body representation. On 2,616 held-out files, the tokenizer obtains body/hand mean-absolute-error ratios of 0.247/0.248 relative to a mean-prediction baseline, and every codebook entry is used (Appendix E.4).

4.2 Supervised and Preference-Based Training

Learning the one-to-many mapping. We first learn the motion vocabulary from caption–motion pairs. This stage initializes motion generation and also supplies the decoder used by the matched caption-cascade baseline. Supervised fine-tuning (SFT) then uses 14,648 supportive condition–motion pairs from ReactMotionNet. Let $\mathcal{D}_{\text{gold}}$ denote this training set and let S specify the available subset of audio, transcript, and emotion cues. The SFT objective is

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(c,m)\sim\mathcal{D}_{\text{gold}},\, S} \sum_{t=1}^{M(m)} \log p_\theta(y_t(m) \mid y_{<t}(m), c_S). \tag{3}$$

Here $y(m)$ is the serialized token sequence for motion m, $M(m)$ is its sequence length, $y_t(m)$ is its t-th token, and c_S is condition c restricted to cues in S. Each supportive motion is a separate example, so the likelihood covers several valid responses to the same utterance without averaging them.

During SFT, full conditioning is selected with probability $q = 0.667$; otherwise, one of $\{A, T, A{+}T, A{+}E, T{+}E, A{+}T{+}E\}$ is sampled uniformly. Consequently, the marginal probability of full conditioning is $q + (1-q)/6 \approx 0.723$. Preference optimization uses all three cues.

Measuring input specificity. A separate compatibility judge outputs $J(m, c) \in [0, 1]$, where a larger value indicates that motion m is a more appropriate response to condition c. The judge is trained on the automatically scored corpus after applying the human-selected thresholds; the human-rated calibration pairs are not judge-training examples. Two independent checks anchor the pipeline to human judgment: the thresholds and preference margin are chosen from human ratings (Section 3.2), and generated motions are rated by humans (Section 5.4). Raw compatibility alone can reward a generic motion that fits many utterances. We therefore compare each generated motion with its

intended condition and with unrelated conditions:

$$G_i = J(m_i, c_i) - \frac{1}{K} \sum_{d \in D_i} J(m_i, c_d), \qquad \text{Gain} = \frac{1}{n_{\text{eval}}} \sum_{i=1}^{n_{\text{eval}}} G_i. \tag{4}$$

For evaluation item i, m_i is the generated motion, c_i is its intended condition, and D_i indexes $K = 48$ distractor conditions c_d drawn from the other held-out utterances; the intended condition is never its own distractor. Gain averages G_i over the n_{eval} evaluated utterances. If a generator ignores its input, intended and distractor compatibility have the same expectation, so expected Gain is zero. Positive Gain therefore measures utterance-specific compatibility; naturalness is assessed separately by human raters.

Refining preferences. Starting from the SFT model, DPO increases the relative likelihood of a preferred motion m^+ over a less preferred motion m^- for the same condition:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E} \log \sigma \left(\beta \left[\log \frac{p_\theta(m^+ \mid c)}{p_{\text{ref}}(m^+ \mid c)} - \log \frac{p_\theta(m^- \mid c)}{p_{\text{ref}}(m^- \mid c)} \right] \right). \tag{5}$$

The expectation is over preference triples (c, m^+, m^-) from the training partition. Here $p_\theta(m \mid c)$ abbreviates the probability of serialized sequence $y(m)$, p_{ref} is the fixed SFT reference policy, σ is the logistic function, and β controls preference strength. Preference pairs are defined by ordering and score difference rather than by tier, so a pair need not be gold versus negative. The training partition provides 19,206 comparisons with an automatic-score difference of at least the human-selected margin 0.35; the principal configuration uses $\beta = 0.3$. Neither preference training nor judge fitting uses the test partition. Because the preference labels and the judge share the calibrated scoring pipeline, Section 5.4 additionally evaluates generated motions with human raters. Appendices B and E.2 report optimization details and parameter sensitivity.

5 Experiments

5.1 Setup and Evaluation

Protocol. All main automatic comparisons use the same 358 held-out utterances and a fixed judge; generator and judge fitting never see test items. ReactMotion uses rank-32 LoRA and supportive-pair SFT (Appendix B). Four captioners—Omni-3B with audio, Omni-3B with $A{+}T{+}E$, and Phi-4 (Abdin et al., 2024) or Qwen3.6-27B with transcript and emotion—provide descriptions to our caption-to-motion model, Kimodo (Rempe et al., 2026), or a 783-clip retrieval library, yielding twelve alternatives (Appendix D).

Complementary measurements. Gain asks whether a motion fits its intended utterance better than unrelated ones. R@1 and R@3 ask whether the intended condition can be retrieved from 32 candidates. J-FID compares generated and paired-reference motion distributions in the judge embedding space; precision measures how often generated motions lie near the reference distribution, and recall how much of it the generator covers. Diversity compares motions across utterances and guards against collapse. Multimodality (MM) compares ten samples for each of 100 fixed utterances and measures within-condition variation; it is zero by construction for deterministic top-1 retrieval (Appendix C).

5.2 Validating the Learned Evaluator

Table 2 examines three requirements on the 17.1M-parameter judge: agreement with human judgments, higher scores for stronger pairings, and near-zero Gain for motions that do not depend on the input.

The judge agrees with the screened human majority on 78.8% of pairs versus 72.8% for the majority-class predictor (AUC 0.775 versus 0.500), which is the level at which the annotators agree with one another (Appendix A.3). Mean scores follow the intended order—0.807 supportive, 0.707 socially appropriate, and 0.412 incorrect—while shuffled, constant, and random controls have near-zero Gain. Section 5.4 adds human appropriateness and naturalness judgments.

Table 2: **Three evaluations of the learned judge.** (a) Agreement with the screened human majority on 633 corpus pairs, with a constant majority-class predictor as reference. (b) Mean scores for 631 supportive, socially appropriate, and incorrect pairs. (c) Mismatched and input-independent controls on 358 held-out utterances. Appendix A.3 provides the complete human-label comparison.

(a) Human agreement

Predictor	Agree.↑	AUC↑
Majority class	0.728	0.500
Learned judge	**0.788**	**0.775**

(b) Tier separation

Tier	Gain	J	n
Supportive	0.472	0.807	144
Socially appr.	0.386	0.707	334
Incorrect	0.168	0.412	153

(c) Specificity controls

Control	Gain	J
Paired reference	0.4032	0.7309
Shuffled pairing	0.0078	0.3385
Constant motion	−0.0012	0.3015
Random motion	0.0012	0.2973

5.3 Direct Generation and the Preference Trade-off

Does direct generation outperform caption mediation? Across twelve caption-mediated systems, direct SFT obtains 0.3392 Gain and 0.5894 R@3 versus best values of 0.1157 and 0.1983, respectively. Changing the decoder does not close the gap: Kimodo reaches at most 0.0517 Gain. The random-library control pairs reference-like diversity with near-zero Gain: distributional similarity does not imply utterance specificity. Learning directly from condition–motion supervision is thus more effective than routing the condition through a generated caption (Appendix D).

What does preference optimization change? DPO raises Gain from 0.3392 to 0.3790 (paired bootstrap 95% interval of the difference $[0.0158, 0.0635]$), R@3 from 0.5894 to 0.6145, and precision/recall from 0.930/0.858 to 0.958/0.869, while MM decreases from 0.978 to 0.843. Preference refinement therefore sharpens the conditional distribution toward utterance-specific reactions at the cost of some within-condition variation. Appendix E.2 reports the sensitivity to β, and Section 5.4 checks the result with human raters.

Cue configurations. Cue ablations (Appendix E.3) show that the transcript and the emotion label carry complementary semantic signal, and that modality dropout makes the model robust to missing cues: withholding audio at inference leaves R@3 essentially unchanged, and the transcript alone retains most of the retrieval accuracy. Preference refinement improves over SFT under every cue subset.

5.4 Human Evaluation of Generated Motions

Twenty screened annotators compare SFT+DPO with retrieval, cascade generation, recorded motion, and SFT on 269 items (2,560 retained judgments). Each trial shows the condition and two anonymously rendered reactions; raters judge appropriateness and naturalness separately, with a tie option.

SFT+DPO is preferred to retrieval and cascade generation in appropriateness (59.9% versus 33.3%; 60.9% versus 33.8%) and naturalness (52.8% versus 18.9%; 53.2% versus 26.2%). Recorded motion leads in appropriateness (51.7% versus 41.5%), whereas naturalness is close (36.4% versus 34.7%, with 28.8% ties). SFT+DPO stays on par with SFT in perceived appropriateness (44.7% versus 46.0%) and naturalness (28.1% versus 33.6%, with 38.3% ties), so preference refinement improves specificity at no perceived cost. Failures are rare and concentrated: SFT+DPO has Gain ≤ 0 on only 7.0% of held-out utterances, mostly those on which even the paired reference scores low (Appendix F).

5.5 Initialization and Reaction-Identity Generalization

Is caption–motion initialization necessary? We compare matched-capacity models with and without caption–motion initialization. Without it, SFT obtains higher Gain (0.3649 versus 0.3392), R@1, mean reciprocal rank (MRR), and J-FID, whereas initialization yields higher recall (0.858 versus 0.832); DPO Gain is similar (0.3790 versus 0.3708). Initialization therefore changes the specificity–coverage balance rather than enabling task learning. We keep the initialized model as primary because it shares its stage-one decoder with the caption-cascade baseline (Appendix E).

Table 3: **Main results** on 358 utterances. Bold marks the best value among the evaluated systems, excluding the reference and the random control; Div./MM are interpreted with Gain rather than ranked. MM is zero by construction for deterministic top-1 retrieval.

System / captioner	Gain↑	R@1↑	R@3↑	J-FID↓	Prec.↑	Rec.↑	Div.	MM
Random library motion	0.0012	0.0279	0.0838	0.1304	0.869	0.897	1.348	1.347
Paired reference	0.4032	0.5559	0.8212	0.0000	1.000	1.000	1.353	–
Cascade: caption → our stage-one decoder								
Omni-3B (A)	0.0225	0.0223	0.0922	0.6769	0.838	0.385	1.051	0.996
Omni-3B (A+T+E)	0.0188	0.0279	0.0978	0.6817	0.874	0.316	1.054	1.019
Phi-4	0.0224	0.0475	0.1285	0.5252	0.779	0.363	1.215	1.070
Qwen3.6-27B	0.0362	0.0503	0.1201	0.5027	0.774	0.385	1.235	1.096
Cascade: same caption → Kimodo								
Omni-3B (A)	0.0485	0.0447	0.1089	0.5074	0.821	0.179	1.205	0.884
Omni-3B (A+T+E)	0.0351	0.0531	0.1201	0.5146	0.855	0.271	1.169	0.904
Phi-4	0.0517	0.0503	0.1257	0.4162	0.735	0.148	1.247	0.911
Qwen3.6-27B	0.0498	0.0419	0.1369	0.4254	0.696	0.226	1.249	0.892
Retrieval: caption → motion library								
Omni-3B (A)	0.0464	0.0698	0.1676	0.5595	0.453	0.629	1.205	0.000
Omni-3B (A+T+E)	0.0627	0.0587	0.1788	0.4981	0.489	0.704	1.201	0.000
Phi-4	0.0871	0.0670	0.1648	0.3260	0.696	0.726	1.313	0.000
Qwen3.6-27B	0.1157	0.0838	0.1983	0.3226	0.799	0.522	1.309	0.000
Direct utterance-to-motion generation								
ReactMotion SFT	0.3392	0.3101	0.5894	**0.0824**	0.930	0.858	1.335	0.978
SFT+DPO	**0.3790**	**0.3715**	**0.6145**	0.0966	**0.958**	**0.869**	1.326	0.843

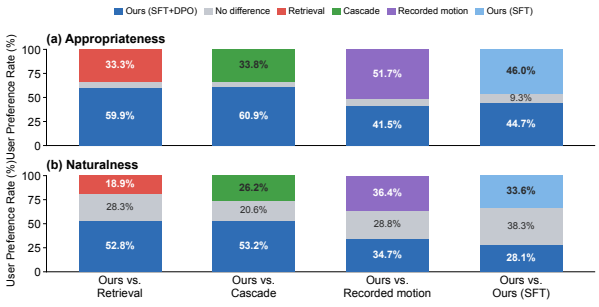

Figure 5: **Human pairwise preference.** SFT+DPO versus each alternative (win / tie / alternative).

5.6 Controlling for Direct Reaction-SFT Exposure

We remove every SFT pair containing 200 designated identities, leaving 9,952 pairs. Gain remains 0.3218 on designated-reference cases and 0.3154 on others ($p = 0.73$), so the model learns the conditional mapping without reaction-SFT examples of those identities (Appendix E.5). On 462 utterances from recorded conversations, the ordering SFT+DPO > SFT > input-independent controls is preserved (Appendix E.6).

6 Discussion and Conclusion

Listener reactions are one-to-many, yet ordinary recordings show one response per utterance. This paper turns that gap into a benchmark, a model, and a metric. ReactMotionNet builds graded one-to-many supervision at scale from a curated reaction inventory, elicited utterances, cross-paired candidates, and human-calibrated compatibility grades. ReactMotion maps the speaker's audio, transcript, and emotion directly to body-and-hand motion tokens, with no intermediate caption, and DPO on same-condition preferences sharpens its reactions. Gain measures whether a motion fits its own utterance better than unrelated ones, validated against human labels and input-independent controls. Direct generation outperforms twelve caption-mediated pipelines by a wide margin, is preferred by human raters, and keeps its ordering on recorded conversation.

AI USE STATEMENT

Generative models were used to construct the dataset, including utterance generation, caption normalization, speech synthesis, motion synthesis, and automatic preference labeling, as described in Section 3 and Appendix A. An AI coding and writing assistant was also used to reorganize and edit the manuscript, check consistency between reported results and claims, assist with interpretation and experimental recommendations, and compile and inspect the LaTeX document. These uses do not constitute independent experimental verification. The authors are responsible for verifying the source data, experimental claims, references, and final submission.

ETHICS STATEMENT

The corpus encodes judgments about plausible listener behavior, not a universal standard of socially correct behavior. Such judgments may depend on culture, context, disability, and individual preference; generated reactions should not be used to infer a person's true feelings or intentions. Synthetic speech and motion may also support misleading representations of conversational behavior.

Human annotation. Two human studies were conducted: the calibration study with 11 annotators (Section 3.2) and the preference study with 20 annotators (Section 5.4). Annotators were recruited through the university and compensated for their time. Before annotating, each annotator received a written manual and a training session that explained the task and the content they would see, and proceeded to the full study only after passing a small qualification set. Annotation ran on a web platform built for the study: items were presented in randomized order and sampled uniformly across automatic-score intervals, so that every annotator saw a balanced range of candidates. No personally identifying information was collected, and the rendered stimuli show anonymous body meshes.

Source assets and release. The reaction inventory is derived from Motion-X and its IDEA400 subset, HAA500, Kimodo, BONES-SEED, and smaller animation collections. Reuse of these assets and the dataset release follow the license of each source.

REPRODUCIBILITY STATEMENT

The appendices document the corpus views, human calibration, architecture, optimization settings, evaluator, metric sampling, baseline grid, and controlled studies. We will release the ReactMotionNet data (split files and graded pairings), the code for construction, training, and evaluation, the H2VQ tokenizer, and the model checkpoints, together with the exact model revisions used for the reported numbers.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report. arXiv preprint arXiv:2412.08905, 2024.

Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, et al. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. arXiv preprint arXiv:2506.22554, 2025.

Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)*, 42 (4):1–20, 2023.

Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics (TOG)*, 41(6):1–19, 2022.

Bones Studio. BONES-SEED: Skeletal everyday embodiment dataset. Hugging Face dataset, https://huggingface.co/datasets/bones-studio/seed, 2026. Optical motion-capture dataset of approximately 288 hours.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18000–18010, 2023.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

Jihoon Chung, Cheng-hsin Wuu, Hsuan-ru Yang, Yu-Wing Tai, and Chi-Keung Tang. Haa500: Human-centric atomic action dataset with curated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13465–13474, 2021.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. arXiv preprint arXiv:2503.19786, 2025.

Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Remos: 3d motion-conditioned reaction synthesis for two-person interactions. In *European Conference on Computer Vision (ECCV)*, pp. 418–437, 2024.

Bishal Ghosh, Emma Li, and Tanaya Guha. Active listener: Continuous generation of listener's head motion response in dyadic interactions. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025. doi: 10.1109/ICASSP49660.2025.10889429.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5152–5161, 2022.

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 (CVPR)*, pp. 1900–1910, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card, 2024.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. MotionGPT: Human motion as a foreign language. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, and Jaehong Kim. Air-act2act: Human–human interaction dataset for teaching non-verbal social behaviors to robots. *The International Journal of Robotics Research*, 40(4-5):691–697, 2021.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, 132(9):3463–3483, 2024. doi: 10.1007/s11263-024-02042-6.

Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1144–1154, 2024.

Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. HumanTOMATO: Text-aligned whole-body motion generation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32939–32977. PMLR, 2024. URL https://proceedings.mlr.press/v235/lu24b.html.

Cheng Luo, Siyang Song, Weicheng Xie, Micol Spitale, Zongyuan Ge, Linlin Shen, and Hatice Gunes. Reactface: Online multiple appropriate facial reaction generation in dyadic interactions. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–18, 2024.

Cheng Luo, Siyang Song, Siyuan Yan, Zhen Yu, and Zongyuan Ge. Reactdiff: Fundamental multiple appropriate facial reaction diffusion model. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pp. 5607–5616, 2025.

Muhammad Hamza Mughal, Rishabh Dabral, Ikhsanul Habibie, Lucia Donatelli, Marc Habermann, and Christian Theobalt. Convofusion: Multi-modal conversational diffusion for co-speech gesture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1388–1398, 2024.

Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20395–20405, 2022.

Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. From audio to photoreal embodiment: Synthesizing humans in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1001–1010, 2024.

Qwen Team. Qwen3-VL. Official model release, https://github.com/QwenLM/Qwen3-VL, 2025.

Qwen Team. Qwen3.6-27B model card. Official model card, April 2026. URL https://huggingface.co/Qwen/Qwen3.6-27B.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Davis Rempe, Mathis Petrovich, Ye Yuan, Haotian Zhang, Xue Bin Peng, Yifeng Jiang, Tingwu Wang, Umar Iqbal, David Minor, Michael de Ruyter, Jiefeng Li, Chen Tessler, Edy Lim, Eugene Jeong, Sam Wu, Ehsan Hassani, Michael Huang, Jin-Bey Yu, Chaeyeon Chung, Lina Song, Olivier Dionne, Jan Kautz, Simon Yuen, and Sanja Fidler. Kimodo: Scaling controllable human motion generation. arXiv preprint arXiv:2603.15546, 2026.

Siyang Song, Micol Spitale, Cheng Luo, Germán Barquero, Cristina Palmero, Sergio Escalera, Michel Valstar, Tobias Baur, Fabien Ringeval, Elisabeth André, and Hatice Gunes. REACT2023: The first multiple appropriate facial reaction generation challenge. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9620–9624, 2023.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:3008–3021, 2020.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SJ1kSyO2jwu.

Zhenzhi Wang, Jingbo Wang, Yixuan Li, Dahua Lin, and Bo Dai. Intercontrol: Zero-shot human interaction generation by controlling every joint. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Liang Xu, Yizhou Zhou, Yichao Yan, Xin Jin, Wenhan Zhu, Fengyun Rao, Xiaokang Yang, and Wenjun Zeng. Regennet: Towards human action-reaction synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1759–1769, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. In *The Association for the Advancement of Artificial Intelligence (AAAI)*, volume 38, pp. 7368–7376, 2024.

Appendix organization. The supplementary material follows the paper's evidence chain: corpus construction and human calibration; generator representation and optimization; evaluator and metric definitions; complete baseline results; controlled studies; and qualitative examples.

A Dataset Construction and Human Calibration

A.1 Corpus Views and Split Statistics

The generator split inventory is the 85,142 pairings in Table 1: training $81{,}618 = 14{,}648$ gold $+53{,}114$ silver $+13{,}856$ ungraded recorded reactions; validation $1{,}467 = 304 + 1{,}163$ gold/silver; test $2{,}057 = 513 + 1{,}544$ gold/silver. The recorded training reactions have no compatibility score and are not used to train the reported generators. Negative-tier candidates used for judge and preference construction are a separate inventory; they are not added to 85,142 and are not the 13,856 recorded clips. The 358-utterance generation grid is unchanged.

The learned judge is fitted on the training data, uses the validation split for validation, and does not use the test split for training. Its corpus labels remain distinct from the screened human ratings used to select the operating points.

The audited inventory contains 48 emotions and 15 named synthesis-speaker identities, together with an unknown-identity marker. Eleven named identities occur only in training, two only in validation, and two only in testing. These split-disjoint counts refer to the available synthesis-identity labels.

Motion-identity overlap is complete across the graded splits: all 593 test and 558 validation identities occur among the 657 training identities. At the pairing level, 2,044 of 2,057 positive-tier test pairings have a complete six-stream motion also in the training pool. These denominators describe the gold/silver view in Table 1. Appendix E.5 reports targeted retraining that evaluates this dependence. Duration summaries ("Asset-h" in Table 1) reuse motion assets across pairings and therefore do not represent unique captured hours.

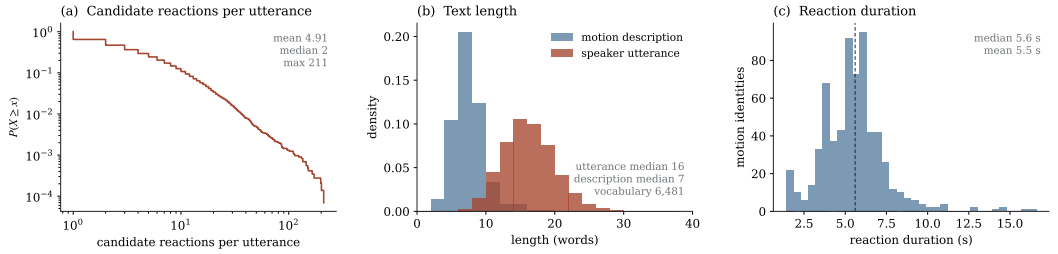

Figure 6: **Corpus profiles supporting one-to-many construction.** Left: candidate reactions per utterance have mean 4.91, median 2, and maximum 211. Center: speaker utterances contain a median of 16 words, compared with 7 for motion descriptions. Right: reactions have median duration 5.6 seconds.

Figure 6 characterizes the supervision produced by the construction pipeline. The long-tailed candidate count supplies multiple alternatives for many conditions and reaches 211 candidates for the most extensively expanded utterances. Speaker utterances are linguistically richer than the short action descriptions used to retrieve motions, while the reaction-duration distribution remains centered on the turn-level behavior modeled in this work.

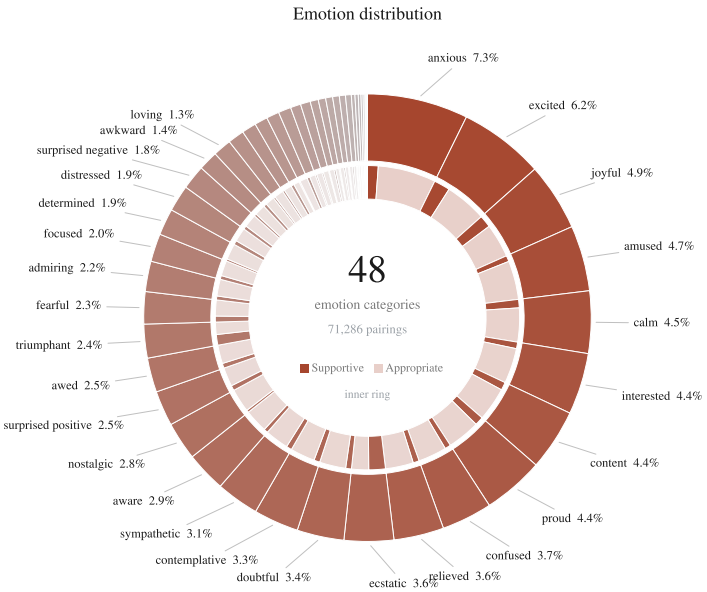

Figure 7: **Emotion coverage of the graded corpus.** The outer ring distributes 71,286 supportive and socially appropriate pairings across 48 emotion categories; the inner ring shows the two supervision tiers within each category. The largest category accounts for 7.3% of pairings, and the five largest together account for 27.6%.

Figure 7 shows that no single emotion dominates the graded corpus. The 48-category inventory spans frequent conversational states as well as less common affective contexts, providing varied conditioning labels for learning utterance-specific reactions. The figure reports coverage rather than claiming a uniform category distribution.

A.2 Construction Procedure

1. Motion curation. The source inventory contains 52,074 clips from BONES-SEED (Bones Studio, 2026), Kimodo (Rempe et al., 2026), Motion-X and its IDEA400 subset (Lin et al., 2023), HAA500 (Chung et al., 2021), and smaller animation/performance collections. All are mapped to the common body-and-hand representation. Screening is performed at the action-group level: an automatic proposal is reviewed by a person, and a group decision is propagated to related clips. Construction records contain 4,772 group-level human annotations, which propagate to verdicts for 22,430 proposed clips and retain 10,795 clips.

A low-motion filter uses mean frame-to-frame change over 619 non-root feature dimensions, with threshold 0.0010, retaining 96.5% of the screened collection. Captions are normalized to observable body/hand movement and omit facial details not represented by the skeleton. The caption-processing report lists 12,607 action groups: 9,587 accepted after text rewriting and 3,020 routed to visual re-captioning using rendered front/side views. Human checking corrected 6.7% of normalized captions.

2. Speaker-condition synthesis. GPT-4o (Hurst et al., 2024) proposes utterances and intended emotions that could elicit a motion. The initial pool contains 47,689 distinct utterances. A conversational-naturalness filter retains 27,392 at a score threshold of 0.5. An additional commonness filter uses three model scores and requires each to be at least 0.5. Speech is synthesized from the retained text and emotion. This produces possible eliciting conditions rather than recovered recordings of the original motion event.

3. Candidate retrieval. An elicitation channel retrieves utterances similar to hypothetical triggers of a motion; a topical channel matches utterances directly to motion captions. Both use sentence

embeddings. The construction summary reports 101,287 deduplicated candidate pairings over 657 normalized reaction descriptions before final grading and subset preparation. Topical similarity alone need not imply that the motion is an appropriate reaction.

4. Graded supervision. Gemma 3 12B (Gemma Team et al., 2025), Qwen3-30B-A3B (Yang et al., 2025), and Qwen3.6-27B (Qwen Team, 2026) independently score each candidate from the utterance text, its emotion label, and the normalized motion caption; their mean defines $s \in [0, 1]$. Scoring reads the caption rather than the motion itself because the captions are normalized to observable body-and-hand actions and human-checked (Step 1), so they describe each reaction precisely. In a pilot we also scored pairs with a vision–language model (Qwen3-VL (Qwen Team, 2025)) given rendered keyframes of the motion together with the caption; its judgments were noticeably less reliable, since keyframe interpretation added errors that the caption does not carry, so the released grades use caption-level scoring. Human plausibility and pairwise judgments determine the operating points: gold ($s \geq 0.67$) is supportive, silver ($0.18 \leq s < 0.67$) is socially appropriate, negative ($s < 0.18$) is incorrect, and a minimum difference of 0.35 defines the principal preference set. SFT uses supportive pairs, while preference construction also uses lower-scoring candidates. The prompts for utterance generation, compatibility scoring, and captioning, the model versions, and the intermediate filtering manifests are released with the code.

A.3 Human-Label Evidence and Its Scope

Figure 3 shows the pooled, per-annotator, and score-gap evidence used to choose these operating points. It calibrates the corpus criteria, whereas Table 2(a) evaluates the learned judge that represents them. Quality-control comparisons are repeated with randomized presentation order, and screening is based on within-annotator consistency rather than agreement with automatic rankings.

The corpus-label study contains 935 sampled pairs and 5,996 judgments from 11 annotators; the four annotators who satisfy the 95% self-consistency criterion contribute the 3,024 decisive labels used for calibration. The retained A/B subset has 458 comparisons and 0.679 agreement with score ordering. Agreement increases from 0.55 for score gaps below 0.05 to 0.88 for gaps above 0.85, supporting the use of a minimum preference margin. The per-annotator curves in Figure 3 provide a direct check that the monotone score–plausibility relationship is not produced by pooling alone.

Figure 8 reports, for each retained annotator, how often the majority of the other three annotators and how often the learned judge agree with that annotator's held-out labels. Peer-majority agreement is the practical ceiling for any automatic scorer on these labels, and the judge reaches it on average.

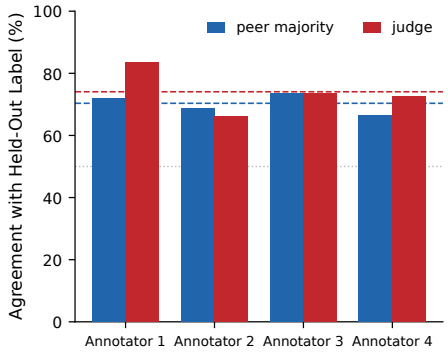

Figure 8: **Leave-one-annotator-out agreement.** For each retained annotator, agreement of the peer majority (blue) and of the learned judge (red) with that annotator's labels; dashed lines are means, the dotted line is chance. Peer-majority agreement bounds what an automatic scorer can attain on these labels.

The binary-majority evaluation of the learned judge contains 633 pairs and uses a different analysis set from the A/B comparisons. The tier-stratified analysis contains 631 pairs ($144 + 334 + 153$).

Table 4: Human-majority agreement and tier-stratified judge diagnostics (gold = supportive, silver = socially appropriate, negative = incorrect). These are corpus-pair evaluations, not human comparisons of generated systems.

Predictor	Agree.↑	AUC↑	Tier	Gain	J	n
Majority class	0.728	0.500	Gold (supp.)	0.472	0.807	144
Learned judge	**0.788**	**0.775**	Silver (appr.)	0.386	0.707	334
			Neg. (inc.)	0.168	0.412	153

Agreement of the learned judge under cue removal is 0.766 (A), 0.776 (T), 0.782 (A+T), 0.779 (A+E), 0.785 (T+E), and 0.788 (A+T+E). This evaluates the *judge's* input robustness; it is distinct from withholding cues from the generator.

B Model, Motion Representation, and Optimization

This section specifies the generator, serialized motion vocabulary, supervised fine-tuning, and preference optimization. The tokenizer and evaluation judge are separate models and are not included in the generator's trainable-parameter count.

B.1 Generator Configuration

Table 5: Reported SFT configuration. Quantized motion steps are distinct from decoded frames.

Component	Setting
Backbone	Qwen2.5-Omni-3B Thinker
Text stack	Width 2,048; 36 layers; 16 heads; original vocabulary 151,936
Audio tower	Width 1,280; 32 layers; frozen
LoRA	Rank 32, scale 64, dropout 0.05
Adapted projections	Query, key, value, output, gate, up, and down projections
Fully trainable modules	Input token embeddings and language-model output projection
Motion tokenizer	H2VQ, 6 streams, 512 entries per stream
Motion representation	623 features; 52 joints; decoded at 20 fps
Optimizer	AdamW; learning rate 3×10^{-4}
Schedule	Warmup ratio 0.05; up to 8 epochs; early-stopping patience 3
Batching	Per-device batch 4; accumulation 4
Validation cadence	Every 250 optimizer steps
Sequence limits	1,536 sequence tokens; 96 quantized motion steps
SFT data	14,648 supportive training pairs; 1,467 validation pairs (304 supportive, 1,163 socially appropriate)
Conditioning	Full-cue override $q = 0.667$, otherwise sampled subset
Target	Duration, interleaved motion codes, and sequence-boundary markers

The supportive-only filter applies to the training split. Validation loss is computed over both validation tiers, with socially appropriate pairs serving as validation targets alongside supportive pairs.

The complete input-embedding matrix and output projection are optimized, rather than only the newly introduced motion-token rows. Together they contain 634.5M parameters; LoRA contributes 82.2M, giving 716.7M trainable generator parameters. The separately trained 32.9M-parameter H2VQ tokenizer remains fixed during generator adaptation, and the 17.1M-parameter evaluation judge is a separate model.

Vocabulary and decoding. The vocabulary extension specifies 3,240 symbols: six disjoint 512-entry motion namespaces, 160 duration markers, and eight prompt/target markers. Of these symbols, 2,969 are absent from the pretrained tokenizer, increasing its vocabulary from 151,936 to 154,905 entries. The difference arises because some requested control strings already have token identifiers in the pretrained vocabulary.

For a reaction with N quantized steps, the target is

$$[\texttt{<Length N>}, \texttt{<Motion Tokens>}, x_1, \ldots, x_N, \texttt{</Motion Tokens>}, \texttt{<EOS>}].$$

Each x_n contains the six codes defined in the main text. At decoding, namespace identifiers separate the streams. For body or hands p, residual embeddings are summed,

$$\hat{z}_{p,n} = \sum_{r=0}^{2} e_{p,r}(x_{p,r,n}), \qquad \hat{m} = \mathcal{D}_{\mathrm{mot}}(\hat{z}_b, \hat{z}_h).$$

Here $x_{p,r,n}$ is the code selected for component $p \in \{b, h\}$, residual level r, and quantized time step n, so that $x_{b,r,n} = b_{r,n}$ and $x_{h,r,n} = h_{r,n}$ in Eq. 1; $e_{p,r}$ is the corresponding codebook embedding; and $\mathcal{D}_{\mathrm{mot}}$ is the frozen motion decoder. The hand decoder also conditions on the body representation. The reported configuration uses all six streams. Decoded outputs enter evaluation only after valid parsing and the stated minimum-length requirement.

Modality dropout. The full-cue override and subset selection are distinct. If the six documented subsets are sampled uniformly in the non-override branch, the marginal probability of $A{+}T{+}E$ is $q + (1-q)/6$, approximately 0.723, not 0.667. A fixed-cue training configuration instead fixes the condition mode. Generator ablations always use the full-cue judge, so changing generator inputs does not also change the scoring rule.

B.2 Preference Optimization

DPO initializes both the trainable policy and its fixed reference from the SFT model. It uses full $A{+}T{+}E$ conditioning, learning rate 5×10^{-7}, gradient accumulation 8, and validation every 50 optimization steps. The principal set contains 19,206 training comparisons and 940 validation comparisons, with score difference at least 0.35 and at most four comparisons per utterance. Every comparison ranks two motions for the same condition; it need not be restricted to a supportive–incorrect tier pair. The test partition is excluded from preference construction, optimization, and model selection. The principal configuration trains for one epoch. Label smoothing, when used, is 0.25. Model selection also constrains the preference margin and KL divergence to 8.0 and 25.0, respectively.

Model selection. The principal SFT model uses only the cross-entropy objective. Because adapter initialization can also determine LoRA capacity, the main-text initialization comparison explicitly holds rank and scale fixed.

C Evaluator and Metric Definitions

C.1 Compatibility Judge

The judge has condition and motion towers, each producing a normalized 256-dimensional embedding. The condition tower consumes precomputed audio features of dimension 2,048, transcript features of dimension 768, and text-encoded emotion features. Each branch applies layer normalization, a 512-unit linear layer, GELU, a 256-unit linear layer, and normalization. A masked mean of available branches is normalized to obtain z_c.

The motion tower embeds each of six code streams using a 512×512 table initialized from the corresponding motion codebook. It sums stream embeddings at each quantized step, adds positional information, and applies a four-layer transformer with width 512, eight heads, and feed-forward width 2,048. Masked pooling and projection produce the normalized motion embedding z_m. The interaction head is

$$J(m, c) = \sigma(\mathrm{MLP}([z_c, z_m, z_c \odot z_m])),$$

with hidden width 256 and a scalar output.

The final judge uses equally weighted pointwise binary cross-entropy and symmetric in-batch contrastive losses. It is optimized for four epochs from a previously trained judge, with no listwise-preference loss. Optimization uses AdamW, learning rate 2×10^{-4}, weight decay 0.01, batch size 128, and gradient clipping at 1.0. No human-rated pair is included in its training.

Hard-negative refinement. We mine motions that score highly across unrelated utterances and add them as condition-specific hard negatives. The final judge assigns 0.3015 to the constant-motion control, compared with 0.7309 for paired references (Table 2), supporting its use for condition-specific evaluation.

C.2 Specificity and Retrieval

Gain follows Eq. 4, with 48 distractor conditions per evaluation item and fixed seed 0. Distractors are drawn with replacement from the other 357 held-out utterances, so the intended condition is never its own distractor, and the draws are shared across all full-coverage systems.

MRR uses the intended condition plus the same 48 distractors. For item i, its reciprocal rank is

$$\left(1 + \sum_{d \in D_i} \mathbf{1}[J(m_i, c_d) > J(m_i, c_i)]\right)^{-1};$$

and MRR is the mean of this quantity over evaluation items. Strict comparison favors the intended condition in ties. R@1 and R@3 instead use cosine similarity $s_{ij} = z_m(m_i)^\top z_c(c_j)$ over a 32-condition candidate set C_i. Defining

$$r_i = 1 + \sum_{j \in C_i \setminus \{i\}} \mathbf{1}[s_{ij} > s_{ii}], \qquad \mathrm{R@}k = \frac{1}{n_{\mathrm{eval}}} \sum_i \mathbf{1}[r_i \le k],$$

the reported values use $k \in \{1, 3\}$. The candidate set contains the intended condition and 31 distinct distractors. MRR and R@k therefore use different scoring functions and candidate pools and are not interchangeable ranks.

All systems in the main comparison produce valid outputs for the full 358-item evaluation set. Full-coverage comparisons reuse the same distractor and reference sets.

C.3 Distribution, Coverage, and Diversity

Let z_i be a normalized 256-dimensional motion embedding. J-FID applies the standard Fréchet formula (Heusel et al., 2017) to empirical means and covariances of generated and paired-reference embeddings:

$$\|\mu_g - \mu_r\|_2^2 + \mathrm{Tr}\Big(\Sigma_g + \Sigma_r - 2(\Sigma_r^{1/2}\Sigma_g\Sigma_r^{1/2})^{1/2}\Big).$$

The paired reference has J-FID zero because it is compared with itself. Reported J-FID values use the same judge embedding, dataset, and sample count across systems.

Manifold precision and recall (Kynkäänniemi et al., 2019) use $k = 3$ nearest-neighbor radii in the same embedding space. Let $\mathcal{M}_r$ and $\mathcal{M}_g$ be the unions of the resulting reference and generated neighborhoods, and let Z_r and Z_g be their embedding sets. Then

$$\mathrm{Precision} = \frac{1}{|Z_g|} \sum_{z \in Z_g} \mathbf{1}[z \in \mathcal{M}_r], \qquad \mathrm{Recall} = \frac{1}{|Z_r|} \sum_{z \in Z_r} \mathbf{1}[z \in \mathcal{M}_g].$$

Precision measures whether generated samples lie near reference motions; recall measures whether generated samples cover the reference distribution. High precision alone can conceal concentration on a small subset.

Diversity averages Euclidean distance over a set $\mathcal{P}$ of 10,000 randomly sampled pairs of distinct generated motions:

$$\mathrm{Div} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \|z_i - z_j\|_2.$$

It is an absolute distance in $[0, 2]$, not a reference-normalized ratio. The reference value 1.353 supplies context, and the random-library control reaches 1.348; the metric therefore has little dynamic range on this library and is used only to detect collapse. A larger value does not automatically indicate better coverage, and a smaller value does not by itself establish collapse.

MM measures variation for a fixed condition. For $U = 100$ utterances and $R = 10$ generated samples per utterance, with embedding $z_{u,r}$ for sample r of utterance u,

$$\mathrm{MM} = \frac{1}{U\binom{R}{2}} \sum_{u=1}^{U} \sum_{1 \le r < r' \le R} \|z_{u,r} - z_{u,r'}\|_2.$$

Autoregressive sampling uses temperature 0.8 and top-p 0.95. Deterministic top-1 retrieval has MM zero by construction, while the single-reference protocol has no MM value. Random-library MM of 1.347 is an unconditioned control rather than a desired value.

C.4 Decoded Motion and Statistical Analysis

The decoded speed diagnostic computes mean joint speed and reports its ratio to the paired-reference mean. It provides a complementary measure of motion activity and helps identify low-amplitude or near-idle generation.

Reported Gain intervals use 5,000 paired bootstrap resamples of utterances. Paired Wilcoxon analyses use signed ranks, and the inference-modality family receives Holm correction. The reported intervals quantify utterance-level uncertainty for each trained model.

D Baseline Protocols and Complete Results

D.1 Caption, Decoder, and Retrieval Settings

The comparison includes four caption sources: Qwen2.5-Omni-3B with audio alone, the same model with $A{+}T{+}E$, Phi-4 (Abdin et al., 2024) with transcript and emotion, and Qwen3.6-27B (Qwen Team, 2026) with transcript and emotion. Caption generation is greedy and limited to 64 new tokens.

The two cascade decoders receive the same caption for a given caption source. Our decoder is the caption–motion model used for ReactMotion initialization, and its motion output is sampled even though caption generation is greedy. Kimodo is a kinematic motion diffusion model (Rempe et al., 2026) that requires a target duration; we set it to 5.6 s, the median reaction duration of the corpus (Figure 6), so that its outputs match the typical reference length, whereas ReactMotion predicts the duration with its duration token.

Caption retrieval queries a 783-clip motion library using caption embeddings and top-1 cosine similarity. The same library supplies the random-motion control.

D.2 Complete Caption-Mediated Grid

Tables 6 and 7 split the same systems into alignment and distribution/variation measurements. They avoid shrinking all metrics into one unreadable table. The main table includes all twelve caption-mediated pipelines and both ReactMotion operating points; the tables here add MRR and decoded-speed diagnostics.

Table 6: **Complete alignment results** on 358 utterances. The organization matches Table 3; MRR and R@k use different scores and candidate pools. Bold marks the best directional value among non-reference systems.

System / captioner	Gain↑	R@1↑	R@3↑	MRR↑	J-FID↓
Random library motion	0.0012	0.0279	0.0838	0.0909	0.1304
Paired reference	0.4032	0.5559	0.8212	0.5470	0.0000
Cascade: caption → our stage-one decoder					
Omni-3B (A)	0.0225	0.0223	0.0922	0.1008	0.6769
Omni-3B (A+T+E)	0.0188	0.0279	0.0978	0.0973	0.6817
Phi-4	0.0224	0.0475	0.1285	0.1246	0.5252
Qwen3.6-27B	0.0362	0.0503	0.1201	0.1204	0.5027
Cascade: same caption → Kimodo					
Omni-3B (A)	0.0485	0.0447	0.1089	0.1112	0.5074
Omni-3B (A+T+E)	0.0351	0.0531	0.1201	0.1220	0.5146
Phi-4	0.0517	0.0503	0.1257	0.1250	0.4162
Qwen3.6-27B	0.0498	0.0419	0.1369	0.1204	0.4254
Retrieval: caption → motion library					
Omni-3B (A)	0.0464	0.0698	0.1676	0.1168	0.5595
Omni-3B (A+T+E)	0.0627	0.0587	0.1788	0.1402	0.4981
Phi-4	0.0871	0.0670	0.1648	0.1524	0.3260
Qwen3.6-27B	0.1157	0.0838	0.1983	0.1660	0.3226
Direct utterance-to-motion generation					
ReactMotion SFT	0.3392	0.3101	0.5894	0.4034	**0.0824**
SFT+DPO	**0.3790**	**0.3715**	**0.6145**	**0.4612**	0.0966

Table 7: **Complete distribution, coverage, and variation results** for the systems in Table 6. Div. and MM are absolute embedding distances; only speed is reference-normalized.

System / captioner	Prec.↑	Rec.↑	Div.	MM	Speed ratio
Random library motion	0.869	0.897	1.348	1.347	0.96
Paired reference	1.000	1.000	1.353	–	1.00
Cascade: caption → our stage-one decoder					
Omni-3B (A)	0.838	0.385	1.051	0.996	0.10
Omni-3B (A+T+E)	0.874	0.316	1.054	1.019	0.11
Phi-4	0.779	0.363	1.215	1.070	0.14
Qwen3.6-27B	0.774	0.385	1.235	1.096	0.24
Cascade: same caption → Kimodo					
Omni-3B (A)	0.821	0.179	1.205	0.884	0.46
Omni-3B (A+T+E)	0.855	0.271	1.169	0.904	0.48
Phi-4	0.735	0.148	1.247	0.911	0.52
Qwen3.6-27B	0.696	0.226	1.249	0.892	0.56
Retrieval: caption → motion library					
Omni-3B (A)	0.453	0.629	1.205	0.000	0.70
Omni-3B (A+T+E)	0.489	0.704	1.201	0.000	0.78
Phi-4	0.696	0.726	1.313	0.000	0.83
Qwen3.6-27B	0.799	0.522	1.309	0.000	0.92
Direct utterance-to-motion generation					
ReactMotion SFT	0.930	0.858	1.335	0.978	0.92
SFT+DPO	**0.958**	**0.869**	1.326	0.843	0.90

E Additional Controlled Studies

These studies isolate five questions: caption–motion initialization, preference-optimization settings, cue availability, exposure to reaction identities during SFT, and generalization to recorded conversation. A separate tokenizer analysis verifies reconstruction and codebook use.

E.1 Caption–Motion Initialization

The matched-rank comparison in Section 5.5 uses $r = 32$ and $\alpha = 64$ for both initializations. No task-specific initialization means starting without the text-to-motion adapter, not training a 3B language model from scratch. The earlier rank-16 comparison changes capacity as well as initialization and is not used to estimate an initialization effect.

The additional metrics for the no-initialization configurations are:

Table 8: No-caption–motion-initialization configurations at matched rank, covering all 358 utterances. Bold compares SFT and DPO.

Stage	Gain↑	R@1↑	R@3↑	MRR↑	J-FID↓	Prec.↑	Rec.↑	Div.	MM
SFT	0.3649	**0.3966**	0.6173	**0.4648**	**0.0759**	0.947	0.832	1.342	0.958
SFT+DPO	**0.3708**	0.3687	**0.6285**	0.4544	0.0773	**0.953**	**0.927**	1.338	0.871

Without caption–motion initialization, DPO increases R@3 from 0.6173 to 0.6285, precision from 0.947 to 0.953, and recall from 0.832 to 0.927. At the same time, MM decreases from 0.958 to 0.871, indicating that the improved matching coverage is accompanied by less variation among repeated samples.

E.2 Preference-Optimization Sensitivity

Every β setting improves Gain over the SFT starting point (0.3392), and $\beta = 0.3$ gives the largest improvement, so it is the primary configuration. Larger β trades some Gain for recall, and label smoothing has little effect. Table 9 reports the comparison.

Table 9: Preference-strength sensitivity relative to initialized SFT (Gain 0.3392, recall 0.858). LS denotes label smoothing (0.25). Bold marks the best value among DPO settings.

Method	Setting	Gain↑	Rec.↑
DPO	$\beta = 0.1$	0.3595	0.852
DPO	$\beta = 0.1$, LS	0.3581	0.846
DPO	$\beta = 0.3$	**0.3790**	0.869
DPO	$\beta = 0.5$	0.3611	**0.888**
DPO	$\beta = 1$	0.3542	**0.888**

E.3 Complete Inference-Cue Diagnostics

Table 10 examines which cues the generator uses, in two complementary ways. The *training-cue* rows are separate SFT models trained without modality dropout on a fixed cue subset and evaluated with the same subset. The *inference-cue* rows regenerate the primary SFT and SFT+DPO checkpoints, which are trained with modality dropout, while withholding the listed cues at inference. All rows are computed in one evaluation run with regenerated samples and their own candidate draws, so absolute values differ slightly from Table 3, and comparisons should be read within the table; the paired reference is included as an anchor.

Three findings follow. First, the transcript and the emotion label carry complementary semantic signal: among fixed-cue models, training on transcript and emotion gives the highest retrieval accuracy (R@3 0.506, MRR 0.365), the transcript alone reaches R@3 0.425, and adding the emotion label to audio raises R@3 from 0.117 to 0.310. The ordering $A < A{+}E < A{+}T < T{+}E$ shows that each added cue contributes. Second, modality dropout makes the primary models robust to missing cues. Withholding audio at inference leaves R@3 essentially unchanged (0.603 versus 0.567 for SFT; 0.609 versus 0.629 for SFT+DPO), and the transcript alone retains 0.548 and 0.573. Audio and transcript without an emotion label, the condition available on recorded conversation (Appendix E.6), retain 0.514 and 0.595. Third, preference refinement improves over SFT under every inference cue subset, for example MRR 0.429 versus 0.391 with all cues and 0.421 versus 0.361 with audio and transcript, so the DPO gain is not tied to one input configuration.

Table 10: **Cue ablation.** J is the mean judge score against the intended utterance; MRR uses 48 distractors; R@k uses 32 candidates. Training-cue rows are separate fixed-cue SFT models; inference-cue rows regenerate the primary checkpoints with the listed cues. Bold marks the best value within each block.

Cues	$J\uparrow$	MRR↑	R@1↑	R@3↑	J-FID↓	Rec.↑
Paired reference	0.731	0.555	0.528	0.835	0.000	1.000
Training cues (fixed-cue SFT models)						
$A{+}T{+}E$	0.623	0.326	0.237	0.455	0.119	0.802
$T{+}E$	**0.642**	**0.365**	**0.260**	**0.506**	0.122	**0.821**
$A{+}T$	0.598	0.317	0.246	0.441	0.146	0.779
T	0.606	0.305	0.218	0.425	0.109	0.807
$A{+}E$	0.555	0.249	0.148	0.310	**0.105**	0.810
A	0.307	0.081	0.039	0.117	0.135	**0.821**
Inference cues, SFT (trained with modality dropout)						
$A{+}T{+}E$	0.673	0.391	**0.332**	0.567	**0.086**	**0.883**
$T{+}E$	**0.678**	**0.404**	0.316	**0.603**	0.102	0.810
$A{+}T$	0.655	0.361	0.274	0.514	0.089	0.858
T	0.657	0.394	0.316	0.548	0.100	0.821
$A{+}E$	0.543	0.213	0.120	0.288	0.123	0.793
A	0.336	0.085	0.028	0.087	0.187	0.805
Inference cues, SFT+DPO (trained with modality dropout)						
$A{+}T{+}E$	**0.704**	**0.429**	**0.377**	**0.629**	0.111	0.810
$T{+}E$	0.699	0.415	0.355	0.609	0.091	**0.891**
$A{+}T$	0.682	0.421	0.330	0.595	0.092	0.874
T	0.675	**0.429**	0.358	0.573	**0.087**	0.844
$A{+}E$	0.580	0.249	0.137	0.299	0.123	0.844
A	0.360	0.110	0.045	0.106	0.206	0.732

E.4 Tokenizer Reconstruction and Codebook Use

We evaluate the H2VQ tokenizer used by all reported generators on a 5% hash-selected held-out partition containing 2,616 files and approximately 97.8 million feature elements.

Table 11: **H2VQ reconstruction.** MAE is measured in mixed feature coordinates. The ratio divides tokenizer MAE by mean-predictor MAE; body and hand are separate feature groups.

Component	MAE	Mean baseline	Ratio	Norm. MAE	Norm. RMSE
Body	0.0329	0.1333	0.247	0.199	0.444
Hand	0.0544	0.2195	0.248	0.199	0.376

All 512 entries are active in the audited codebooks, with perplexity 342–371 and maximum single-code occupancy at most 5.8%. These usage statistics show broad codebook utilization. Because the motion features combine position, velocity, and rotation, the ratio to a mean predictor provides a dimensionless reconstruction reference.

E.5 Reaction-SFT Reservation

The intervention designates 200 reaction identities and removes them from reaction SFT, leaving 9,952 of 14,648 training pairs (32.1% removed). DPO filtering rejects a preference comparison if either motion is designated, leaving 9,087 of 19,206 comparisons (52.7% removed). Other settings follow the main recipe. The unchanged 358-utterance evaluation is grouped by whether its paired-reference identity was designated: 138 designated-reference and 220 other cases. This isolates exposure as a reaction-training target; the assets remain in tokenizer learning and caption–motion initialization.

Table 12: **Complete reaction-SFT reservation analysis.** Entries are Gain. Differences are designated minus other, recomputed from displayed means. Reference-group labels remain fixed across rows.

System / training pool	Designated (138)	Other (220)	Δ	p
SFT / reserved	0.3218	0.3154	0.0064	0.73
SFT+DPO / reserved	0.3008	0.2979	0.0029	0.85
SFT / full	0.3518	0.3312	0.0206	0.27
SFT+DPO / full	0.3665	0.3868	−0.0203	0.79
Qwen3.6-27B retrieval	0.1027	0.1239	−0.0212	0.44
Paired reference	0.4228	0.3910	0.0318	0.15
Random motion	−0.001	0.003	−0.004	0.42

Reaction-reserved SFT yields similar Gain on designated and other reference identities (0.3218 versus 0.3154), showing that direct reaction-SFT exposure is not required for the conditional mapping. The full-training and baseline rows provide reference-group context.

E.6 Generalization to Recorded Conversation

The training split inventories 13,856 recorded listener reactions with speaker audio (Section 3.1). They carry no transcript, emotion label, or caption, and no reported model is trained on them, so they provide an evaluation on recorded conversation that is disjoint from the constructed corpus. We sample at most two clips from each of the 425 recording sessions (500 clips), transcribe the speaker audio with Whisper-small.en (Radford et al., 2023), drop transcripts shorter than three characters, and de-duplicate on the transcript so that every condition is distinct, leaving 462 conditions with a median transcript length of 8 words. Because no emotion label exists, generators are conditioned on audio and transcript only, a setting under which the dropout-trained models retain most of their retrieval accuracy (Appendix E.3). Gain uses the same judge and protocol as the main evaluation, with 48 distractors drawn from the same pool. Input-independent controls pair the recorded clips with unrelated utterances (a derangement of the pool checked for zero self-pairings, and a shuffled pairing) or play them backwards.

Table 13: **Recorded-conversation evaluation** on 462 transcribed utterances. Generators receive audio and transcript only. J is the mean judge score against the intended utterance; MRR uses the intended condition and 48 distractors. Controls pair recorded clips with unrelated utterances or reverse them in time.

System	Gain↑	J	MRR↑
ReactMotion SFT	0.1291	0.448	0.226
SFT+DPO	**0.1536**	0.473	**0.252**
Random pairing	−0.0062	0.336	0.098
Shuffled pairing	0.0046	0.347	0.099
Time-reversed reaction	−0.0016	0.340	0.093

Table 13 shows that the system ordering from the constructed benchmark is preserved under this domain shift: SFT+DPO obtains higher Gain and MRR than SFT, and both are well above the input-independent controls, which stay at zero as intended. The generated reactions therefore depend on the recorded utterances even though transcripts come from automatic speech recognition and the emotion cue is absent. Because the judge is fitted to the constructed corpus, Gain on recorded conversation is read as a measure of utterance dependence relative to the input-independent controls; scoring agreement with the recorded responses themselves would require a judge fitted to recorded dyadic data, which we leave to future work.

F QUALITATIVE RESULTS



Figure 9: **Additional qualitative results.** Three speaker conditions with motions generated by SFT+DPO, SFT, caption retrieval, and the cascade baseline. The first and third examples compare the four systems for anxious and embarrassed utterances, respectively. The middle example shows the resulting motion sequences for an amused utterance. Frames are sampled uniformly from each generated sequence; video renderings of these and further examples accompany the supplementary material.

Where the model fails. Figure 10 quantifies failures on the 358 held-out utterances using per-utterance Gain. SFT+DPO scores Gain ≤ 0 on 7.0% of utterances (SFT: 10.3%; paired reference: 5.6%). These failures track the difficulty of the utterance: stratified by the Gain that the paired reference itself achieves, the failure rate is 16.7% in the hardest quintile and 1.4% in the easiest, and paired-reference and SFT+DPO Gain are rank-correlated (Spearman $\rho = 0.47$, $p = 5 \times 10^{-21}$). Ranked by emotion over the 21 categories with at least eight test utterances, mean Gain is highest for high-arousal positive utterances (ecstatic, triumphant) and lowest for reflective, low-arousal ones (focused, contemplative). We report this ordering as an observation about where the benchmark is hard.

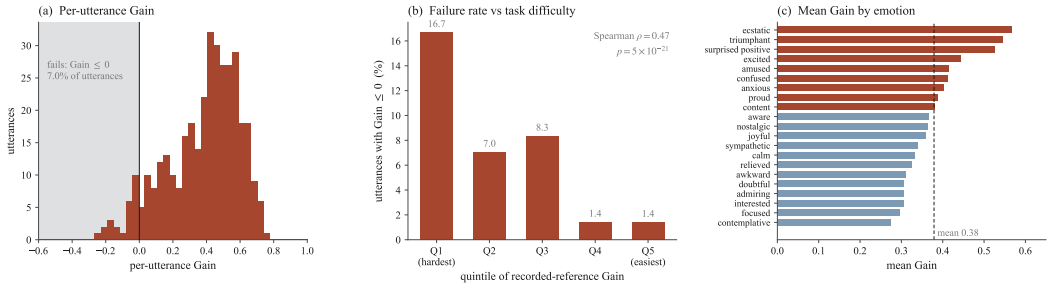

Figure 10: **Quantitative failure analysis of SFT+DPO on the 358 held-out utterances.** (a) Distribution of per-utterance Gain; the shaded region marks Gain ≤ 0. (b) Failure rate by quintile of the Gain achieved by the paired reference, from hardest to easiest. (c) Mean Gain by emotion for the 21 categories with at least eight utterances; the dashed line is the overall mean.