CONGRETS: LOW-RESOURCE, LOW-LATENCY CO-SPEECH GESTURE SYNTHESIS VIA CONTRASTIVE LATENT PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

How small can a complete 3D co-speech gesture system be while scoring on par with state-of-the-art generators? The question matters most for LLM-powered chatbots and dialogue systems, where a response exists as text before speech is synthesized: gestures could be generated without waiting for audio, yet most generators are driven by audio the agent does not yet have. ConGRets separates motion and speaker representation learning from text-conditioned gesture prediction. We learn the motion representation contrastively rather than through a VAE or VQ-VAE, then freeze the encoder before decoder training. A separate contrastive encoder derives a global speaker descriptor from reference motion, enabling new-speaker conditioning without generator retraining. A fine-tuned text encoder and fusion head predict the next 10-dimensional motion latent from response text, speaker conditioning, and previous motion, trained by contrastive alignment with the motion space. Thus one predictor supports memory-based synthesis by selecting and blending recorded units, hybrid synthesis by decoding selected unit latents, and direct synthesis by decoding predicted latents. On holdout takes from five BEAT speakers excluded from training, and without audio, memory-based synthesis achieves 58% lower position-space Fréchet Gesture Distance (FGD) than the lowest-FGD baseline among seven audio-conditioned systems, with pose diversity close to recorded motion. Hybrid synthesis achieves FGD close to that baseline; direct synthesis achieves lower FGD than three baselines. All modes use at most 12 million generation-time parameters, 27$\times$ fewer than the lowest-FGD baseline. After enrollment, loading, and warm-up, they process the transcript of a 69-second utterance in 23–31 ms on CPU and 15–19 ms on GPU, where they are over 30 times faster than the fastest evaluated baseline on the same input. These results demonstrate that compact text-conditioned prediction can deliver strong distributional gesture quality at low generation cost, providing a practical design for interactive conversational agents and motivating deployment on resource-constrained edge devices. (Demo video is attached as supplementary file)

1 INTRODUCTION

An embodied dialogue agent must decide how to move as well as what to say. Co-speech gestures, chiefly upper-body and hand movements accompanying speech, help convey meaning (McNeill, 1992; Kendon, 2004). Unlike text-to-motion prompts specifying actions or paths (Guo et al., 2022), a dialogue utterance leaves its gesture unspecified: the same words can accompany different movements (Li et al., 2021; Ferstl et al., 2021), and speakers gesture differently (Neff et al., 2008; Ginosar et al., 2019) (Figure 1). The task here is to predict joint positions in 3-D space over time, rather than generate 2-D video frames.

Early virtual humans selected recordings from response text using rules (Cassell et al., 2001; Lee & Marsella, 2006); learned generators often require speech audio (Ginosar et al., 2019; Yoon et al., 2020), including recent speech-encoder and diffusion systems (Yang et al., 2023b; Cheng & Huang, 2025). Audio input makes them wait for text-to-speech (Alexanderson et al., 2020). Low latency matters in conversational turn-taking (Levinson & Torreira, 2015; Maslych et al., 2025): three

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.

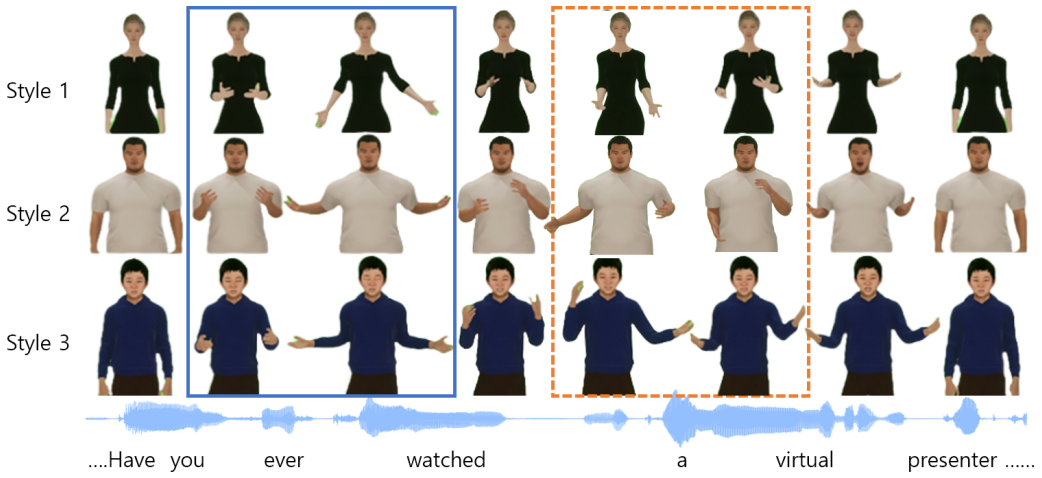


Figure 1: Illustrative ConGRets output for one utterance in three speaker styles. Solid boxes mark the same gesture in different styles; dashed boxes mark different gestures for the same words. The waveform is illustrative; ConGRets does not use audio or predict speech-aligned timing.

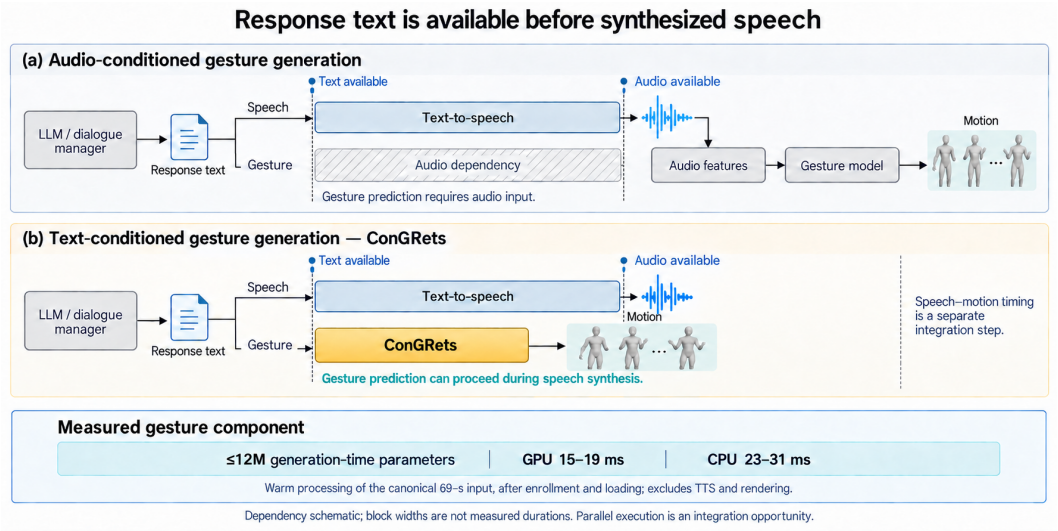


Figure 2: Response text can drive ConGRets before synthesized audio is available, whereas audio-conditioned gesture models require an audio input. Blocks are schematic; the inset's size and warm-up time cover only the gesture component after enrollment and loading. Speech synchronization is post-processed.

diffusion systems in our seven-baseline evaluation take 16–76 seconds of warm GPU time for a 69-second utterance (Table 1).

ConGRets instead starts from response text (Figure 2). Offline motion and speaker representation learning leaves a compact predictor to choose one motion code per gesture unit. Captured motion can supply natural movement: GestureMaster's motion graph had the highest median human-likeness rating in GENEA 2022 (Zhou et al., 2022; Kucherenko et al., 2024), and ACT2G retrieves gestures by contrastive text matching (Teshima et al., 2023). A predicted code can select a recording or drive a decoder.

ConGRets embeds two-second gesture units contrastively in a 10-D space, then freezes the encoder before decoder training (Figure 3); this is within the 2–3-second range of a user-evaluated text-to-

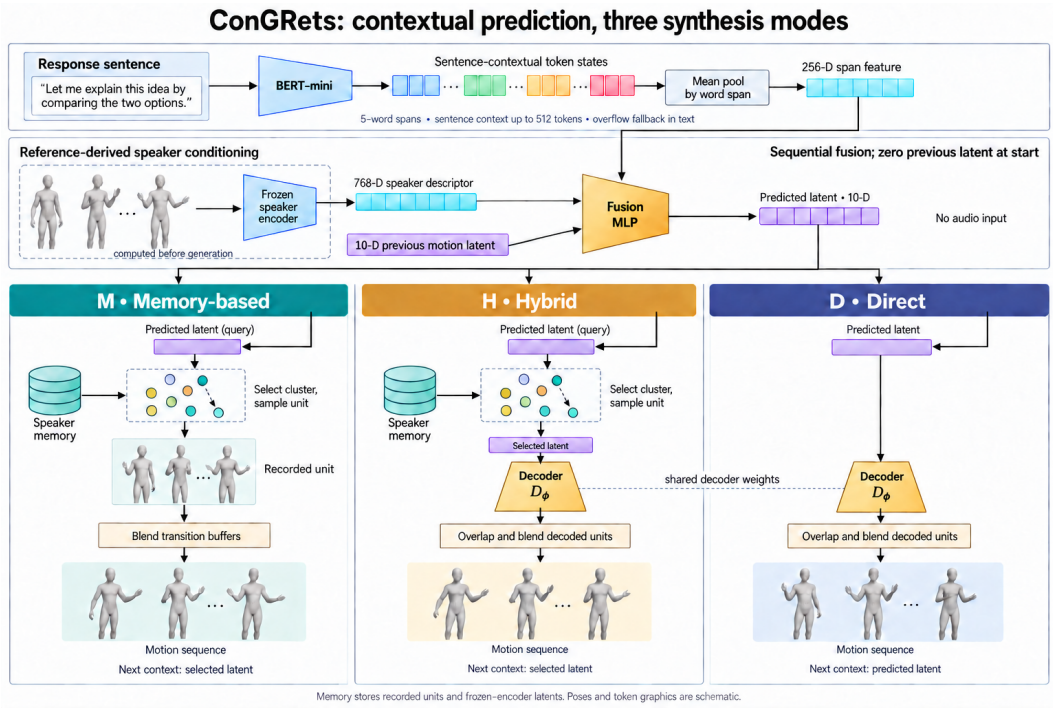


Figure 3: ConGRets predicts a motion latent from sentence-contextual text, an enrolled speaker descriptor, and the previous motion latent. M selects recorded units, H decodes selected latents, and D decodes predicted latents. Enrollment precedes generation; poses and embeddings are schematic.

gesture system (Ali et al., 2025b). Single-speaker models capture individual style (Ginosar et al., 2019), while multi-speaker models risk averaging it (Ahuja et al., 2020); prior work uses speaker IDs or embeddings (Ahuja et al., 2020; Yoon et al., 2020) and reference examples (Ghorbani et al., 2023; Fares et al., 2023). ConGRets takes the reference route, with text supplying content cues: a separate contrastive encoder pools one ten-minute reference window into a global descriptor; new-speaker enrollment needs at least ten minutes of motion but no predictor update. Fine-tuned BERT-mini (Turc et al., 2019) and a small head predict the next latent from text, the descriptor, and previous motion against the locked target, as in LiT (Zhai et al., 2022). The space serves as prediction target, memory index, and decoder input: **memory-based synthesis (M)** replays selected units, **hybrid synthesis (H)** decodes selected latents, and **direct synthesis (D)** decodes predicted latents.

On five BEAT speakers held out from training but enrolled with reference motion (Liu et al., 2022), M's position-space Fréchet Gesture Distance (FGD; lower is better) (Yoon et al., 2020) is 58% below the lowest-FGD adapted baseline; H is close to that baseline and D beats three baselines (Table 1). With BEAT-trained weights frozen and target enrollment, M/H beat four transferred baselines on Talking With Hands, Audio2Photoreal, and Embody3D (Lee et al., 2019; Ng et al., 2024; McLean et al., 2025); on Trinity (Ferstl & McDonnell, 2018), D leads while M/H do not (Table 2). M/H reuse enrolled motion and baseline reference access differs. Each mode has at most 12 million generation-time parameters versus 91–824 million for six baselines; after enrollment, loading, and warm-up, a 69-second transcript takes 23–31 ms on CPU or 15–19 ms on GPU, over 30 times faster on GPU than the fastest baseline (Table 1; Appendix D). We contribute a compact text- and speaker-conditioned predictor, three realization modes, and a quality–cost evaluation with metric limits.

2 Related Work

Learned co-speech gesture generation. Recent generators (Nyatsanga et al., 2023) use diffusion or flow matching (Alexanderson et al., 2023; Yang et al., 2023b; Mughal et al., 2024; Cheng & Huang, 2025), motion tokens (Yi et al., 2023; Liu et al., 2024), speech encoders such as WavLM

(Chen et al., 2022), or fine-tuned LLMs (Chen et al., 2025). Most need finished speech audio, some modeling its prosody and rhythm (Kucherenko et al., 2020; Ao et al., 2022); MIBURI streams from a speech–text dialogue model's tokens (Mughal et al., 2026), whereas ConGRets needs no speech model and provides no prosodic alignment. Latency has been cut by fewer sampling steps (Liu et al., 2025; Cheng et al., 2025), state-space models (Xu et al., 2024), and real-time or streaming designs (Chen et al., 2024; Saleem et al., 2026); since protocols differ and parameter count does not determine time, we time all systems on one input and machine.

Text-driven gesture. Rule-based agents planned behavior from response text, BEAT using speech timing only at realization (Cassell et al., 2001; Lee & Marsella, 2006); BML, SmartBody, and the Virtual Human Toolkit (Kopp et al., 2006; Thiebaux et al., 2008; Hartholt et al., 2013) supported virtual-patient applications (Kenny et al., 2009; Swartout, 2010). Learned systems generate gestures from text for robots (Yoon et al., 2019), emotive motion (Bhattacharya et al., 2021), and 2-D poses (Asakawa et al., 2022), or use LLM features (Windle et al., 2024) and LLM selection (Hensel et al., 2023; Ghanad Torshizi et al., 2025; Gao et al., 2024). Others synthesize speech and gesture jointly (Wang et al., 2021; Mehta et al., 2023; 2024), unify speech- and text-driven motion (Yang et al., 2024), or route text through speech synthesis (Alexanderson et al., 2020; Nagy et al., 2021). Closest, ACT2G contrastively retrieves gestures per sentence and fits them to speech duration (Teshima et al., 2023); ConGRets predicts speaker-conditioned latent sequences in a motion-only space and can decode them.

Recorded motion. Motion graphs and motion matching replay recordings in animation (Kovar et al., 2002; Clavet, 2016) and gesture (Stone et al., 2004; Zhou et al., 2022); learned motion matching replaces database operations with networks (Holden et al., 2020). Gesture systems search databases with speech audio (Ferstl et al., 2021; Habibie et al., 2022; Yang et al., 2023a), combine retrieval with generation (Zhang et al., 2024; Mughal et al., 2025; Guo et al., 2026), or match motion for LLM-driven humans (Guo et al., 2025). Text-to-gesture rule maps were derived automatically (Ali et al., 2020), extended by contrastive pose matching (Ali & Hwang, 2022), and given a learned fallback (Ali et al., 2025a). ConGRets compares replay and decoding under one predictor.

Speaker style and representations. Style has been profiled or learned per speaker (Neff et al., 2008; Ginosar et al., 2019), from labels or embeddings of known speakers (Yoon et al., 2020; Ahuja et al., 2020; Liu et al., 2022), text prompts (Ao et al., 2023), or adaptation (Ahuja et al., 2022), and for new speakers from example clips (Ghorbani et al., 2023) or multimodal references (Fares et al., 2023); ConGRets uses a global enrollment descriptor, as in speaker verification (Wan et al., 2018). Motion latents are often learned by reconstruction (Kingma & Welling, 2014; van den Oord et al., 2017; Chen et al., 2023; Zhang et al., 2023), while contrastive losses (Chen et al., 2020) align motion with text for retrieval (Petrovich et al., 2023) and evaluation (Guo et al., 2022) or condition gesture generators (Ao et al., 2023; Deichler et al., 2023). As in LiT (Zhai et al., 2022), only the conditioning tower learns against a locked target; unlike CPC and I-JEPA, whose targets are learned jointly or as a moving average (van den Oord et al., 2018; Assran et al., 2023), our target is trained beforehand, frozen, and also indexes the memory and feeds the decoder.

3 METHOD

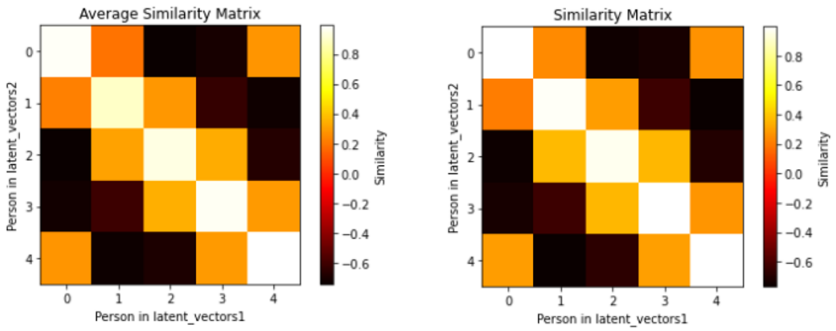
Let $q_{1:T}$ be response-text chunks and x_t a recorded two-second gesture unit. Offline, ConGRets trains a motion encoder $E_m(x_t) = z_t \in \mathbb{R}^{10}$ and a speaker encoder E_s. Enrolling speaker s computes a descriptor $g_s = E_s(R_s) \in \mathbb{R}^{768}$ from a ten-minute reference window R_s. At generation, a predictor estimates the next motion latent,

$$\hat{z}_t = f_\theta\big(e_t, g_s, z_{t-1}^{\mathrm{ctx}}\big)\,, \tag{1}$$

from a BERT-mini text feature e_t (Devlin et al., 2019; Turc et al., 2019) and previous-motion context z_{t-1}^{ctx}. The same predictor drives memory-based (M), hybrid (H), and direct (D) synthesis.

3.1 CONTRASTIVE MOTION AND SPEAKER REPRESENTATIONS

Motion is root-centred 3-D positions of 50 joints at 30 fps, cut into 60-frame units aligned to word-group starts. The motion encoder is trained on random 2-, 3-, or 4-second windows, pairing each clean window with a noisy, crop-shifted copy (Appendix A.2), and then frozen. The speaker encoder is trained separately: two ten-minute chunks of one speaker form a positive pair, other speakers are



a) Similarity Matrix With 1 minute of style encoding

b) Similarity Matrix With 10 minute of style encoding.

Figure 4: Speaker-encoder similarity matrices for five speakers, from (a) one-minute and (b) ten-minute reference motion.

negatives, and ten one-minute embeddings are pooled as a set into g_s (Appendix A.3; Figure 4). No ID label reaches the predictor, and a new speaker is enrolled without updating it. M and H also need the speaker's motion memory; D's output does not.

Both encoders and the predictor use the symmetric NT-Xent loss (Chen et al., 2020). For normalized paired embeddings $u = [a_1, \ldots, a_N, b_1, \ldots, b_N]$, where $p(i)$ indexes the pair of i, and temperature $\tau = 0.1$,

$$\mathcal{L}_{\text{NTX}} = -\frac{1}{2N} \sum_{i=1}^{2N} \log \frac{\exp(\bar{u}_i^\top \bar{u}_{p(i)}/\tau)}{\sum_{k \neq i} \exp(\bar{u}_i^\top \bar{u}_k/\tau)}. \tag{2}$$

3.2 Contrastive next-latent prediction

Training examples hold three consecutive units of one take; BERT-mini encodes each unit's text separately and mean-pools it to $e_t \in \mathbb{R}^{256}$, and a two-layer MLP maps

$$[e_t; g_s; z_{t-1}] \in \mathbb{R}^{1034} \quad \longmapsto \quad \hat{z}_t \in \mathbb{R}^{10}, \tag{3}$$

with z_{t-1} teacher-forced from the recorded predecessor. Only BERT-mini and the MLP are trained, with

$$\mathcal{L}_{\text{pred}} = \mathcal{L}_{\text{NTX}}\big(\{\hat{z}_t\}, \{z_t\}\big); \tag{4}$$

reconstruction losses appear only in ablations. Validation "retrieval accuracy" is in-batch alignment, not retrieval from a speaker's memory (Appendix H.1).

3.3 Speaker motion memory and three synthesis modes

Following the cluster-then-sample selection of Ali et al. (2025b), each speaker's unit latents are clustered with bisecting k-means, $k = \text{round}(n_s/12.5)$ for n_s units (Appendix A.5). Selection takes the centroid most cosine-similar to $\hat{z}_t$ and samples an eligible member uniformly, giving unit x_{j_t} with latent z_{j_t}. The modes differ in output and recurrence:

$$\begin{array}{lll} \text{M:} & y_t = x_{j_t}, & z_t^{\text{ctx}} = z_{j_t}, \\ \text{H:} & y_t = D_\phi(z_{j_t}), & z_t^{\text{ctx}} = z_{j_t}, \\ \text{D:} & y_t = D_\phi(\hat{z}_t), & z_t^{\text{ctx}} = \hat{z}_t. \end{array} \tag{5}$$

M replays recordings, H decodes the selected latent, and D exposes the predictor's own trajectory. The variational decoder D_ϕ shared by H and D is fitted after the motion encoder is frozen; with its posterior mean at inference, D is deterministic, while H stays stochastic through selection.

At generation, BERT-mini encodes each whole sentence and pools its contextual token states over consecutive five-word spans, which are processed in order. M cross-fades up to ten recorded buffer frames between unmodified 60-frame cores, while H and D overlap adjacent decoded clips by ten frames with a decaying pose-offset alignment. Output length therefore follows the number of chunks, not prosody or speech duration: ConGRets does not predict speech-aligned timing. Truncation handling and exact frame counts are in Appendix H.1.

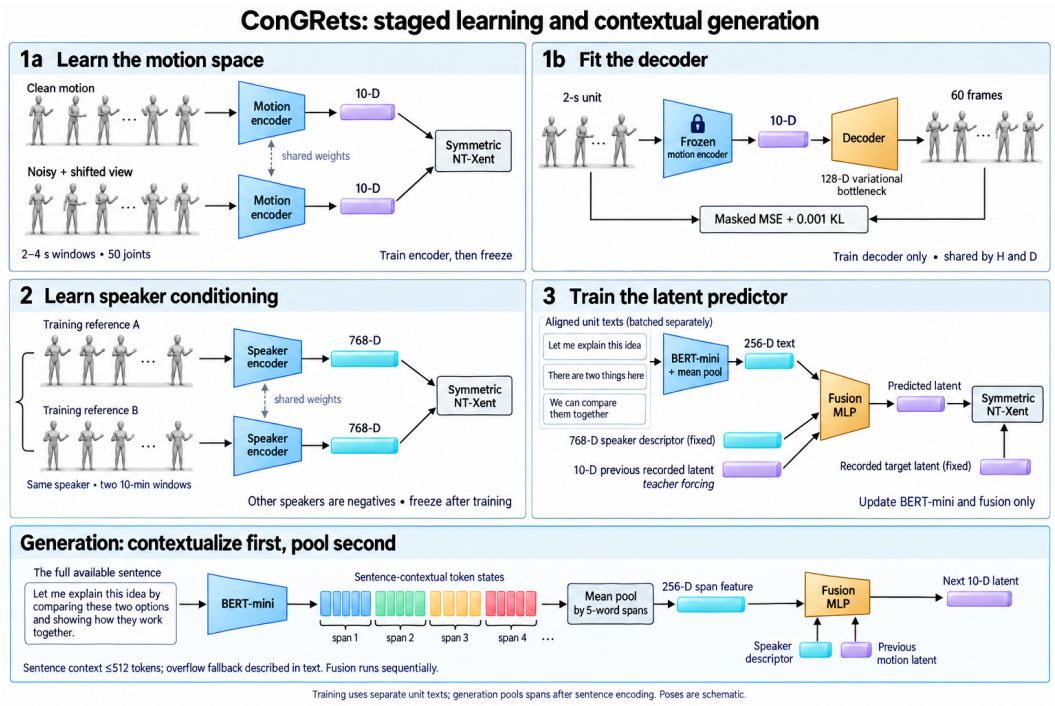


Figure 5: Staged training and generation. (1a) Clean and perturbed views train the motion encoder, which is then frozen; (1b) a decoder is fitted to its fixed latents. (2) Same-speaker reference windows train the speaker descriptor. (3) BERT-mini and the fusion head align predictions to the fixed motion targets with teacher-forced previous latents. Bottom: generation encodes a sentence, pools span features, and predicts sequentially. Poses, text, and token cells are schematic.

4 Experimental Setup

Data and reference access. We use BEAT (Liu et al., 2022) with 17 training, five validation, and five test speakers, scored as 50 root-centered joints at 30 fps on a common panel of 138 test takes; ablations use their full support, normally 149 takes, and are never mixed with it. ConGRets excludes test speakers from training but enrolls them from reference motion, which every tested mode requires. While a take is generated, M and H cannot select its units, but the memory's centroids and the descriptor may still reflect it (Appendix A.7). The split is speaker-disjoint, not text-disjoint, and evaluation excludes locomotion and facial motion.

Baselines. We adapt the seven audio-conditioned systems of Table 1 (Yang et al., 2023b; Liu et al., 2024; 2025; Cheng & Huang, 2025; Mughal et al., 2025; Habibie et al., 2022; Fares et al., 2023). Apart from Habibie et al. and ZS-MSTM, rest of the baselines are from GENEA Leaderboard. we choose Habibie et al. for retrieval candidate and ZS-MSTM for zero-shot speaker style transfer. Habibie et al. and ZS-MSTM are our re-implementations. Frozen author-supplied components keep their pretraining (Appendix B).

Metrics. Following GENEA Leaderboard, we used same FGD model architecture as our main metric. FGD uses an independently trained, frozen autoencoder in one root-centered position space for all systems (Appendix D.1), so it is not comparable with rotation-space FGD reported elsewhere. Pose diversity is defined in Appendix D.2. Controls are recorded motion, random unit selection, and a static pose from the scored take (an oracle diagnostic, not a generator); neither metric measures human preference or text-gesture fit.

Selection and cost. Checkpoints maximize validation in-batch accuracy; seed 1 is the report run by convention, not by test score, and all seeds appear in Appendix C. Cost uses trained weights and one 69-second input with warm GPU and CPU timing (Appendix D), counting generation networks but not offline encoders, reference preparation, LLM inference, TTS, or rendering.

Table 1: BEAT common panel (138 takes). Lower FGD is better; diversity is read against recorded motion. Baseline budgets, pretraining, skeleton conversion, and reference access differ (Appendix B). Parameters count generation networks, excluding enrollment networks and memory storage. Warm GPU time is for the 69-second input after each system's preparation, rounded to 0.1 ms (full precision in Appendix D).

System	FGD ↓	Pose diversity	Parameters (M)	Warm GPU (ms)
ConGRets-M	0.1567	0.07945	11.438	19.2
ConGRets-H	0.3856	0.04476	11.989	18.4
ConGRets-D	0.9220	0.03527	11.989	14.6
DiffuseStyleGesture+	0.3742	0.07052	328.1	75500.0
Habibie et al.	0.6941	0.08608	11.5	77900.0
EMAGE	0.8046	0.05044	153.1	651.0
RAG-Gesture	0.8451	0.05618	823.6	15800.0
GestureLSM	0.9543	0.04234	91.4	706.0
HoloGest	0.9719	0.10121	338.5	24000.0
ZS-MSTM	1.1171	0.00550	253.0	3058.0
Recorded-motion reference	0.00	0.07967	—	—

5 Results

5.1 Distributional quality and generation cost

With leave-one-take-out selection, M has the lowest FGD of all generators, 58% below DSG+, with pose diversity close to recorded motion (0.0795 vs. 0.0797). H is numerically close to DSG+ (0.386 vs. 0.374), and D has lower FGD only than GestureLSM, HoloGest, and ZS-MSTM. The controls qualify this: text-blind random units reach FGD 0.042, below every generator, and a static pose reaches 0.566 with near-zero diversity, so FGD indicates distributional similarity, not semantic appropriateness.

Every mode uses under 12 million generation-time parameters, about 27 times fewer than DSG+ (excluding memory and enrollment networks), and runs in 14.59–19.25 ms warm on GPU, over 30 times faster than EMAGE (651 ms), the fastest baseline; warm CPU means are 22.82–30.46 ms. Output length follows the text chunks, not the 69-second input. Start-up costs are in Appendix D; Figures 15–16 (supplementary) plot quality–cost and all ablations.

5.2 Design studies and final ablations

In search for the smallest complete system, we performed extensive design studies and ablations. We covered depth and breadth of this problem. Complete details are available in supplementary section. Appendix C reports all 40 final runs (24 recipes, grouped descriptively, not preregistered) with validity flags. On the full ablation panel (normally 149 takes), the three main seeds span FGD 0.072–0.184 for M, 0.369–0.390 for H, and 0.576–0.869 for D. Later alternatives score better on test FGD. A 32-D fine-tuned seed-1 run, a width that development did not retain, has lower FGD than the 10-D seed-1 reference in all three modes (M 0.064 vs. 0.152, H 0.259 vs. 0.373, D 0.462 vs. 0.869) with nearly identical parameters but lower validation accuracy (0.0169 vs. 0.0178), so compactness does not stem from the 10-D width. Frozen BERT-mini lowers M and H FGD (0.059, 0.313) but diverges in D. We treat both as post-hoc observations (Section 6). BERT-tiny (4.6–5.2 M parameters) stays within all seed ranges. Two objective families use a different motion space, so cross-family FGD cannot isolate loss effects (Appendix H.2); Appendix C also records twelve H/D cells invalidated by a decoder-loading defect and divergent D runs outside the main seeds, and speaker studies support a global descriptor, not a temporal-style model (Appendix E).

A full-memory, teacher-forced query diagnostic finds above-chance but low alignment (speaker 0.447 vs. chance 0.200; paired cluster 0.0450 vs. 0.00973; paired unit 0.00479 vs. 0.000655) and does not isolate text or descriptor effects (Appendix H.3).

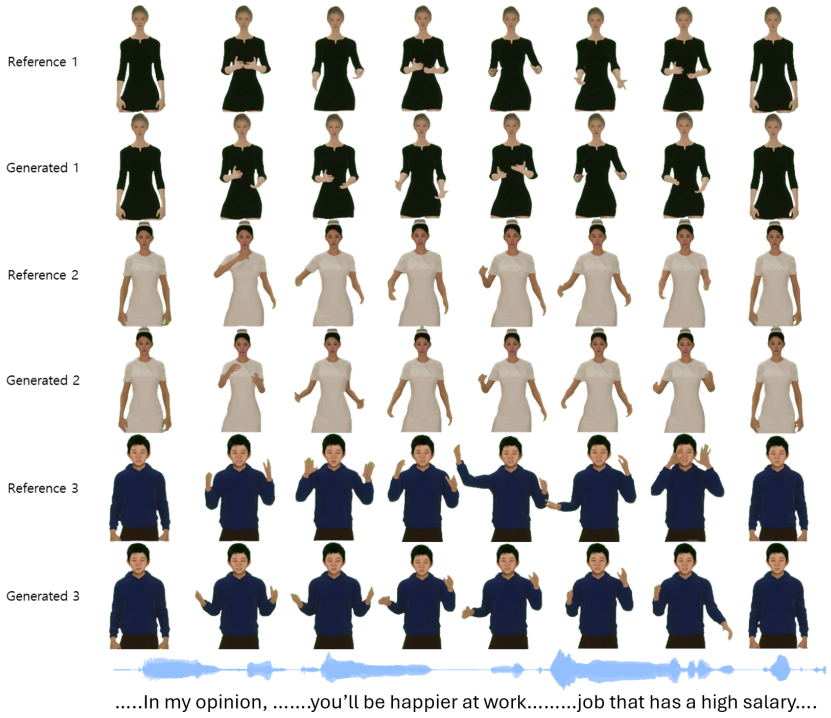


Figure 6: Reference motion and illustrative ConGRets output for three speakers on the same utterance.

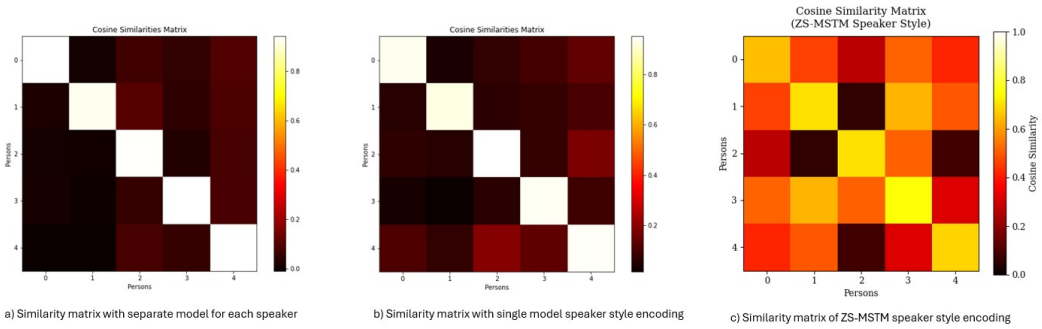


Figure 7: Cosine-similarity matrices of speaker-style embeddings for five speakers: (a) a separate model per speaker, (b) a single model with speaker-style encoding, and (c) ZS-MSTM's style encoding.

Figures 1, 6, and 7 illustrate speaker conditioning outside the quantitative evaluation.

5.3 Frozen-weight transfer with target-speaker enrollment

With ConGRets' BEAT-trained weights frozen and target speakers enrolled, and four baselines under their differing conditioning (one seed; not target-data-free generation), M and H rank first and second among the seven system arms on TWH, Audio2Photoreal, and Embody3D, while on Trinity D leads and M/H rank last; choosing an arm per corpus would be post hoc. These single-seed FGDs are compared only within a pack, because output support differs and the BEAT feature extractor is unvalidated on these domains (Table 2; Appendix F).

Table 2: External-corpus FGD (↓) with ConGRets' BEAT-trained weights frozen and target speakers enrolled. Bold: best of the seven system arms. Coverage and conditioning differ across methods, so compare values within a column (Appendix F).

System / control	Trinity	TWH	Audio2-Photoreal	Embody3D
EMAGE	2.899	3.654	3.310	4.666
ZS-MSTM	1.946	4.353	3.011	4.109
HoloGest	1.935	3.190	2.303	2.342
DSG+	1.037	3.573	1.937	3.817
ConGRets-M	7.140	**1.304**	**0.343**	**1.410**
ConGRets-H	5.497	1.546	0.920	1.740
ConGRets-D	**0.775**	3.794	3.382	3.599

6 Discussion and conclusion

Design and selection. No configuration is best on every metric. The reported configuration was fixed before final test scoring and not changed afterward: checkpoints maximize validation alignment, and seed 1 is a reporting convention rather than our best seed, since seed 0 has lower M FGD. The 32-D model (lower FGD in all three modes) and frozen BERT-mini (lower M and H FGD) are therefore post-hoc candidates for a future main configuration; since validation alignment and FGD can disagree, future selection should also score generated motion.

Limitations. D is trained on teacher-forced latents but fed its own predictions, whose norm cosine alignment leaves unconstrained, plausibly causing divergent D runs. We ran no human study, so ConGRets lacks perceptual validation; our aim was a numerically competitive, low-cost system compared with seven published baselines under one protocol. RAG-Gesture and HoloGest underwent the independent GENEA benchmark (Nagy et al., 2026), DSG+ was evaluated in the GENEA Challenge 2023 (Kucherenko et al., 2023), and recorded-motion systems have been rated favorably by human evaluators (Zhou et al., 2022; Kucherenko et al., 2024; Ali et al., 2025b); these evaluations concern the original systems, not our adaptations or ConGRets. FGD rewards distributional similarity, and M and H reuse the enrolled speaker's other takes while baseline reference access differs. Externally, D leads on Trinity where M and H fail. Frozen-component, or reimplemented baselines limit conclusions about the originals. ConGRets removes the audio prerequisite but do speech synchronization in post.

Conclusion. ConGRets pairs frozen contrastive motion and speaker representations with a compact text-conditioned predictor realized in three modes. On held-out BEAT speakers, memory-based synthesis has the lowest FGD among seven adapted baselines with under 12 million generation-time parameters and millisecond-scale warm generation; transfer is mode-dependent, and semantic conditioning, speech synchronization, and perceptual validation remain open.

REPRODUCIBILITY STATEMENT

The main text specifies the method, conditioning assumptions, split, checkpoint selection, and measurement boundaries. The extended appendices report every final run, including invalid and unfavorable cells, together with development studies, baseline adaptations and exclusions, evaluator configuration and hash, external-corpus selection and coverage and representation controls (Appendices F and G), and a reproduction procedure. These records make the systematic configuration search and each reported score auditable. Reproduction also requires the corresponding datasets, model checkpoints, and enrolled motion artifacts.

ETHICS STATEMENT

The experiments use existing motion and speech data and do not introduce a new human-participant study. Speaker conditioning uses reference motion and can reproduce aspects of a person's motion distribution. Applications should obtain appropriate rights and consent for reference recordings and avoid presenting a synthetic character as an authentic recording of that person. The benchmark covers a limited set of speakers and does not establish performance across demographic groups, languages, or accessibility needs.

AI-USE STATEMENT

AI tools were utilized exclusively in an assistive capacity during the preparation of this manuscript. Specifically, AI tools were used for:

- **Draft Polishing & Critical Feedback:** Editing prose for clarity, phrasing, and grammatical fluency, as well as simulating reviewer-style critical feedback (acting as an interactive paper reading and critique assistant).
- **Coding & Data Processing:** Assisting with implementation scripts, debugging, and programmatically parsing and tabulating raw experimental data into LaTeX tables for the supplementary material.

All foundational research contributions, including conceptual formulation, theoretical development, model architecture design, ablation studies, and scientific interpretations, were conceived and executed independently by the authors without generative assistance. The authors have personally reviewed, verified, and take full intellectual and ethical responsibility for all text, code, and empirical findings presented in this work.

REFERENCES

Chaitanya Ahuja, Dong Won Lee, Yukiko I. Nakano, and Louis-Philippe Morency. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *Proceedings of the European Conference on Computer Vision*, 2020. URL https://arxiv.org/abs/2007.12553.

Chaitanya Ahuja, Dong Won Lee, and Louis-Philippe Morency. Low-resource adaptation for personalized co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20566–20576, 2022.

Simon Alexanderson, Éva Székely, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. Generating coherent spontaneous speech and gesture from text. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, 2020. doi: 10.1145/3383652.3423874.

Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics*, 42(4), 2023. doi: 10.1145/3592458.

Ghazanfar Ali and Jae-In Hwang. Improving co-speech gesture rule-map generation via wild pose matching with gesture units. In *SIGGRAPH Asia 2022 Posters*. Association for Computing Machinery, 2022. doi: 10.1145/3550082.3564185. URL https://doi.org/10.1145/3550082.3564185.

Ghazanfar Ali, Myungho Lee, and Jae-In Hwang. Automatic text-to-gesture rule generation for embodied conversational agents. *Computer Animation and Virtual Worlds*, 31(4–5):e1944, 2020. doi: 10.1002/cav.1944.

Ghazanfar Ali, Hwang Youn Kim, and Jae-In Hwang. RIDGE: Rule-infused deep learning for realistic co-speech gesture generation. *Computer Animation and Virtual Worlds*, 36(4), 2025a. doi: 10.1002/cav.70034. URL https://doi.org/10.1002/cav.70034.

Ghazanfar Ali, Woojoo Kim, Muhammad Shahid Anwar, Jae-In Hwang, and Ahyoung Choi. Expanding multilingual co-speech interaction: The impact of enhanced gesture units in text-to-gesture synthesis for digital humans. *IEEE Access*, 13:145144–145157, 2025b. doi: 10.1109/ACCESS.2025.3596328.

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*, 41(6), 2022. doi: 10.1145/3550454.3555435. URL https://arxiv.org/abs/2210.01448.

Tenglong Ao, Zeyi Zhang, and Libin Liu. GestureDiffuCLIP: Gesture diffusion model with CLIP latents. *ACM Transactions on Graphics*, 42(4):1–18, 2023. doi: 10.1145/3592097. URL https://arxiv.org/abs/2303.14613.

Eiichi Asakawa, Naoshi Kaneko, Dai Hasegawa, and Shinichi Shirakawa. Evaluation of text-to-gesture generation model using convolutional neural network. *Neural Networks*, 151:365–375, 2022. doi: 10.1016/j.neunet.2022.03.041.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023. doi: 10.1109/CVPR52729.2023.01499.

Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces*, 2021. doi: 10.1109/VR50410.2021.00037.

Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. BEAT: The behavior expression animation toolkit. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 477–486, 2001. doi: 10.1145/383259.383315. URL https://www.justinecassell.com/publications/siggraph2001.final.PDF.

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

Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. DiffSHEG: A diffusion-based approach for real-time speech-driven holistic 3D expression and gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6), 2022. doi: 10.1109/JSTSP.2022.3188113.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 1597–1607. PMLR, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

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

Yongkang Cheng and Shaoli Huang. HoloGest: Decoupled diffusion and motion priors for generating holisticly expressive co-speech gestures. *arXiv preprint arXiv:2503.13229*, 2025. URL https://arxiv.org/abs/2503.13229.

Yongkang Cheng, Shaoli Huang, Xuelin Chen, Jifeng Ning, and Mingming Gong. DIDiffGes: Decoupled semi-implicit diffusion models for real-time gesture generation from speech. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2464–2472, 2025. doi: 10.1609/aaai.v39i3.32248.

Simon Clavet. Motion matching and the road to next-gen animation. Talk at the Game Developers Conference (GDC), 2016. URL https://gdcvault.com/play/1023280/Motion-Matching-and-The-Road.

Anna Deichler, Shivam Mehta, Simon Alexanderson, and Jonas Beskow. Diffusion-based co-speech gesture generation using joint text and audio representation. In *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI)*, 2023. doi: 10.1145/3577190.3616117.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Mireille Fares, Catherine Pelachaud, and Nicolas Obin. ZS-MSTM: Zero-shot style transfer for gesture animation driven by text and speech using adversarial disentanglement of multimodal style encoding. *Frontiers in Artificial Intelligence*, 6, 2023. doi: 10.3389/frai.2023.1142997. URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.1142997/full.

Ylva Ferstl and Rachel McDonnell. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pp. 93–98. Association for Computing Machinery, 2018. doi: 10.1145/3267851.3267898. URL https://www.scss.tcd.ie/rachel.mcdonnell/papers/IVA2018b.pdf.

Ylva Ferstl, Michael Neff, and Rachel McDonnell. ExpressGesture: Expressive gesture generation from speech through database matching. *Computer Animation and Virtual Worlds*, 32(3-4):e2016, 2021. doi: 10.1002/cav.2016. URL https://www.cs.ucdavis.edu/~neff/papers/ExpressGesture_Ferstl2021_preprint.pdf.

Nan Gao, Zeyu Zhao, Zhi Zeng, Shuwu Zhang, Dongdong Weng, and Yihua Bao. GesGPT: Speech gesture synthesis with text parsing from ChatGPT. *IEEE Robotics and Automation Letters*, 9(3), 2024. doi: 10.1109/LRA.2024.3359544.

Parisa Ghanad Torshizi, Laura B. Hensel, Ari Shapiro, and Stacy C. Marsella. Large language models for virtual human gesture selection. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2025.

Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F. Troje, and Marc-André Carbonneau. ZeroEGGS: Zero-shot example-based gesture generation from speech. *Computer Graphics Forum*, 42(1):206–216, 2023. doi: 10.1111/cgf.14734. URL https://arxiv.org/abs/2209.07556.

Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3497–3506, 2019. URL https://arxiv.org/abs/1906.04160.

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*, pp. 5152–5161, 2022. doi: 10.1109/CVPR52688.2022.00509.

Yueqian Guo, Tianzhao Li, Xin Lyu, Jiehaolin Chen, Zhaohan Wang, Sirui Xiao, Yurun Chen, Yezi He, Helin Li, and Fan Zhang. TRiMM: Transformer-based rich motion matching for real-time multi-modal interaction in digital humans. arXiv preprint arXiv:2506.01077, 2025.

Yueqian Guo, Tianzhao Li, Xin Lv, Jiehaolin Chen, Zhaohan Wang, Yurun Chen, Sirui Xiao, Yezi He, Helin Li, and Fan Zhang. NeRAG: Neuro-explicit retrieval-augmented generation for real-time interaction in digital humans. In *Proceedings of the 2026 International Conference on Multimedia Retrieval*, pp. 1842–1850, 2026. doi: 10.1145/3805622.3810842. URL https://doi.org/10.1145/3805622.3810842.

Ikhsanul Habibie, Mohamed Elgharib, Kripasindhu Sarkar, Ahsan Abdullah, Simbarashe Nyatsanga, Michael Neff, and Christian Theobalt. A motion matching-based framework for controllable gesture synthesis from speech. In *ACM SIGGRAPH 2022 Conference Proceedings*. Association for Computing Machinery, 2022. doi: 10.1145/3528233.3530750. URL https://vcai.mpi-inf.mpg.de/projects/SpeechGestureMatching/.

Arno Hartholt, David R. Traum, Stacy C. Marsella, Ari Shapiro, Giota Stratou, Anton Leuski, Louis-Philippe Morency, and Jonathan Gratch. All together now – introducing the virtual human toolkit. In *Intelligent Virtual Agents*, volume 8108 of *Lecture Notes in Computer Science*, pp. 368–381. Springer, 2013. doi: 10.1007/978-3-642-40415-3_33.

Laura B. Hensel, Nutchanon Yongsatianchot, Parisa Torshizi, Elena Minucci, and Stacy Marsella. Large language models in textual analysis for gesture selection. In *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI)*, 2023.

Daniel Holden, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa. Learned motion matching. *ACM Transactions on Graphics*, 39(4):53:1–53:12, 2020. doi: 10.1145/3386569.3392440.

Adam Kendon. *Gesture: Visible Action as Utterance*. Cambridge University Press, Cambridge, UK, 2004.

Patrick G. Kenny, Thomas D. Parsons, and Albert Rizzo. Human computer interaction in virtual standardized patient systems. In *Human-Computer Interaction. Interacting in Various Application Domains*, volume 5613 of *Lecture Notes in Computer Science*, pp. 514–523. Springer, 2009. doi: 10.1007/978-3-642-02583-9_56.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N. Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R. Thórisson, and Hannes Högni Vilhjálmsson. Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent Virtual Agents*, volume 4133 of *Lecture Notes in Computer Science*, pp. 205–217. Springer, 2006. doi: 10.1007/11821830_17.

Lucas Kovar, Michael Gleicher, and Frédéric Pighin. Motion graphs. *ACM Transactions on Graphics*, 21(3):473–482, 2002. doi: 10.1145/566654.566605.

Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 242–250, 2020. doi: 10.1145/3382507.3418815.

Taras Kucherenko, Rajmund Nagy, Youngwoo Yoon, Jieyeon Woo, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. The GENEA challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pp. 792–801, 2023. doi: 10.1145/3577190.3616120.

Taras Kucherenko, Pieter Wolfert, Youngwoo Yoon, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. Evaluating gesture generation in a large-scale open challenge: The GENEA challenge 2022. *ACM Transactions on Graphics*, 43(3):1–28, 2024. doi: 10.1145/3656374.

Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S. Srinivasa, and Yaser Sheikh. Talking with hands 16.2M: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 763–772, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/papers/Lee_Talking_With_Hands_16.2M_A_Large-Scale_Dataset_of_Synchronized_Body-Finger_ICCV_2019_paper.pdf.

Jina Lee and Stacy Marsella. Nonverbal behavior generator for embodied conversational agents. In *Intelligent Virtual Agents*, volume 4133 of *Lecture Notes in Computer Science*, pp. 243–255. Springer, 2006. doi: 10.1007/11821830_20.

Stephen C. Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6:731, 2015. doi: 10.3389/fpsyg.2015.00731.

Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. Audio2Gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11293–11302, 2021. URL https://arxiv.org/abs/2108.06720.

Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. BEAT: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *Computer Vision – ECCV 2022*, volume 13667 of *Lecture Notes in Computer Science*, pp. 612–630. Springer, 2022. doi: 10.1007/978-3-031-20071-7_36. URL https://link.springer.com/chapter/10.1007/978-3-031-20071-7_36.

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*, pp. 1144–1154, 2024. doi: 10.1109/CVPR52733.2024.00115. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_EMAGE_Towards_Unified_Holistic_Co-Speech_Gesture_Generation_via_Expressive_Masked_CVPR_2024_paper.html.

Pinxin Liu, Luchuan Song, Junhua Huang, Haiyang Liu, and Chenliang Xu. GestureLSM: Latent shortcut based co-speech gesture generation with spatial-temporal modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10929–10939, 2025. doi: 10.1109/ICCV51701.2025.01017. URL https://openaccess.thecvf.com/content/ICCV2025/html/Liu_GestureLSM_Latent_Shortcut_based_Co-Speech_Gesture_Generation_with_Spatial-Temporal_Modeling_ICCV_2025_paper.html.

Mykola Maslych, Mohammadreza Katebi, Christopher Lee, Yahya Hmaiti, Amirpouya Ghasemaghaei, Christian Pumarada, Janneese Palmer, Esteban Segarra Martinez, Marco Emporio, Warren Snipes, Ryan P. McMahan, and Joseph J. LaViola, Jr. Mitigating response delays in free-form conversations with LLM-powered intelligent virtual agents. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, 2025. doi: 10.1145/3719160.3736636.

Claire McLean, Makenzie Meendering, Tristan Swartz, Orri Gabbay, Alexandra Olsen, Rachel Jacobs, Nicholas Rosen, Philippe de Bree, Tony Garcia, Gadsden Merrill, Jake Sandakly, Julia Buffalini, Neham Jain, Steven Krenn, Moneish Kumar, Dejan Markovic, Evonne Ng, Fabian Prada, Andrew Saba, Siwei Zhang, Vasu Agrawal, Tim Godisart, Alexander Richard, and Michael Zollhoefer. Embody 3D: A large-scale multimodal motion and behavior dataset. Technical Report arXiv:2510.16258, arXiv, 2025. URL https://arxiv.org/abs/2510.16258.

David McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago, IL, 1992.

Shivam Mehta, Siyang Wang, Simon Alexanderson, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Diff-TTSG: Denoising probabilistic integrated speech and gesture synthesis. In *12th ISCA Speech Synthesis Workshop (SSW)*, 2023. doi: 10.21437/SSW.2023-24.

Shivam Mehta, Ruibo Tu, Simon Alexanderson, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Unified speech and gesture synthesis using flow matching. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16578–16588, 2025. doi: 10.1109/CVPR52734.2025.01545. URL https://openaccess.thecvf.com/content/CVPR2025/html/Mughal_Retrieving_Semantics_from_the_Deep_an_RAG_Solution_for_Gesture_CVPR_2025_paper.html.

M. Hamza Mughal, Rishabh Dabral, Vera Demberg, and Christian Theobalt. MIBURI: Towards expressive interactive gesture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

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)*, 2024.

Rajmund Nagy, Taras Kucherenko, Birger Moell, André Pereira, Hedvig Kjellström, and Ulysses Bernardet. A framework for integrating gesture generation models into interactive conversational agents. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2021.

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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Findings Track*, 2026. URL https://arxiv.org/abs/2511.01233.

Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics*, 27(1): 5:1–5:24, 2008. doi: 10.1145/1330511.1330516.

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*, pp. 1001–1010, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ng_From_Audio_to_Photoreal_Embodiment_Synthesizing_Humans_in_Conversations_CVPR_2024_paper.html.

Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. A comprehensive review of data-driven co-speech gesture generation. *Computer Graphics Forum (Proc. Eurographics)*, 2023. doi: 10.1111/cgf.14776.

Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Muhammad Usama Saleem, Mayur Jagdishbhai Patel, Ekkasit Pinyoanuntapong, Zhongxing Qin, Li Yang, Hongfei Xue, Ahmed Helmy, Chen Chen, and Pu Wang. LiveGesture: Streamable co-speech gesture generation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2264–2273, 2026.

Matthew Stone, Doug DeCarlo, Insuk Oh, Christian Rodriguez, Adrian Stere, Alyssa Lees, and Chris Bregler. Speaking with hands: creating animated conversational characters from recordings of human performance. *ACM Transactions on Graphics*, 23(3):506–513, 2004. doi: 10.1145/1015706.1015753.

William R. Swartout. Lessons learned from virtual humans. *AI Magazine*, 31(1):9–20, 2010. doi: 10.1609/aimag.v31i1.2284.

Hitoshi Teshima, Naoki Wake, Diego Thomas, Yuta Nakashima, Hiroshi Kawasaki, and Katsushi Ikeuchi. ACT2G: Attention-based contrastive learning for text-to-gesture generation. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(3), 2023. doi: 10.1145/3606940. URL https://arxiv.org/abs/2309.16162.

Marcus Thiebaux, Andrew Marshall, Stacy Marsella, and Marcelo Kallmann. Smartbody: Behavior realization for embodied conversational agents. In *Proceedings of Autonomous Agents and Multi-Agent Systems*, 2008. URL https://ccs.neu.edu/~marsella/publications/abstracts/smartbody08-abstract.html.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*, 2019. URL https://arxiv.org/abs/1908.08962.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

Siyang Wang, Simon Alexanderson, Joakim Gustafson, Jonas Beskow, Gustav Eje Henter, and Éva Székely. Integrated speech and gesture synthesis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pp. 177–185, 2021. doi: 10.1145/3462244.3479914.

J. Windle, I. Matthews, and S. Taylor. Llanimation: Llama driven gesture animation. *Computer Graphics Forum*, 43(8):e15167, 2024. doi: 10.1111/cgf.15167.

Zunnan Xu, Yukang Lin, Haonan Han, Sicheng Yang, Ronghui Li, Yachao Zhang, and Xiu Li. MambaTalk: Efficient holistic gesture synthesis with selective state space models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 20055–20080, 2024. doi: 10.52202/079017-0633. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/23c9c94227f937cfb50592a15e7fbb63-Abstract-Conference.html.

Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, and Haolin Zhuang. QPGesture: Quantization-based and phase-guided motion matching for natural speech-driven gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2321–2330, 2023a. URL https://openaccess.thecvf.com/content/CVPR2023/html/Yang_QPGesture_Quantization-Based_and_Pha

se-Guided_Motion_Matching_for_Natural_Speech-Driven_Gesture_CVPR_2023_paper.html.

Sicheng Yang, Haiwei Xue, Zhensong Zhang, Minglei Li, Zhiyong Wu, Xiaofei Wu, Songcen Xu, and Zonghong Dai. The DiffuseStyleGesture+ entry to the GENEA challenge 2023. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pp. 779–785. Association for Computing Machinery, 2023b. doi: 10.1145/3577190.3616114. URL https://arxiv.org/abs/2308.13879.

Sicheng Yang, Zunnan Xu, Haiwei Xue, Yongkang Cheng, Shaoli Huang, Mingming Gong, and Zhiyong Wu. FreeTalker: Controllable speech and text-driven gesture generation based on diffusion models for enhanced speaker naturalness. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7945–7949, 2024. doi: 10.1109/ICASSP48485.2024.10447978. URL https://ieeexplore.ieee.org/document/10447978.

Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J. Black. Generating holistic 3D human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *International Conference on Robotics and Automation (ICRA)*, pp. 4303–4309, 2019. doi: 10.1109/ICRA.2019.8793720.

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics*, 39(6), 2020. doi: 10.1145/3414685.3417838. URL https://arxiv.org/abs/2009.02119.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18123–18133, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Zhai_LiT_Zero-Shot_Transfer_With_Locked-Image_Text_Tuning_CVPR_2022_paper.html.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2M-GPT: Generating human motion from textual descriptions with discrete representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Zeyi Zhang, Tenglong Ao, Yuyao Zhang, Qingzhe Gao, Chuan Lin, Baoquan Chen, and Libin Liu. Semantic gesticulator: Semantics-aware co-speech gesture synthesis. *ACM Transactions on Graphics*, 43(4), 2024. doi: 10.1145/3658134. URL https://pku-mocca.github.io/Semantic-Gesticulator-Page/.

Chi Zhou, Tengyue Bian, and Kang Chen. GestureMaster: Graph-based speech-driven gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pp. 764–770. Association for Computing Machinery, 2022. doi: 10.1145/3536221.3558063.

A Architecture, optimization, and checkpoint details

A.1 Data and units

The selected configuration uses BEAT English motion at 30 fps, retains 50 joints after removing toes, and subtracts the hips independently in every frame. The speaker split is 17 train, five validation, and five test speakers. Three additional training candidates with frozen mocap were excluded before learning. Both monologue and conversation recordings are retained.

Training and memory units are fixed two-second windows (60 frames) anchored at word-group starts; overlap between units is allowed. Memory storage adds ten frames (one third second) on each side when available. The latent always encodes only the 60-frame core. The fusion dataset constructs every run of three consecutive chunk indices within a take and samples one of four stored 10-minute speaker descriptors for each item.

A.2 Motion encoder

The motion encoder takes frame tokens of width 150 and applies sinusoidal positions and three Transformer encoder layers with five attention heads, 512-dimensional feed-forward blocks, dropout 0.1, masked mean pooling, and a Linear(150,10)-BatchNorm-LeakyReLU head. The main pretraining arm, with 3-D augmented views, draws 20,000 random windows per epoch from the training takes, with lengths sampled from 2, 3, and 4 s. One view is clean; the other receives Gaussian noise whose variance is sampled from $\{0.001, 0.01, 0.1\}$, followed by a temporal shift that places a two-thirds crop at a random location in a same-length canvas. The batch size is 512. Optimization is AdamW with learning rate 5×10^{-4}, weight decay 10^{-4}, cosine annealing, temperature 0.1, and a nominal 1,000-epoch cap with patience 10. The main arm uses contrastive loss only. After checkpoint selection, the encoder is exported, loaded in evaluation mode with gradients disabled, and used in batches of 256 to encode unit cores.

A.3 Speaker encoder and enrollment

The main speaker-encoder arm, which reads raw frames without a convolutional stride, downsamples motion to 15 fps. Each one-minute segment therefore has 900 frame tokens. A per-frame linear projection maps 150 channels to width 256; four Transformer layers use eight heads and feed-forward width 1,024. Attention pooling reduces each minute to one vector. Across the ten minutes in a training chunk, one-query attention pooling acts as a set aggregator without positional encoding, followed by a 768-dimensional projection and BatchNorm. Input standardization statistics come only from the training split and are stored in the checkpoint.

Each batch contains five distinct speaker identities and two distinct 10-minute chunks for each identity. The loss is symmetric NT-Xent at temperature 0.1. The selected raw arm uses AdamW at 2×10^{-4}, weight decay 0.01, cosine annealing, and 500 epochs with early stopping disabled. The encoder is then frozen. Enrollment samples four contiguous windows for each requested reference duration and stores four 768-D vectors; the main predictor uses ten-minute vectors. The current generator chooses one vector by a descriptor index (default zero), rather than averaging the four.

A.4 Predictor and validation selection

The text model is BERT-mini (four layers, hidden width 256; identifier in Appendix A.6) and is fine-tuned. Attention-masked mean pooling produces one text vector. Concatenated text, style, and previous-latent features pass through Linear(1034,256)-GELU-Linear(256,10). Optimization uses AdamW with learning rate 5×10^{-4}, weight decay 0.01, cosine annealing, temperature 0.1, training batch 128, validation batch 32, a 1,000-epoch cap, and patience 10.

The designated run is seed 1 (numeric seed 43). Its manifest resolves the predictor to the epoch-2 checkpoint; the run stopped after 13 epochs and the selected epoch's validation retrieval accuracy is 0.0177951417863369. Selection maximizes that metric within the designated run.

For clarity, the validation metric is computed by flattening a batch of 32 three-step sequences into 96 predicted and 96 target latents. It counts whether each prediction's cosine-nearest target in that

flattened batch is its paired target. It does not search a speaker's clustered memory. The final partial validation batch may contain fewer candidates.

A.5 Memory and decoder

For a speaker with n_s units, bisecting k-means uses $k = \max(2, \min(n_s, \text{round}(n_s/12.5)))$ on normalized 10-D latents with the largest-cluster splitting strategy. Members of clusters smaller than five are successively reassigned to their nearest surviving centroid, after which centroids are recomputed and normalized. Retrieval ranks centroids by cosine similarity, skips blocklisted members, and uniformly samples one member of the first nonempty cluster.

The shared H/D decoder maps the 10-D latent to mean and log-variance vectors of width 128. During training it samples the variational bottleneck; in evaluation it uses the mean. A two-layer Transformer decoder with four heads and feed-forward width 512 receives 60 sinusoidal position queries and one latent memory token, then projects to 150-D frames. It is trained separately on the frozen main latent space, using training-speaker unit cores and validation-speaker unit cores for validation. AdamW uses learning rate 5×10^{-4}, weight decay 0.01, batch 128, cosine annealing, a 300-epoch cap, and patience 10. The loss is masked coordinate MSE plus 10^{-3} KL. The designated manifest pairs H/D with the separately trained main decoder over the same main latent space as the predictor's memory. No checkpoint-specific epoch or validation value is attributed to that decoder.

A.6 Reproduction entry points

Reproduction begins from configuration `configs/v2_beat30.yaml` with seed 43. The designated artifacts are the text model `google/bert_uncased_L-4_H-256_A-4`; predictor checkpoint `best-epoch002.ckpt` of run `main/s1`; latent space `main-bf21a9` from motion-encoder run `clr/main/s0`; speaker-encoder checkpoint `best-epoch049.ckpt`; decoder `decoder/main`; and the FGD evaluator with SHA-256 `ef8814eec97ac5c4b10cabd0b294da0971ea89a8c003713564c7c6a6d57e0daa`. The configuration uses the main contrastive latent space, the raw speaker-encoder arm, and fine-tuned BERT-mini. Predictor checkpoint selection maximizes validation retrieval accuracy within the designated run. The shared H/D decoder must be trained on that same latent space. Generation accepts text or a text file, a speaker key for enrolled-artifact lookup, a reference-descriptor index, and a synthesis mode; it accepts no audio input.

A.7 Test-time memory and length protocol

The evaluated test speaker is enrolled from that speaker's reference motion. Its memory contains units from the evaluation corpus, but while a take is generated, every unit originating from that same take is blocklisted; units from the speaker's other takes remain eligible. The driver restores the blocklist after each take and records the excluded and available counts. This leave-one-take-out rule applies while all three modes are run, although selected IDs do not affect D's output.

Generation itself follows text-chunk length. For reference-aligned evaluation only, an output longer than its source take is truncated to the take's frame count; shorter output is neither padded nor stretched. The sidecar records generated, reference, and written frame counts. Task G's positional FGD reads the resulting root-centred 50-joint positions; it does not invoke the optional native rotation/root recovery path. Resource timing calls the same predictor, retrieval/decoding, and blending implementation directly, with the full loaded library and without Task G's take-specific blocklist, truncation, or export conversion. It reports the 69-second input duration separately from generated motion duration.

B Comparison protocol and exclusions

B.1 SPEAKER PARTITIONS AND REFERENCE ACCESS

Table 3: Speaker partitions and reference access

Partition	BEAT speaker identifiers
Training	1, 2, 3, 4, 6, 7, 8, 9, 11, 13, 14, 15, 16, 21, 23, 24, 30
Validation	5, 10, 18, 20, 27
Test	12, 17, 22, 25, 28
Excluded from the configured split	19, 26, 29

BEAT2-native EMAGE, GestureLSM, and RAG-Gesture adaptations use 15 training speakers because speakers 8 and 14 are unavailable in the corresponding BEAT2 data. Author-pretrained tokenizers are retained where specified below; the speaker-disjoint claim describes adaptation training and does not certify the absence of every test speaker from every frozen component's pretraining. The common 138-take comparison is the intersection of available exports. Full-support ablation results are reported separately.

ConGRets uses reference-derived global speaker conditioning in every mode. Memory-based and hybrid synthesis select within that speaker's enrolled motion under the leave-one-take-out rule of Appendix A.7. The generation tables use the leave-one-take-out exports; earlier full-memory generation exports are archived and excluded from these tables. This is leave-one-take-out selection within an enrolled test speaker, not generation for an entirely unobserved speaker without reference data. The memory's remaining takes need not belong to a separately collected enrollment session. The speaker descriptor also comes from reference motion; the protocol does not establish a separate-session or take-disjoint descriptor enrollment. Habibie's motion-matching database instead uses training-speaker motion, while ZS-MSTM is given a local reference clip per utterance. These are different reference conditions.

The take exclusion above applies to sampled units. Cluster centroids are fitted once on the complete speaker memory, including latents of units from the scored take. The ten-minute descriptor reference window is also not recomputed per scored take and may overlap it in all three modes. No experiment isolates the effect of either overlap; the protocol must not be described as completely take-disjoint enrollment.

B.2 BASELINE IMPLEMENTATION AND CONDITIONING

Table 4: Baseline implementation and conditioning

Evaluated system	Implementation and adaptation	Conditioning / frozen components	Material qualification
DiffuseStyle Gesture+	Upstream diffusion sampling path; 8.5 GPU-h adaptation, cap-limited	Audio; unseen-speaker style-dropout path	Generated with unconditional style rather than assigning a training-speaker identity. Normalization statistics were rebuilt on the same training frame set.
EMAGE	BEAT2 generator adaptation; 8.5 GPU-h, cap-limited	Audio; five author-supplied frozen VQ/VAE tokenizers	Training-cache construction differs from the other BEAT2 adaptations; early-stopping threshold semantics also differ.

Continued on next page

Table 4 (continued)

Evaluated system	Implementation and adaptation	Conditioning / frozen components	Material qualification
GestureLSM	BEAT2 generator adaptation; recorded 0.41 GPU-h, early-stopped	Audio, transcript, word timings, seed pose, speaker conditioning; three author-supplied frozen RVQ-VAEs	Its inference path seeds the initial window using 16 ground-truth motion frames and passes some non-body channels through. Root-centered body-joint evaluation removes root/face comparisons, but not the initial-pose advantage.
HoloGest	Upstream sampling architecture with a locally implemented trainer; recorded 4.54 GPU-h	Audio, transcript, and frozen text heads	The original ASR and semantic-discrimination tools were unavailable; transcript conditioning and substitute heads are documented adaptations. Native data/rig provenance is not fully resolved in the manifest, limiting strict matched-data claims.
RAG-Gesture	Four part VAEs and diffusion adapted on BEAT2; stage-specific limits, recorded 16 + 5 GPU-h	Audio, transcript, precomputed discourse/prominence annotations	The full chain did not share a single 8.5-hour total cap. Annotation-generation tooling is not included in the timing. Its prebuilt retrieval database is an additional preparation dependency.
Habibie et al.	Local kNN + conditional-GAN reimplementation; recorded 1.2 GPU-h for cGAN training	Audio; training-speaker motion database; k = 1 in the evaluated run	Comparison is with this documented reimplementation, not an assertion of reproducing all author results.
ZS-MSTM	Local reimplementation; recorded 0.54 GPU-h	Audio, transcript, timings, per-utterance style reference; frozen AST and BERT	Local reference conditioning differs from ConGRets' persistent global descriptor.

Recorded GPU-hours are the benchmark training records, not normalized estimates of equivalent work across different GPU models or a complete accounting of upstream pretraining. Early-stopping criteria and training caps differ by method and stage. DSG+, EMAGE, and RAG-Gesture are reported as cap-limited, not converged. The common panel matches take identities, output coordinate convention, frame rate, and metric implementations; it does not equalize all information, pretraining, or training budgets.

SMPL-X-native outputs are converted through the body model, mirrored into the common coordinate convention, and mapped to the 50-joint comparison skeleton. This introduces a representation floor absent from the direct mocap path. A separate geometric-distance diagnostic on 141 takes reports 0.1439 for BEAT2 ground truth through that conversion. This is a geometric FD value, not a position-FGD correction, and is not subtracted from Table 1.

B.3 Excluded and separately reported systems

Table 5: Excluded and separately reported systems

System or checkpoint	Reason it does not enter the seven-system ranking
ConvoFusion local retrain	Preprocessing unit and forward-kinematics defects led to a collapsed VAE and near-static output. Its measurements describe a failed local retrain, not the original method's capability.
ConvoFusion author weights	Available exports cover 112 of 149 takes unevenly. Four test speakers overlap the authors' training list. The separately scored unseen speaker-25 subset contains 20 takes and obtains FGD 1.086 and pose diversity 0.03793; it is not take-matched to Table 1 and is not ranked there. No timing from the collapsed local retrain is paired with these author-weight scores.
PersonaGest, MECo, miburi, SemTalk local artifacts	Available checkpoints implement codecs or partial branches rather than complete speech-to-gesture generation. Reconstruction/component timing is not a generator comparison.
ZeroEGGS	No executable adapter was established for this evaluation.
Seamless	No evaluated BEAT output path was established.
AMUSE	Not included in the completed adaptation/quality protocol; usable author weights were unavailable in that workflow.
Semantic Gesticulator	Outside the completed evaluation scope.

Absence from the ranking is not evidence of inferior quality or efficiency. Related-work coverage is broader than the set of executable benchmark configurations.

C Complete final ablations

The inventory contains 24 recipes expanded to 40 predictor runs. All 40 runs were measured in each of the three modes; flagged cells remain in the accounting but are excluded from intended-configuration ranking. M = memory-based synthesis, H = hybrid synthesis, D = direct synthesis. `DIV` marks a divergent direct-latent trajectory; `DEC` marks an E12 run evaluated with the unfine-tuned decoder. Such rows are retained for accounting, not ranked as valid outputs of the intended configuration. Their stored finite measurements are shown with the status, including reduced take support where applicable. Zero flagged failures occurred for direct synthesis in the three main runs.

The generation results use the ablation test panel, normally 149 takes; the exact FGD support is given per row. A common number of takes does not imply identical generated duration. Pose diversity is the stored mean over the available takes. GPU/CPU measurements are the September 18 records on the canonical input, shown to one decimal millisecond from the rounded cost ledger; Appendix D retains the main repeat-level precision. Training checkpoint verification is separate from the generation-run validity flags.

C.1 Main reference

Table 6: Main reference

Run	Val. in-batch top-1 ↑	Epoch
main/s0	0.01890432	2
main/s1	0.01779514	2
main/s2	0.01904900	1

Table 7: Main reference

Run	Mode	FGD ↓	Pose diversity	Takes	Params M	GPU ms	CPU ms	Status
main/s0	M	0.07169	0.06317	149	11.438	17.8	22.9	Valid
main/s0	H	0.36906	0.03162	149	11.989	18.4	32.3	Valid
main/s0	D	0.57557	0.02768	149	11.989	14.0	27.8	Valid
main/s1	M	0.15225	0.07836	149	11.438	19.2	22.8	Valid
main/s1	H	0.37336	0.04445	149	11.989	18.4	30.5	Valid
main/s1	D	0.86949	0.03485	149	11.989	14.6	26.7	Valid
main/s2	M	0.18372	0.08059	149	11.438	17.8	23.0	Valid
main/s2	H	0.39008	0.04472	149	11.989	18.8	30.0	Valid
main/s2	D	0.65273	0.03200	149	11.989	13.2	26.7	Valid

C.2 TEXT ENCODER ALTERNATIVES

Table 8: Text encoder alternatives

Run	Val. in-batch top-1 ↑	Epoch
e5/albert_base	0.01557677	2
e5/albert_base_frozen	0.01740934	2
e5/bert_128	0.01769869	11
e5/bert_base	0.01658951	26
e5/bert_base_frozen	0.01842208	8
e5/bert_mini_frozen	0.01875965	29
e5/bert_small	0.01933836	11
e5/bert_tiny	0.01832562	1
e5/bge_small	0.01494985	0
e5/bge_small_frozen	0.01880787	20
e5/distilgpt2	0.01919368	2
e5/electra_small	0.01692708	40
e5/minilm_l6	0.01827739	5
e5/minilm_l6_frozen	0.01861497	18

Table 9: Text encoder alternatives

Run	Mode	FGD ↓	Pose diversity	Takes	Params M	GPU ms	CPU ms	Status
e5/albert_base	M	0.21915	0.06744	149	12.082	21.2	84.6	Valid
e5/albert_base	H	0.55320	0.03632	149	12.634	21.5	88.6	Valid
e5/albert_base	D	0.97537	0.02259	149	12.634	16.7	80.2	Valid
e5/albert_base_frozen	M	0.12778	0.07999	149	12.082	20.9	77.0	Valid
e5/albert_base_frozen	H	0.38695	0.04810	149	12.634	21.6	91.6	Valid
e5/albert_base_frozen	D	0.42693	0.05342	149	12.634	16.4	85.1	Valid
e5/bert_128	M	0.09951	0.06708	149	5.017	18.9	17.9	Valid
e5/bert_128	H	0.42846	0.03666	149	5.569	19.2	26.1	Valid
e5/bert_128	D	1.16094	0.04984	149	5.569	14.3	23.1	DIV
e5/bert_base	M	0.21691	0.05233	149	109.88	20.6	77.2	Valid
e5/bert_base	H	0.56153	0.02271	149	110.43	22.2	83.1	Valid
e5/bert_base	D	2.67253	0.05499	131	110.43	17.1	84.8	DIV
e5/bert_base_frozen	M	0.13549	0.08298	149	109.88	21.3	77.6	Valid
e5/bert_base_frozen	H	0.37752	0.05219	149	110.43	21.0	112.0	Valid
e5/bert_base_frozen	D	0.80454	0.04215	149	110.43	15.9	83.5	Valid

Continued on next page

Table 9 (continued)

Run	Mode	FGD ↓	Pose diversity	Takes	Params M	GPU ms	CPU ms	Status
e5/bert_mini_frozen	M	0.05939	0.08097	149	11.438	17.6	30.7	Valid
e5/bert_mini_frozen	H	0.31256	0.05254	149	11.989	18.6	31.3	Valid
e5/bert_mini_frozen	D	2.07071	0.05166	135	11.989	13.2	28.0	DIV
e5/bert_small	M	0.07446	0.07072	149	29.097	19.5	55.1	Valid
e5/bert_small	H	0.36455	0.03436	149	29.648	18.7	36.1	Valid
e5/bert_small	D	1.00554	0.03226	149	29.648	13.3	32.6	DIV
e5/bert_tiny	M	0.14727	0.08533	149	4.621	17.1	68.5	Valid
e5/bert_tiny	H	0.37166	0.05188	149	5.172	17.9	24.2	Valid
e5/bert_tiny	D	0.67450	0.04289	149	5.172	12.7	20.5	Valid
e5/bge_small	M	0.18387	0.06892	149	33.66	21.6	54.0	Valid
e5/bge_small	H	0.50895	0.03267	149	34.212	22.2	56.1	Valid
e5/bge_small	D	1.02686	0.01096	149	34.212	17.7	45.8	Valid
e5/bge_small_frozen	M	0.10040	0.06349	149	33.66	22.2	49.3	Valid
e5/bge_small_frozen	H	0.39842	0.03508	149	34.212	22.3	50.5	Valid
e5/bge_small_frozen	D	0.64333	0.03339	144	34.212	16.9	130.0	DIV
e5/distilgpt2	M	0.12151	0.08568	149	82.311	18.5	52.1	Valid
e5/distilgpt2	H	0.35334	0.05331	149	82.862	19.6	90.8	Valid
e5/distilgpt2	D	1.97489	0.05785	149	82.862	14.0	56.8	Valid
e5/electra_small	M	0.16910	0.05677	149	13.751	20.6	36.8	Valid
e5/electra_small	H	0.51688	0.02284	149	14.302	21.5	42.2	Valid
e5/electra_small	D	2.08966	0.05457	38	14.302	16.5	37.6	DIV
e5/minilm_l6	M	0.10726	0.06389	149	23.014	21.7	29.2	Valid
e5/minilm_l6	H	0.43880	0.02736	149	23.565	20.5	37.6	Valid
e5/minilm_l6	D	1.62592	0.03392	149	23.565	15.7	36.3	DIV
e5/minilm_l6_frozen	M	0.07532	0.07034	149	23.014	20.4	31.7	Valid
e5/minilm_l6_frozen	H	0.37220	0.03967	149	23.565	21.3	36.0	Valid
e5/minilm_l6_frozen	D	0.87741	0.05952	149	23.565	15.6	34.4	DIV

C.3 Width and text adaptation

Table 10: Width and text adaptation

Run	Val. in-batch top-1 ↑	Epoch
w32/finetuned/s0	0.01639661	10
w32/finetuned/s1	0.01687886	9
w32/finetuned/s2	0.01678241	16
w32/frozen/s0	0.01567322	33
w32/frozen/s1	0.01514275	12
w32/frozen/s2	0.01538387	17

Table 11: Width and text adaptation

Run	Mode	FGD ↓	Pose diversity	Takes	Params M	GPU ms	CPU ms	Status
w32/finetuned/s0	M	0.04514	0.07605	149	11.449	17.9	21.7	Valid
w32/finetuned/s0	H	0.22527	0.05189	149	12.006	19.2	30.1	Valid
w32/finetuned/s0	D	0.36897	0.02451	149	12.006	13.3	27.8	Valid
w32/finetuned/s1	M	0.06432	0.07483	149	11.449	17.5	22.4	Valid
w32/finetuned/s1	H	0.25928	0.04904	149	12.006	19.6	30.2	Valid
w32/finetuned/s1	D	0.46199	0.02169	149	12.006	14.0	27.3	Valid

Continued on next page

Table 11 (continued)

Run	Mode	FGD ↓	Pose diversity	Takes	Params M	GPU ms	CPU ms	Status
w32/finetuned/s2	M	0.04623	0.06993	149	11.449	17.5	22.7	Valid
w32/finetuned/s2	H	0.25514	0.04425	149	12.006	18.9	30.5	Valid
w32/finetuned/s2	D	0.36096	0.02359	149	12.006	13.3	299.0	DIV
w32/frozen/s0	M	0.10425	0.08784	149	11.449	18.2	22.4	Valid
w32/frozen/s0	H	0.27553	0.06635	149	12.006	19.9	29.8	Valid
w32/frozen/s0	D	0.49614	0.06460	135	12.006	13.2	26.7	DIV
w32/frozen/s1	M	0.06497	0.07612	149	11.449	17.5	22.0	Valid
w32/frozen/s1	H	0.28568	0.05504	149	12.006	18.4	30.5	Valid
w32/frozen/s1	D	0.78061	0.04339	149	12.006	13.3	28.4	DIV
w32/frozen/s2	M	0.05781	0.08446	149	11.449	18.0	22.7	Valid
w32/frozen/s2	H	0.24311	0.06471	149	12.006	18.7	29.0	Valid
w32/frozen/s2	D	0.61604	0.05294	148	12.006	13.3	26.6	DIV

C.4 Prediction objective

Table 12: Prediction objective

Run	Val. in-batch top-1 ↑	Epoch
e11/mse	0.01620370	12
e11/nt_xent_mse/s0	0.01789159	2
e11/nt_xent_mse/s1	0.01803627	0
e11/nt_xent_mse/s2	0.01808449	5
e11/nt_xent_perceptual/s0	0.01837384	1
e11/nt_xent_perceptual/s1	0.01774691	0
e11/nt_xent_perceptual/s2	0.01779514	5
e11/perceptual	0.01601081	4

Table 13: Prediction objective

Run	Mode	FGD ↓	Pose diversity	Takes	Params M	GPU ms	CPU ms	Status
e11/mse	M	0.10886	0.05366	149	11.438	17.5	23.8	Valid
e11/mse	H	0.47684	0.02101	149	11.989	18.3	29.0	Valid
e11/mse	D	1.50383	0.02903	135	11.989	13.3	27.3	DIV
e11/nt_xent_mse/s0	M	0.09134	0.07143	149	11.438	18.0	21.7	Valid
e11/nt_xent_mse/s0	H	0.36239	0.03743	149	11.989	19.6	29.3	Valid
e11/nt_xent_mse/s0	D	0.46120	0.02377	149	11.989	13.3	26.8	Valid
e11/nt_xent_mse/s1	M	0.16187	0.05973	149	11.438	17.5	22.3	Valid
e11/nt_xent_mse/s1	H	0.47956	0.02596	149	11.989	18.9	32.1	Valid
e11/nt_xent_mse/s1	D	0.57324	0.01635	149	11.989	13.4	25.2	Valid
e11/nt_xent_mse/s2	M	0.06525	0.07010	149	11.438	18.6	22.4	Valid
e11/nt_xent_mse/s2	H	0.35314	0.03840	149	11.989	19.0	29.6	Valid
e11/nt_xent_mse/s2	D	0.65512	0.02373	149	11.989	14.3	30.4	Valid
e11/nt_xent_perceptual/s0	M	0.08399	0.06421	149	11.438	17.7	22.5	Valid
e11/nt_xent_perceptual/s0	H	0.41564	0.03160	149	11.989	18.3	30.4	Valid
e11/nt_xent_perceptual/s0	D	0.51815	0.01812	149	11.989	13.3	30.3	Valid
e11/nt_xent_perceptual/s1	M	0.20136	0.07158	149	11.438	18.3	22.7	Valid
e11/nt_xent_perceptual/s1	H	0.51834	0.03574	149	11.989	19.5	30.2	Valid
e11/nt_xent_perceptual/s1	D	0.59373	0.02138	149	11.989	13.6	25.1	Valid
e11/nt_xent_perceptual/s2	M	0.11719	0.07999	149	11.438	17.8	23.1	Valid

Continued on next page

Table 13 (continued)

Run	Mode	FGD ↓	Pose diversity	Takes	Params M	GPU ms	CPU ms	Status
e11/nt_xent_perceptual/s2	H	0.37321	0.04909	149	11.989	18.5	30.9	Valid
e11/nt_xent_perceptual/s2	D	0.44145	0.03165	149	11.989	13.5	26.7	Valid
e11/perceptual	M	0.17173	0.05597	149	11.438	17.9	23.3	Valid
e11/perceptual	H	0.57641	0.02056	149	11.989	18.5	30.4	Valid
e11/perceptual	D	1.39268	0.02144	135	11.989	16.1	26.4	DIV

C.5 Decoder adaptation

Table 14: Decoder adaptation

Run	Val. in-batch top-1 ↑	Epoch
e12/nt_xent_mse_ft/s0	0.01765046	7
e12/nt_xent_mse_ft/s1	0.01822917	0
e12/nt_xent_mse_ft/s2	0.01871142	3
e12/nt_xent_perceptual_ft/s0	0.01731289	13
e12/nt_xent_perceptual_ft/s1	0.01808450	3
e12/nt_xent_perceptual_ft/s2	0.01721644	1

Table 15: Decoder adaptation

Run	Mode	FGD ↓	Pose diversity	Takes	Params M	GPU ms	CPU ms	Status
e12/nt_xent_mse_ft/s0	M	0.04959	0.06871	149	11.438	17.6	22.3	Valid
e12/nt_xent_mse_ft/s0	H	0.32623	0.03658	149	11.989	18.4	30.4	DEC
e12/nt_xent_mse_ft/s0	D	0.46920	0.02532	149	11.989	13.2	34.3	DIV, DEC
e12/nt_xent_mse_ft/s1	M	0.14671	0.06332	149	11.438	17.9	22.6	Valid
e12/nt_xent_mse_ft/s1	H	0.42756	0.03178	149	11.989	18.5	30.2	DEC
e12/nt_xent_mse_ft/s1	D	0.60425	0.01674	149	11.989	13.3	269.0	DEC
e12/nt_xent_mse_ft/s2	M	0.04647	0.07155	149	11.438	18.5	22.3	Valid
e12/nt_xent_mse_ft/s2	H	0.29660	0.04138	149	11.989	18.5	29.6	DEC
e12/nt_xent_mse_ft/s2	D	0.44273	0.02827	149	11.989	13.6	26.4	DEC
e12/nt_xent_perceptual_ft/s0	M	0.05109	0.06675	149	11.438	18.8	22.4	Valid
e12/nt_xent_perceptual_ft/s0	H	0.33883	0.03409	149	11.989	18.3	29.8	DEC
e12/nt_xent_perceptual_ft/s0	D	0.58930	0.03178	137	11.989	13.2	25.9	DIV, DEC
e12/nt_xent_perceptual_ft/s1	M	0.14770	0.07948	149	11.438	17.7	22.1	Valid
e12/nt_xent_perceptual_ft/s1	H	0.36994	0.04405	149	11.989	18.6	29.4	DEC
e12/nt_xent_perceptual_ft/s1	D	0.62002	0.02845	149	11.989	13.7	26.3	DIV, DEC
e12/nt_xent_perceptual_ft/s2	M	0.08524	0.07457	149	11.438	18.3	22.8	Valid
e12/nt_xent_perceptual_ft/s2	H	0.33183	0.04226	149	11.989	18.4	30.4	DEC
e12/nt_xent_perceptual_ft/s2	D	0.56750	0.03226	149	11.989	14.3	25.4	DEC

C.6 DECODER PROVENANCE

Table 16: Decoder provenance

Run	Val. in-batch top-1 ↑	Epoch
e2d/nt_xent_mse_jointdec/s0	0.01765047	13
e2d/nt_xent_mse_jointdec/s1	0.01929012	0
e2d/nt_xent_mse_jointdec/s2	0.01842207	0

Table 17: Decoder provenance

Run	Mode	FGD ↓	Pose diversity	Takes	Params M	GPU ms	CPU ms	Status
e2d/nt_xent_mse_jointdec/s0	M	0.08910	0.07320	149	11.438	17.5	22.8	Valid
e2d/nt_xent_mse_jointdec/s0	H	0.51937	0.02829	149	11.989	18.4	29.9	Valid
e2d/nt_xent_mse_jointdec/s0	D	2.54276	0.04646	148	11.989	13.2	26.8	DIV
e2d/nt_xent_mse_jointdec/s1	M	0.17565	0.06928	149	11.438	17.8	22.3	Valid
e2d/nt_xent_mse_jointdec/s1	H	0.54046	0.02633	149	11.989	18.4	29.7	Valid
e2d/nt_xent_mse_jointdec/s1	D	0.69181	0.01428	149	11.989	13.3	26.1	Valid
e2d/nt_xent_mse_jointdec/s2	M	0.11842	0.07579	149	11.438	18.4	23.4	Valid
e2d/nt_xent_mse_jointdec/s2	H	0.55333	0.02863	149	11.989	18.6	30.1	Valid
e2d/nt_xent_mse_jointdec/s2	D	0.67475	0.02071	149	11.989	13.1	35.5	Valid

D METRIC AND RESOURCE REPRODUCIBILITY

D.1 POSITION-SPACE FGD

Let the frozen evaluator produce features with empirical means and covariances (μ_r, Σ_r) and (μ_g, Σ_g) for reference and generated motion. We compute

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

The skeleton-aware convolutional autoencoder follows the architectural pattern of the EMAGE evaluator (Liu et al., 2024) but is independently implemented and trained on the comparison positions. Each input frame has 50 joints with three coordinates in metres and is root-centered. The encoder has four blocks, per-edge widths 6, 6, 12, 12, kernel size 4, skeleton neighborhood distance 2, and a 204-dimensional output feature. Group-normalization group counts use valid divisors for each channel width. Training uses 64-frame windows with stride 20, Adam with learning rate 10^{-4}, betas (0.5, 0.999), batch size 32, seed 2021, and a 1,000-epoch ceiling. The reconstruction objective combines position error with velocity and acceleration terms; the checkpoint is selected by validation reconstruction. The retained evaluator is epoch 117 with validation MPJPE 0.005241 m.

At scoring time, each take is cropped to a multiple of 32 frames and encoded as a whole sequence. The network emits one feature per 16 frames; all features are pooled across takes. The generation and reference feature populations can have different lengths. At least twice as many features as feature

dimensions are required for either population. The evaluator is pinned by the SHA-256 listed in Appendix A.6. This evaluator is separate from the ConGRets motion encoder and its generation decoder.

D.2 Diversity, retrieval, and temporal support

Pose diversity samples up to 1,000 frame pairs per take under seed 0 and computes mean absolute difference per coordinate, then averages over takes. Sampling uses independently drawn indices with self-pairs redirected to the next frame; it is not exhaustive all-pairs diversity. Larger values are not necessarily better: motion corruption can also increase the score.

Validation retrieval accuracy normalizes the predicted and target motion latents, computes their cosine-similarity matrix, and records whether each prediction's nearest target is its paired target. With validation batches of 32 three-unit items, the candidate pool contains up to 96 targets. This is a teacher-forced in-batch diagnostic, not full-memory retrieval or a human semantic judgment. Batch composition and the number of candidates are part of its protocol. Distinct valid motions may correspond to the same text, so an incorrect paired-target decision is not equivalent to an inappropriate gesture.

Generation selection excludes the scored take, while retrieval diagnostics retain the full memory. These answer different questions. The position-space FGD pools available generated features; time-aligned diagnostics, where reported, use the shared temporal support. Exports truncate a stream exceeding the take's duration and do not pad shorter streams. The decoder modes produce shorter streams than the reference takes under the fixed unit/crossfade schedule. Timing and quality therefore do not establish exact utterance-duration reproduction or audio-beat alignment.

Table 3 uses the separate full-memory query diagnostic, with teacher-forced recorded history and independently encoded unit texts. For speaker s with U_s units, K_s centroids, and cluster sizes n_{sc}, its analytical reference levels are $K_s/\sum_r K_r$ for SP, $\sum_c (n_{sc}/U_s)^2$ for CC, and $1/U_s$ for CG. SP corresponds to a uniformly sampled pooled centroid; CC and CG correspond to uniform unit selection. The reported values average the per-speaker levels equally. They are not simulated free-running random gestures. The checkpoint and ten-minute descriptor context are held fixed across the three reported query metrics.

D.3 Resource measurement boundaries

The canonical input is test take `12_zhao_0_1_1`: 69 seconds, 2,070 frames at 30 fps, and 188 transcript words. ConGRets forms 38 text chunks. The native assembled output is about 88.3 seconds for M and 63.7 seconds for H/D under the respective buffer/overlap rules. All system timings use the canonical input, but their native output schedules differ. Resource measurements time text encoding, sequential prediction, retrieval/decoding, and blending through the normal generation path. They use the full enrolled memory and exclude the quality protocol's take-specific blocklist and post-generation length truncation.

Table 18: Resource measurement boundaries

ConGRets mode	Generation parameters (M)	GPU warm ms, mean ± SD	CPU warm ms, mean ± SD	Peak allocated GPU memory (GB)	CPU peak RAM (GB)
Memory-based	11.438	19.25 ± 1.23	22.82 ± 0.33	0.0572	0.989
Hybrid	11.989	18.44 ± 0.12	30.46 ± 0.16	0.0697	0.935
Direct	11.989	14.59 ± 0.54	26.73 ± 0.53	0.0698	0.934

The main GPU figures use five measured repeats and the CPU figures three, with population standard deviation over repeats, not confidence intervals over utterances. The text encoder contributes approximately 11.171 M parameters, the fusion head 0.268 M, and the decoder 0.551 M. Motion and speaker encoders operate outside the warm generation path; the speaker encoder alone contributes

approximately 3.924 M additional parameters when enrollment is performed. The canonical speaker memory has 1,645 units, with recorded motion occupying about 75.3 MiB and latents/centroids about 0.07 MiB. Generation-time model size is not the total training-system footprint or motion-memory storage.

Hardware is an NVIDIA RTX A6000 and an AMD EPYC 9124 CPU with 16 inference threads. The September 18 ConGRets records use GPU cards of the same model as the baseline measurements, with different physical cards and measurement times. The machine was shared, and system load varied. Baseline warm measurements use two or three repeats, depending on method. Hence ratios describe the recorded runs and input, not a statistical speed guarantee over hardware or utterances.

Peak memory is whole-process peak over cold and warm phases. Cold timing includes loading and the first generation call; it is process-cold, not necessarily filesystem-cache-cold. Representative main M process-cold totals are approximately 5.59 seconds on GPU and 5.79 seconds on CPU, substantially longer than warm generation. Reference acquisition, motion encoding, speaker enrollment, and database construction remain separate preparation costs. The RAG-Gesture annotation pipeline is outside its recorded timing; this omission is disclosed rather than filled with an estimate.

D.4 Reproduction procedure

1. Use the recorded BEAT partitions and native data preparations for each system. Build the root-centered 50-joint exports at 30 fps, preserving per-take metadata and actual output lengths.
2. Train or load the motion encoder, speaker encoder, and separately fitted decoder described in Appendix A. Construct motion units, their latents, and per-speaker descriptors with the matching encoder weights.
3. Train each predictor recipe with its specified seed and select the saved checkpoint using validation in-batch retrieval accuracy. Ensure decoder, latent space, and motion memory belong to that run's declared family.
4. Generate M/H/D outputs with the declared feedback rules. Exclude the scored take's units for generation selection and retain full memory only for the separate retrieval task. Preserve divergence and decoder-selection diagnostics.
5. Use the pinned position evaluator for FGD and the specified pose-diversity implementation. Report the common 138-take baseline panel separately from full-support ablations. Compare recorded, static, and random-unit controls in the same scoring space.
6. Measure each model in its compatible environment on the canonical real input. Record weights, input/output lengths, warm-up and repeat counts, synchronization, resident parameter count, cold/warm time, whole-process memory, device, thread count, and machine load.

The experiment records and implementation define these steps. Reproducing numerical scores additionally requires the corresponding datasets, checkpoints, and motion-memory artifacts. Their anonymous distribution and third-party licensing must be checked before a submission package can claim artifact availability.

E Development studies

These measurements belong to the earlier development phase. They are not substituted for final checkpoint-aligned metrics. Module objectives, supports, and protocols differ; numerical scales must not be compared across different metrics. The tables below transcribe recorded outcomes and do not assert that a chosen configuration wins every later evaluation.

E.1 Motion representation, reconstruction, and speaker modules

Table 19: Motion representation, reconstruction, and speaker modules

Recorded module run	Pair accuracy	Reconstruction / validation MSE	Validation loss	Epochs
clr/e11_decoder_src	0.828	0.00911418	3.17574	45
clr/e1_w128	0.952	—	1.03506	57
clr/e1_w16	0.907	—	2.59346	26
clr/e1_w32	0.954	—	1.65539	44
clr/e1_w4	0.4495	—	4.65943	31
clr/e1_w64	0.948	—	1.32824	50
clr/e1_w8	0.7535	—	3.59516	17
clr/e2_contrastive_recon_w10	0.8205	—	3.24109	49
clr/e2_contrastive_recon_w128	0.949	—	1.20984	35
clr/e2_contrastive_recon_w32	0.956	—	1.65456	48
clr/e2_vae_w10	—	—	0.00555778	45
clr/e2_vae_w128	—	—	0.00527124	47
clr/e2_vae_w32	—	—	0.00522503	63
clr/main/s0	0.816	—	3.32119	16
clr/main/s1	0.802	—	3.24187	29
clr/main/s2	0.8235	—	3.21718	29
clr/pairmode_2d3d	0.786	—	3.13193	38
decoder/e2d	—	0.00454989	0.00482031	36
m3/decoder_w10	—	0.00474085	0.00497384	73
m3/decoder_w128	—	0.00395936	0.00430163	71
m3/decoder_w16	—	0.00409989	0.004336	52
m3/decoder_w32	—	0.00391751	0.00420476	42
m3/decoder_w4	—	0.00981751	0.0101328	55
m3/decoder_w64	—	0.00416689	0.0045011	99
m3/decoder_w8	—	0.00614534	0.00643141	40
style/raw/s0	1	—	0.0937042	500
style/raw/s1	1	—	0.0614942	500
style/raw/s2	1	—	0.131347	500
style/stem/s0	1	—	0.0864455	500
style/stem/s1	1	—	0.405561	500
style/stem/s2	1	—	0.796118	500

Reported module metrics can be extrema reached at different epochs. A dash means that metric was not stored under this key, not a zero result. E1 encoder training ended at different epochs under early stopping, so capacity and training duration are not fully isolated. M3 compares separately fitted unit decoders at all seven widths. The E2b/E2c runs compare contrastive-plus-reconstruction and VAE objectives at widths 10, 32, and 128. VQ-VAE was not an executed comparison.

E.2 Historical downstream selection

Table 20: Historical downstream selection

Historical predictor run	Maximum validation retrieval	Epoch of maximum	Epochs trained
albert_base	0.0159626	19	22
albert_base_frozen	0.0174093	2	11
bert_128	0.0174576	9	11
bert_base	0.016059	16	22

Continued on next page

Table 20 (continued)

Historical predictor run	Maximum validation retrieval	Epoch of maximum	Epochs trained
bert_base_frozen	0.0184221	8	13
bert_mini/s0	0.0180363	1	11
bert_mini/s1	0.0202064	9	11
bert_mini/s2	0.018615	1	11
bert_mini_frozen	0.01712	11	13
bert_small	0.017988	7	11
bert_tiny	0.0183256	1	12
bge_small	0.0157697	13	14
bge_small_frozen	0.0178916	8	15
distilgpt2	0.0191937	2	11
e11_mse	0.0162037	12	13
e11_nt_xent_mse	0.0177951	10	12
e11_nt_xent_mse/s0	0.0178916	2	12
e11_nt_xent_mse/s1	0.0180363	0	11
e11_nt_xent_mse/s2	0.0184221	0	12
e11_nt_xent_mse_ft	0.0176505	7	12
e11_nt_xent_mse_ft/s0	0.0177951	9	12
e11_nt_xent_mse_ft/s1	0.0175058	3	12
e11_nt_xent_mse_ft/s2	0.0192901	1	11
e11_nt_xent_mse_jointdec	0.0171682	5	11
e11_nt_xent_mse_jointdec/s0	0.0171682	5	11
e11_nt_xent_mse_jointdec/s1	0.0184221	3	12
e11_nt_xent_mse_jointdec/s2	0.0177469	0	11
e11_nt_xent_perceptual	0.0183738	1	12
e11_nt_xent_perceptual/s0	0.017988	11	12
e11_nt_xent_perceptual/s1	0.0188079	1	11
e11_nt_xent_perceptual/s2	0.0170235	3	11
e11_nt_xent_perceptual_ft	0.0171682	3	12
e11_nt_xent_perceptual_ft/s0	0.0181809	3	12
e11_nt_xent_perceptual_ft/s1	0.0188561	3	12
e11_nt_xent_perceptual_ft/s2	0.0172164	1	11
e11_perceptual	0.015191	11	13
electra_small	0.0151427	8	14
minilm_l6	0.0182774	5	12
minilm_l6_frozen	0.0180845	8	14
w32/fusion_finetuned/s0	0.0163966	10	11
w32/fusion_finetuned/s1	0.0168789	9	11
w32/fusion_finetuned/s2	0.0161555	3	11
w32/fusion_frozen/s0	0.0137442	10	11
w32/fusion_frozen/s1	0.0132137	9	12
w32/fusion_frozen/s2	0.0140818	7	11

Except for rows beginning with `w32/`, run names omit the common prefix `fusion/`; prepending it recovers the exact run identifier. The earlier runs retained checkpoints by validation loss while reporting peak retrieval accuracy. Some reported maxima therefore correspond to unsaved weights. The final evaluation uses the later retrieval-selected checkpoints in Appendix C. This table records design history and must not be paired with final FGD as if both described the same weights. Unsuffixed development trials and later seeded runs are preserved as separate recorded jobs, not asserted to be independent recipes.

E.3 SPEAKER NORMALIZATION AND LOSS CONTROLS

Nine variants were trained for 500 steps with batch 8, seed 342, learning rate 0.0005, weight decay 0.01, and temperature 0.1. Each input contains three segments of 150 frames. These micro-scale diagnostics use 12 training and four validation speakers and predate the overlap-masked protocol below. `dump` denotes the recovered earlier encoder/loss; `repo` the then-current implementation; std and bn denote input standardization and output BatchNorm.

Table 21: Speaker normalization and loss controls

Variant	Train loss	Train pair accuracy	Validation NN	Off-diagonal cosine
dump-enc / dump-loss	2.09139	0.125	0.208333	1
dump-enc / repo-loss	2.71693	0.125	0.208333	1
repo-enc / dump-loss	0.000707984	1	0.958333	0.225189
repo-enc / repo-loss	0.000686538	1	0.958333	0.269397
repo-raw / repo-loss	1.27433	0.25	0.666667	0.739891
dump+std / repo-loss	0.00798935	1	0.791667	0.0467347
dump+bn / repo-loss	0.974684	0.875	0.833333	0.778805
dump+both / repo-loss	0.0035913	1	0.875	0.240309
dump+both / dump-loss	0.00203836	1	0.833333	-0.0610365

E.4 SPEAKER TEMPORAL-STRIDE AND LEARNING-RATE SCREEN

All 12 recorded runs are shown. These use the earlier four-speaker evaluation (23 disjoint chunks); the static reference has NN 0.869565 and margin 0.052687. They must not be pooled with the five-speaker masked protocol in E.5. The raw arm has no convolutional stride. Frame cosine compares the descriptor before/after frame permutation; values near one indicate order insensitivity.

Table 22: Speaker temporal-stride and learning-rate screen

Strides	Learning rate	Epochs	Margin	NN	Frame cosine
2/2/2	0.001	26	0.582494	0.869565	0.973206
2/2/2	0.0005	19	0.331693	0.782609	0.979183
2/2/2	0.0002	19	0.464746	0.869565	0.99757
2/2	0.001	30	0.405275	0.869565	0.982222
2/2	0.0005	19	0.733005	0.782609	0.995453
2/2	0.0002	22	0.60978	0.956522	0.997926
2	0.001	14	0.2304	0.652174	0.999302
2	0.0005	14	0.315065	0.956522	0.997902
2	0.0002	25	0.27465	0.913043	0.994602
raw	0.001	26	0.376594	0.956522	0.999995
raw	0.0005	22	0.419421	0.956522	0.999989
raw	0.0002	15	0.434909	1	0.999975

E.5 SPEAKER STRIDE × NORMALIZATION × SEED STUDY

All 24 runs use five validation speakers and 200 reference chunks with overlap masking: candidate pairs sharing segments are excluded. The static reference has NN 0.835 and margin 0.027644. std uses training-set standardization; pcs adds per-chunk standardization. Margins compare within-speaker and between-speaker descriptor similarity. This is a speaker-representation diagnostic, not synthesized-gesture quality.

Table 23: Speaker stride × normalization × seed study

Strides	Normalization	Seed	Margin	NN	Frame cosine
2/2/2	std	0	0.0738603	0.475	0.999817
2/2/2	std	1	0.18298	0.475	0.994138
2/2/2	std	2	0.469925	0.68	0.997167
2/2	std	0	0.48987	0.73	0.995979
2/2	std	1	0.684116	0.8	0.999603
2/2	std	2	0.166724	0.875	0.998397
2	std	0	0.217761	0.815	0.998373
2	std	1	0.501195	0.82	0.997287
2	std	2	0.332711	0.835	0.995179
raw	std	0	0.457128	0.915	0.99999
raw	std	1	0.574931	0.86	0.99998
raw	std	2	0.315903	0.91	0.999977
2/2/2	pcs	0	0.187007	0.42	0.799202
2/2/2	pcs	1	0.139841	0.345	0.747738
2/2/2	pcs	2	0.145097	0.25	0.844674
2/2	pcs	0	0.0552809	0.31	0.902197
2/2	pcs	1	0.200322	0.31	0.930617
2/2	pcs	2	0.0258911	0.605	0.983835
2	pcs	0	0.0922327	0.34	0.99668
2	pcs	1	0.10672	0.58	0.963684
2	pcs	2	0.200657	0.45	0.621445
raw	pcs	0	0.107868	0.69	0.996801
raw	pcs	1	0.309076	0.78	0.997363
raw	pcs	2	0.120237	0.625	0.999271

E.6 Matched retraining of the earlier speaker architecture

The earlier 29.06 M-parameter v2 architecture was retrained in three seeds under the masked protocol. This check separates an earlier favorable checkpoint from the architecture comparison; it is not an additional final predictor family.

Table 24: Matched retraining of the earlier speaker architecture

Seed	Epochs	Margin	NN	Frame cosine
0	20	0.23917	0.88	1
1	13	0.123956	0.925	1
2	17	0.56496	0.735	1

E.7 Production-scale speaker and decoder checks

The S3 speaker comparison reports raw versus convolutional-stem encoders over three seeds each. These are module results, not three independent synthesis-mode style evaluations.

Table 25: Production-scale speaker and decoder checks

Encoder	Parameters M	Margin, mean ± SD	Masked LOO-NN, mean ± SD	Frame cosine
Raw	3.92	0.629 ± 0.108	0.973 ± 0.015	0.99998
Stem 2/2/2	4.61	0.646 ± 0.070	0.743 ± 0.186	0.94658
Static reference	0	0.028	0.835	Invariant

The separately recorded decoder comparison uses the same real E2b latents over 27 speakers and 76,998 units: the window-trained joint decoder has masked MSE 0.00721 and the separately fitted unit decoder 0.00466. This is an all-corpus reconstruction diagnostic, not a held-out generalization score. The earlier micro comparison of clean versus augmented reconstruction targets reports joint-decoder MSE 0.01913 versus 0.01957; it does not explain most of the joint-versus-unit decoder difference. The main decoder job has no separately verified decoder validation value in the tracked automatic ledger; the M3 width-10 decoder result is not silently assigned to that artifact.

E.8 Motion-selection development

A micro-scale study swept target units per cluster across 17 speakers, then compared small-cluster merge floors. The selected setting used approximately 12.5 units per cluster and merge floor 5. This is a study of selection structure, not semantic relevance or an established optimum after full representation training.

Table 26: Motion-selection development

Merge floor	Mean minimum cluster size	Mean maximum cluster size	Clusters in size band 10–20	Units displaced
2	2.1	20.8	71%	0.07%
5	5.1	21.8	75%	1.32%
10	10.0	27.2	86%	11.04%

The broader proposed encoder-architecture, fusion-conditioning, clustering-algorithm, buffer/overlap, automated-HPO, and downstream-style sweeps were not completed in the evidence used here. They are not counted as experimental support.

F External-corpus transfer with target-speaker enrollment

F.1 Question, packs, and reference access

We evaluate frozen BEAT-trained weights on Trinity, Talking With Hands (TWH), Audio2Photoreal, and Embody3D. The packs contain 15 speakers, 47 enrollment takes, and 76 scoring takes, comprising 360.1 scoring minutes. The 123 packed takes include both splits. This is transfer with target-speaker enrollment for methods that require it: ConGRets receives target motion for its descriptor and retrieval memory, ZS-MSTM receives a local motion-reference clip from the enrollment split, and the audio-conditioned baselines receive the target audio. The experiment does not establish equally provisioned, target-data-free generation. No weight, motion encoder, speaker encoder, decoder, or input-normalization statistic is fitted to a target pack.

Eligible takes have at least 20 transcript words and at least 40 words per minute. Per speaker, whole takes in recording order enter enrollment until at least ten minutes are accumulated. Subsequent eligible whole takes enter scoring, skipping any take that would exceed the 30-minute cap and stopping after at least 20 minutes. Selection is deterministic; the recorded selection seed 0 introduces no random selection. Enrollment and scoring occupy separate processed roots, with disjoint take identifiers. Every scoring take is generated in all ten rows: four baseline systems, three ConGRets arms, and three controls. Each system therefore has the same take identities within a pack, although output durations differ.

Trinity has one actor: release I supplies enrollment and audio-bearing release II takes supply scoring; release-II take `tsg2:2-6` has no audio and is excluded. TWH selects the five eligible participants with the most worded minutes, requiring at least 50 worded minutes and finger capture; the selected participants have enough fingered material to use only fingered takes. Audio2Photoreal includes all four persons in scene order. Embody3D selects the first five eligible participants in the ranking by worded minutes from speech sections only, with at most 5% missing mocap. Its higher-ranked CKG723 has only 5.0 eligible minutes after the word-density filter; AQK592 has 25.7, below the enrollment-plus-scoring requirement. Both are excluded by these eligibility rules.

Two takes were excluded after an audio/word-alignment probe exposed upstream transcript defects: TWH deep4 session16/take12, whose optimum shift was +2.40 s, and an Embody3D ATM352 AIAGENT capture, whose optimum was -4.70 s and whose word frames were quieter than silence frames at zero shift. The exclusions and source identifiers are retained in the pack recipe and manifests. ASR word boundaries in retained takes can still lag speech energy by a few tenths of a second. The corpora use Whisper large-v3 transcripts, with faster-whisper large-v3-turbo for Embody3D.

All source motions use the same canonical 52-joint omnimo rest skeleton at 30 fps; the audit finds the same bone lengths as BEAT speaker 12 to floating-point precision. Evaluation uses the root-centered 50-joint comparison positions. This removes a rest-skeleton scale difference, but does not remove domain shifts in behavior, posture, recording, or conversion. Trinity release II has no world-root translation. Embody3D contains seated and lying motion; speakers 403 and 404 spend much of the scoring material seated. These differences are retained.

Audio provenance. Every pack WAV is 16 kHz mono PCM16, start-aligned and trimmed or padded to the motion duration. Trinity uses 44.1 kHz session audio with the negative of the recorded clapper offset; TWH uses each participant's 44.1 kHz close microphone, retaining partner bleed and source privacy-masked zero spans. Both use polyphase resampling by $160/441$. Audio2Photoreal uses channel 0 (the current person, not the partner) of the 48 kHz stereo audio; Embody3D uses its 48 kHz per-participant source-separated track. The latter two use resampling by $1/3$ from 48 kHz. Source-separated audio is an additional qualification for audio-conditioned systems. In dyadic and polyadic recordings, each transcript covers only that person's turns, while the motion also includes listening.

Frozen ConGRets artifacts and enrollment. The predictor, motion encoder, speaker encoder, and decoder are the designated frozen artifacts listed in Appendix A.6; the predictor checkpoint was selected by validation retrieval accuracy. Style standardization uses the BEAT-training statistics stored in its checkpoint. The target enrollment train/validation lists are empty. Enrollment uses seed 42 for clustering and reference-window draws to match the BEAT artifact-building recipe; the predictor configuration retains training seed 43. Generation draws use seed 0 for every system and control.

Enrollment produces 2,823 two-second word-aligned units and 209 clusters across the 15 speakers. Libraries contain 117–275 units; the smallest retained cluster has five units. The descriptor uses row 0 of four sampled contiguous ten-minute contexts from the concatenated enrollment stream. Seven speakers have only one available ten-minute context, so all four stored vectors are identical for those speakers. Five-minute contexts are also stored; 20- and 30-minute contexts are unavailable under the approximately 10–15-minute external enrollment budgets and were not requested. Enrollment and scoring roots are disjoint, so all 228 ConGRets sidecars record zero units excluded by the take blocklist. This external protocol differs from the within-corpus enrollment overlap described in Appendix B; the inherited caveat that the BEAT memory contains test takes does not apply to these external rows.

Table 27: Pack speaker mapping and split support. Takes are enrollment/scoring; durations are minutes. Windows are the protocol's reference feature-support estimate; exact scored feature counts appear in the metric-support tables. A2P denotes Audio2Photoreal.

ID	Source	Takes	Enroll min	Score min	Score frames	Windows
101	Trinity I/II	1 / 2	10.22	22.50	40491	2530
201	TWH deep5	1 / 2	10.18	24.31	43750	2734
202	TWH deep4	1 / 2	12.32	29.25	52649	3290
203	TWH shallow28	1 / 3	12.08	22.50	40492	2530
204	TWH shallow18	1 / 3	11.73	28.66	51582	3223
205	TWH shallow19	1 / 2	10.95	23.58	42443	2652
301	A2P GQS883	3 / 5	14.95	24.86	44742	2796
302	A2P PXB184	2 / 5	10.02	24.87	44771	2798

Continued on next page

Table 27 (continued)

ID	Source	Takes	Enroll min	Score min	Score frames	Windows
303	A2P RLW104	3 / 5	14.51	24.94	44891	2805
304	A2P TXB805	2 / 5	10.00	24.59	44257	2766
401	Embody3D JON169	7 / 9	10.58	20.08	36141	2258
402	Embody3D ATM352	8 / 6	10.25	21.98	39566	2472
403	Embody3D IKF796	3 / 8	14.99	21.23	38220	2388
404	Embody3D QZX685	4 / 6	14.92	22.84	41111	2569
405	Embody3D SZM479	9 / 13	11.50	23.92	43051	2690

F.2 Implementation and metric protocol

The comparison retains EMAGE, ZS-MSTM, HoloGest, and DiffuseStyleGesture+ (DSG+), plus ConGRets memory-based (M), hybrid (H), and direct (D) synthesis. A1 assembles reference word-aligned motion units; it is not the untouched raw reference used by FGD. A3 holds the scoring take's mean pose still. A4 randomly draws target-speaker units from the enrollment library. These are controls with privileged reference access, not competing trained systems. A4 is not a certified floor or mathematical lower bound.

ConvoFusion, RAG-Gesture, and the Habibie reimplementation are outside this frozen transfer scope. GestureLSM is excluded because its evaluated implementation seeds from BEAT2 ground-truth motion unavailable in the external packs. Thus a rank refers only to the seven evaluated system arms. All 760 exports (76 takes times ten rows) were produced successfully. No baseline source code was changed; the following driver and scoring overrides are recorded rather than treated as unchanged inference assumptions:

D1: reference root. ZS-MSTM obtains its first valid 128-frame reference window from another take under the enrollment root, rather than accessing scoring takes.

D2: initialization sample. When the canonical BEAT sample cannot be resolved, driver initialization binds to the first planned external take. Sidecars record the resolved roots.

D3: transcript format. DSG+ falls back from a missing TextGrid to the pack's timestamped metadata words. The existing TextGrid path remains in use when present.

D4: unseen speaker IDs. HoloGest extends its existing unnamed-speaker rule to external IDs: a deterministic unit vector from a random generator seeded by speaker ID, rather than a newly learned speaker embedding.

D5: SMPL-X shape. EMAGE and HoloGest conversion use fixed BEAT speaker-12 shape parameters, consistent with the canonical rest skeleton used to construct the packs. This is a declared conversion choice, not target-body shape estimation.

D6: random-library pool. A4 receives the enrollment root as its explicit unit pool, matching ConGRets' target-motion access.

E1: WAV loader location. The scoring launcher points its resource-loader root at the staged copy containing the WAV reader; the metric implementation is unchanged.

E2: speaker gallery. Spk-Sim's candidate-speaker parameter is set to the IDs of the current pack. It is part of the saved metric parameters and cache key. No Spk-Sim is computed for the one-speaker Trinity pack.

FGD, GT-Sim, and Spk-Sim use the same frozen BEAT-trained positional feature extractor as the main evaluation: epoch 117, validation MPJPE 0.005241 m, SHA-256 as listed in Appendix A.6. Scoring is on CPU at 30 fps with root-centered 50-joint positions. These are out-of-domain feature measurements. FGD values should be compared within a dataset; neither raw FGD nor normalization by a local control validates cross-dataset comparability or perceptual validity.

FGD independently crops each generated and raw-reference sequence to a multiple of 32 frames, encodes it, and pools one feature per 16 frames; it does not equalize their duration. The raw whole scoring takes, not A1, supply the reference population. Div-Pose samples up to 1,000 frame pairs

within each output. GT-Sim maps each text chunk to an output interval using that system's timeline or assembly rule, clips to available frames, and leaves intervals shorter than eight frames unscored. Spk-Sim compares pooled whole-output embeddings to centroids of whole raw takes. Its identification chance is $1/5$ for TWH and Embody3D, $1/4$ for Audio2Photoreal; it is undefined for Trinity.

Coverage is the arithmetic mean across takes of output/reference frame ratio, with the minimum and maximum also reported. It is not speech-alignment accuracy. ConGRets exports truncate overlong streams to the reference length, but do not pad or stretch short streams. A1/A3/A4 also follow assembled-unit schedules and can be shorter than the raw take; small ratios above one reflect their boundary/window arithmetic. Feature counts quantify numerical support, not independent samples or statistical confidence. Frame coverage and GT-Sim chunk support differ: on Trinity, ConGRets has full written-frame coverage but many later text chunks lie beyond the truncated output.

F.3 Complete per-dataset measurements

The tables below retain every evaluated row and control. FGD and Div-Pose are rounded to four decimals, GT-Sim and Spk-Sim to four decimals, and coverage to three decimals; unrounded values remain in the underlying CSVs and append-only ledgers. A3's tiny numerical Div-Pose residuals are printed in scientific notation rather than presented as measured zeros. Takes denotes FGD scoring takes. Each primary table is immediately followed by feature, chunk, and coverage support plus the two eligible temporal diagnostics. BC is beat consistency; jerk is the saved jerk statistic. Neither is a ranking of perceived quality: short text-driven streams can be compared to an audio prefix rather than aligned speech intervals, and static or oversmoothed motion can appear favorable under smoothness statistics. A4's chunk-to-frame mapping can drift by ten frames per recording-boundary unit because individual crossfade widths were not stored (at most 20 frames observed on Trinity).

Table 28: Trinity: full metric panel with FGD take support and mean frame coverage. M/H/D are defined in the protocol; controls are excluded from system ranks.

System/control	FGD ↓	Div-Pose	GT-Sim	Takes	Coverage
A1 assembled	0.0167	0.1010	0.8222	2	1.002
A3 static	1.1583	6.21×10^{-10}	0.0010	2	1.002
A4 random	7.9513	0.1264	-0.0002	2	1.001
EMAGE	2.8986	0.0267	-0.0381	2	1.000
ZS-MSTM	1.9462	0.0070	0.0073	2	0.968
HoloGest	1.9352	0.1467	0.0181	2	1.000
DSG+	1.0374	0.0678	-0.0142	2	1.000
ConGRets M	7.1396	0.1405	0.0200	2	1.000
ConGRets H	5.4969	0.1181	0.0266	2	1.000
ConGRets D	0.7745	0.0997	0.0871	2	1.000

Table 29: Trinity: diagnostics and support. FGD features are generated/reference; GT-Sim chunks are scored/total; coverage range is the per-take minimum–maximum. Dashes mark ineligible A3 BC/jerk, not zeros.

System/control	BC	Jerk	FGD features	GT-Sim chunks	Coverage range
A1 assembled	0.9087	74.50	2534/2530	1054/1066	1.002–1.002
A3 static	—	—	2534/2530	1060/1066	1.002–1.002
A4 random	0.9097	118.88	2532/2530	1052/1066	1.001–1.002
EMAGE	0.9163	19.89	2530/2530	1060/1066	1.000–1.000
ZS-MSTM	0.9244	50.07	2448/2530	1045/1066	0.967–0.969
HoloGest	0.9165	15.93	2528/2530	1060/1066	0.999–1.000
DSG+	0.9158	45.08	2530/2530	1060/1066	1.000–1.000

Continued on next page

Table 29 (continued)

System/control	BC	Jerk	FGD features	GT-Sim chunks	Coverage range
ConGRets M	0.9074	108.89	2530/2530	576/1066	1.000–1.000
ConGRets H	0.9077	57.62	2530/2530	805/1066	1.000–1.000
ConGRets D	0.9104	26.48	2530/2530	805/1066	1.000–1.000

Table 30: TWH: full metric panel with FGD take support and mean frame coverage. M/H/D are defined in the protocol; controls are excluded from system ranks.

System/control	FGD ↓	Div-Pose	GT-Sim	Spk-Sim	Takes	Coverage
A1 assembled	0.0559	0.0985	0.9200	1.0000	12	0.687
A3 static	0.8001	1.95×10^{-9}	-0.0114	1.0000	12	0.687
A4 random	1.1338	0.0663	0.0220	0.7500	12	0.687
EMAGE	3.6536	0.0315	0.0168	0.2500	12	1.000
ZS-MSTM	4.3528	0.0068	0.0140	0.1667	12	0.966
HoloGest	3.1902	0.1120	0.0076	0.2500	12	1.000
DSG+	3.5731	0.0446	0.0008	0.1667	12	1.000
ConGRets M	1.3037	0.0643	0.0327	0.5833	12	0.837
ConGRets H	1.5460	0.0507	0.0324	0.5000	12	0.639
ConGRets D	3.7937	0.1077	0.0382	0.1667	12	0.639

Table 31: TWH: diagnostics and support. FGD features are generated/reference; GT-Sim chunks are scored/total; coverage range is the per-take minimum–maximum. Dashes mark ineligible A3 BC/jerk, not zeros.

System/control	BC	Jerk	FGD features	GT-Sim chunks	Coverage range
A1 assembled	0.7782	136.95	9904/14422	2819/2885	0.381–0.994
A3 static	—	—	9904/14422	1936/2885	0.381–0.994
A4 random	0.7810	141.12	9904/14422	2819/2885	0.381–0.994
EMAGE	0.7962	16.38	14422/14422	2830/2885	1.000–1.000
ZS-MSTM	0.7927	43.89	13944/14422	2711/2885	0.962–0.969
HoloGest	0.7859	9.81	14420/14422	2830/2885	1.000–1.000
DSG+	0.7929	58.05	14422/14422	2830/2885	1.000–1.000
ConGRets M	0.7912	139.20	11938/14422	2680/2885	0.480–1.000
ConGRets H	0.7805	32.84	8988/14422	2822/2885	0.343–1.000
ConGRets D	0.7790	28.06	8988/14422	2822/2885	0.343–1.000

Table 32: Audio2Photoreal: full metric panel with FGD take support and mean frame coverage. M/H/D are defined in the protocol; controls are excluded from system ranks.

System/control	FGD ↓	Div-Pose	GT-Sim	Spk-Sim	Takes	Coverage
A1 assembled	0.0257	0.1062	0.8964	0.8500	20	0.669
A3 static	0.7517	1.14×10^{-9}	0.0112	0.7500	20	0.669
A4 random	0.2456	0.0985	0.0101	0.9000	20	0.669
EMAGE	3.3102	0.0496	-0.0165	0.2500	20	1.000
ZS-MSTM	3.0114	0.0104	0.0216	0.2500	20	0.963
HoloGest	2.3030	0.0914	0.0278	0.1000	20	1.000
DSG+	1.9366	0.0833	-0.0267	0.2500	20	1.000
ConGRets M	0.3429	0.1041	0.0170	0.8500	20	0.834
ConGRets H	0.9200	0.0745	0.0198	0.5000	20	0.611
ConGRets D	3.3824	0.1071	0.0200	0.2500	20	0.611

Table 33: Audio2Photoreal: diagnostics and support. FGD features are generated/reference; GT-Sim chunks are scored/total; coverage range is the per-take minimum–maximum. Dashes mark ineligible A3 BC/jerk, not zeros.

System/control	BC	Jerk	FGD features	GT-Sim chunks	Coverage range
A1 assembled	0.6812	127.29	7446/11146	2119/2178	0.534–0.856
A3 static	—	—	7446/11146	1427/2178	0.534–0.856
A4 random	0.6779	151.50	7444/11146	2119/2178	0.534–0.856
EMAGE	0.6969	16.96	11146/11146	2138/2178	1.000–1.000
ZS-MSTM	0.6775	42.57	10736/11146	2057/2178	0.956–0.969
HoloGest	0.6557	7.51	11144/11146	2135/2178	0.999–1.000
DSG+	0.6637	44.84	11146/11146	2138/2178	1.000–1.000
ConGRets M	0.6820	147.73	9292/11146	2094/2178	0.480–1.000
ConGRets H	0.6729	43.41	6800/11146	2138/2178	0.345–0.914
ConGRets D	0.6778	28.48	6800/11146	2138/2178	0.345–0.914

Table 34: Embody3D: full metric panel with FGD take support and mean frame coverage. M/H/D are defined in the protocol; controls are excluded from system ranks.

System/control	FGD ↓	Div-Pose	GT-Sim	Spk-Sim	Takes	Coverage
A1 assembled	0.0437	0.0995	0.9441	0.6429	42	0.522
A3 static	0.5209	1.20×10^{-9}	0.0088	0.4524	42	0.522
A4 random	1.2245	0.1995	0.0033	0.3810	42	0.522
EMAGE	4.6662	0.0299	0.0223	0.1429	42	1.000
ZS-MSTM	4.1094	0.0095	0.0048	0.2143	42	0.952
HoloGest	2.3418	0.0890	0.0083	0.1190	42	0.998
DSG+	3.8166	0.0610	-0.0102	0.1905	42	1.000
ConGRets M	1.4103	0.1452	0.0015	0.2619	42	0.626
ConGRets H	1.7400	0.0964	0.0036	0.2381	42	0.458
ConGRets D	3.5992	0.0920	0.0087	0.1429	42	0.458

Table 35: Embody3D: diagnostics and support. FGD features are generated/reference; GT-Sim chunks are scored/total; coverage range is the per-take minimum–maximum. Dashes mark ineligible A3 BC/jerk, not zeros.

System/control	BC	Jerk	FGD features	GT-Sim chunks	Coverage range
A1 assembled	0.6384	65.88	6562/12348	1764/1818	0.297–0.946
A3 static	—	—	6562/12348	1064/1818	0.297–0.946
A4 random	0.6295	203.40	6554/12348	1763/1818	0.297–0.945
EMAGE	0.6396	16.69	12348/12348	1780/1818	0.999–1.000
ZS-MSTM	0.5957	43.01	11826/12348	1725/1818	0.906–0.969
HoloGest	0.5603	8.69	12324/12348	1777/1818	0.987–1.000
DSG+	0.5907	62.60	12348/12348	1780/1818	0.999–1.000
ConGRets M	0.6159	93.07	7794/12348	1761/1818	0.362–1.000
ConGRets H	0.6261	40.74	5666/12348	1780/1818	0.263–0.837
ConGRets D	0.6255	31.25	5666/12348	1780/1818	0.263–0.837

Interpretation with controls. Among the seven evaluated system arms, M and H rank first and second on TWH, Audio2Photoreal, and Embody3D; D ranks first on Trinity, while M and H rank last and second-last there. No single ConGRets arm wins all four packs, and choosing an arm per test dataset is post hoc. A4 random retrieval obtains lower FGD than learned M retrieval on all three conversational packs. A3 static pose obtains lower FGD than every evaluated system on TWH and Embody3D. The near-zero Div-Pose of A3 makes clear why these FGD outcomes cannot be interpreted as overall motion quality or semantic-alignment superiority.

Trinity contains only two scoring takes from one actor. Its A4 FGD of 7.951, alongside M's 7.140 and H's 5.497, is consistent with an enrollment/scoring mismatch across releases, but does not isolate its causal contribution or remove the observed retrieval transfer failure. D avoids motion-library selection; its result does not independently demonstrate effective text conditioning. In the conversational packs, ConGRets has markedly shorter written outputs than the baselines, with H/D covering on average 0.639, 0.611, and 0.458 of the raw take. A matched-temporal-support sensitivity analysis was not performed.

GT-Sim values for generated motion are small relative to assembled A1. Without independent uncertainty estimates or external-domain validation, this does not support a reliable semantic ranking, nor does it prove the scores are noise. Spk-Sim likewise requires its controls: M scores 0.583/0.850/0.262 on TWH/Audio2Photoreal/Embody3D, compared with A3 at 1.000/0.750/0.452 and A4 at 0.750/0.900/0.381. Library identity and pose can account for identification; these measurements do not isolate use of the predictor's style input. On TWH, M remains below A1's 1.000; on Audio2Photoreal, both score 0.850.

F.4 Per-speaker measurements

The following matrices transcribe every primary per-speaker GT-Sim and Spk-Sim value in the saved per-speaker result files, including controls and unfavorable values. Each cell gives the score, followed by its scored chunk count (GT-Sim) or take count (Spk-Sim) in parentheses. These are descriptive breakdowns, not independent repeated trials. FGD has no per-speaker breakdown in this record: it is computed over each pack's pooled features. No Trinity Spk-Sim matrix is provided because identification is undefined with one candidate speaker.

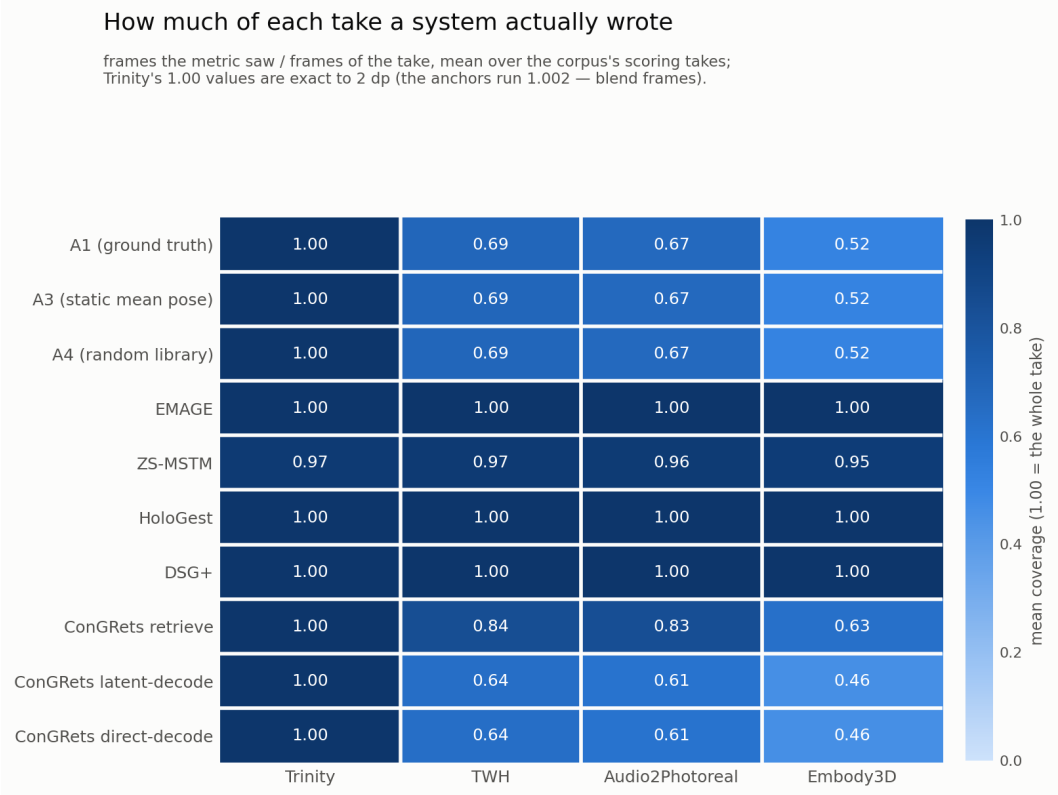


Figure 8: Mean written-frame coverage for every external system arm and control, from the saved per-take export/align records. This is an output/reference length ratio, not temporal or speech alignment accuracy. The corresponding minima, maxima, feature counts, and chunk support are reported beside each dataset's metric panel.

Table 36: Trinity: per-speaker GT-Sim. Parentheses give scored chunks.

System/control	101
A1 assembled	0.8222 (1054)
A3 static	0.0010 (1060)
A4 random	-0.0002 (1052)
EMAGE	-0.0381 (1060)
ZS-MSTM	0.0073 (1045)
HoloGest	0.0181 (1060)
DSG+	-0.0142 (1060)
ConGRets M	0.0200 (576)
ConGRets H	0.0266 (805)
ConGRets D	0.0871 (805)

Table 37: TWH: per-speaker GT-Sim. Parentheses give scored chunks.

System/control	201	202	203	204	205
A1 assembled	0.8307	0.9009	0.9237	0.9694	0.9466
	(403)	(610)	(592)	(581)	(633)
A3 static	0.0082	-0.0041	-0.0004	-0.0238	-0.0302
	(220)	(440)	(469)	(380)	(427)
A4 random	-0.0323	0.0537	0.0406	-0.0036	0.0322
	(403)	(610)	(592)	(581)	(633)
EMAGE	0.0220	0.0156	0.0459	0.0003	0.0027
	(406)	(612)	(594)	(582)	(636)
ZS-MSTM	0.0529	0.0153	-0.0097	0.0009	0.0219
	(393)	(582)	(576)	(557)	(603)
HoloGest	-0.0323	0.0442	0.0222	-0.0135	0.0035
	(406)	(612)	(594)	(582)	(636)
DSG+	-0.0453	0.0249	-0.0010	0.0061	0.0038
	(406)	(612)	(594)	(582)	(636)
ConGRets M	0.0003	0.0116	0.0054	-0.0041	0.1402
	(406)	(612)	(515)	(577)	(570)
ConGRets H	-0.0093	0.0074	0.0126	0.0115	0.1202
	(406)	(612)	(586)	(582)	(636)
ConGRets D	0.0641	0.0266	0.0246	-0.0079	0.0875
	(406)	(612)	(586)	(582)	(636)

Table 38: TWH: per-speaker Spk-Sim. Parentheses give scored takes.

System/control	201	202	203	204	205
A1 assembled	1.0000	1.0000	1.0000	1.0000	1.0000
	(2)	(2)	(3)	(3)	(2)
A3 static	1.0000	1.0000	1.0000	1.0000	1.0000
	(2)	(2)	(3)	(3)	(2)
A4 random	1.0000	1.0000	1.0000	0.0000	1.0000
	(2)	(2)	(3)	(3)	(2)
EMAGE	0.0000	0.5000	0.6667	0.0000	0.0000
	(2)	(2)	(3)	(3)	(2)
ZS-MSTM	0.0000	1.0000	0.0000	0.0000	0.0000
	(2)	(2)	(3)	(3)	(2)
HoloGest	0.0000	0.5000	0.3333	0.0000	0.5000
	(2)	(2)	(3)	(3)	(2)
DSG+	0.0000	1.0000	0.0000	0.0000	0.0000
	(2)	(2)	(3)	(3)	(2)
ConGRets M	1.0000	1.0000	0.3333	0.0000	1.0000
	(2)	(2)	(3)	(3)	(2)
ConGRets H	0.0000	1.0000	0.6667	0.0000	1.0000
	(2)	(2)	(3)	(3)	(2)
ConGRets D	0.5000	0.5000	0.0000	0.0000	0.0000
	(2)	(2)	(3)	(3)	(2)

Table 39: Audio2Photoreal: per-speaker GT-Sim. Parentheses give scored chunks.

System/control	301	302	303	304
A1 assembled	0.8742	0.9449	0.8752	0.8839
	(419)	(572)	(495)	(633)
A3 static	0.0068	0.0328	0.0034	0.0019
	(261)	(375)	(321)	(470)
A4 random	-0.0052	0.0241	0.0208	-0.0008
	(419)	(572)	(495)	(633)
EMAGE	-0.0577	-0.0972	-0.0017	0.0722
	(419)	(578)	(504)	(637)
ZS-MSTM	-0.0063	0.0517	0.0434	-0.0043
	(407)	(554)	(483)	(613)
HoloGest	-0.0463	0.0409	0.0750	0.0277
	(419)	(578)	(501)	(637)
DSG+	0.0086	-0.0530	-0.0415	-0.0143
	(419)	(578)	(504)	(637)
ConGRets M	0.0207	0.0027	0.0297	0.0175
	(419)	(578)	(504)	(593)
ConGRets H	0.0207	0.0101	0.0064	0.0387
	(419)	(578)	(504)	(637)
ConGRets D	0.0556	-0.0002	0.0256	0.0105
	(419)	(578)	(504)	(637)

Table 40: Audio2Photoreal: per-speaker Spk-Sim. Parentheses give scored takes.

System/control	301	302	303	304
A1 assembled	1.0000	0.8000	0.8000	0.8000
	(5)	(5)	(5)	(5)
A3 static	1.0000	0.8000	0.4000	0.8000
	(5)	(5)	(5)	(5)
A4 random	1.0000	0.8000	0.8000	1.0000
	(5)	(5)	(5)	(5)
EMAGE	0.0000	0.0000	0.0000	1.0000
	(5)	(5)	(5)	(5)
ZS-MSTM	0.0000	0.0000	0.0000	1.0000
	(5)	(5)	(5)	(5)
HoloGest	0.0000	0.0000	0.0000	0.4000
	(5)	(5)	(5)	(5)
DSG+	0.0000	0.0000	0.0000	1.0000
	(5)	(5)	(5)	(5)
ConGRets M	0.6000	1.0000	0.8000	1.0000
	(5)	(5)	(5)	(5)
ConGRets H	0.0000	0.0000	1.0000	1.0000
	(5)	(5)	(5)	(5)
ConGRets D	0.0000	0.0000	0.0000	1.0000
	(5)	(5)	(5)	(5)

Table 41: Embody3D: per-speaker GT-Sim. Parentheses give scored chunks.

System/control	401	402	403	404	405
A1 assembled	0.9305	0.9386	0.9127	0.9726	0.9512
	(286)	(385)	(283)	(425)	(385)
A3 static	0.0219	0.0021	-0.0116	0.0063	0.0224
	(164)	(248)	(137)	(294)	(221)
A4 random	0.0166	0.0019	-0.0028	0.0042	-0.0016
	(286)	(384)	(283)	(425)	(385)
EMAGE	0.0089	0.0398	0.0134	0.0372	0.0053
	(292)	(387)	(285)	(426)	(390)
ZS-MSTM	0.0278	-0.0074	0.0097	-0.0038	0.0053
	(283)	(372)	(284)	(413)	(373)
HoloGest	0.0070	0.0482	-0.0237	0.0139	-0.0132
	(292)	(386)	(284)	(426)	(389)
DSG+	0.0010	-0.0640	0.0196	-0.0076	0.0099
	(292)	(387)	(285)	(426)	(390)
ConGRets M	0.0572	-0.0561	0.0477	-0.0378	0.0239
	(292)	(375)	(285)	(420)	(389)
ConGRets H	0.0580	-0.0702	0.0729	-0.0507	0.0448
	(292)	(387)	(285)	(426)	(390)
ConGRets D	0.0270	0.0030	0.0206	-0.0142	0.0168
	(292)	(387)	(285)	(426)	(390)

Table 42: Embody3D: per-speaker Spk-Sim. Parentheses give scored takes.

System/control	401	402	403	404	405
A1 assembled	0.3333	0.8333	0.8750	0.3333	0.7692
	(9)	(6)	(8)	(6)	(13)
A3 static	0.0000	0.0000	0.8750	0.1667	0.8462
	(9)	(6)	(8)	(6)	(13)
A4 random	0.6667	1.0000	0.3750	0.1667	0.0000
	(9)	(6)	(8)	(6)	(13)
EMAGE	0.0000	1.0000	0.0000	0.0000	0.0000
	(9)	(6)	(8)	(6)	(13)
ZS-MSTM	0.3333	1.0000	0.0000	0.0000	0.0000
	(9)	(6)	(8)	(6)	(13)
HoloGest	0.0000	0.1667	0.0000	0.6667	0.0000
	(9)	(6)	(8)	(6)	(13)
DSG+	0.4444	0.3333	0.0000	0.3333	0.0000
	(9)	(6)	(8)	(6)	(13)
ConGRets M	0.4444	0.1667	0.5000	0.3333	0.0000
	(9)	(6)	(8)	(6)	(13)
ConGRets H	0.5556	0.8333	0.0000	0.0000	0.0000
	(9)	(6)	(8)	(6)	(13)
ConGRets D	0.3333	0.5000	0.0000	0.0000	0.0000
	(9)	(6)	(8)	(6)	(13)

F.5 Rank and local-control context

Ranks refer to the same seven system arms on each pack; anchors are not ranked. The BEAT column reuses the existing seven-arm subset of the BEAT evaluation and is not a new experiment. Rank changes summarize this restricted roster and do not measure effect size or significance. The ratio table retains all three controls and shows FGD divided by the local A3 and A4 values. They are

exploratory context only: dividing by A3 does not validate cross-dataset comparability, and A4 is not a certified floor. Ratios to near-zero A1 and GT-Sim rank tables are not used.

Table 43: FGD ranks among seven evaluated system arms. A2P abbreviates Audio2Photoreal.

System	BEAT	Trinity	TWH	A2P	Embody3D
EMAGE	4	5	5	6	7
ZS-MSTM	7	4	7	5	6
HoloGest	6	3	3	4	3
DSG+	3	2	4	3	5
ConGRets M	1	7	1	1	1
ConGRets H	2	6	2	2	2
ConGRets D	5	1	6	7	4

Table 44: Exploratory FGD ratios to the same pack's local controls. Each cell lists the ratio to A3 first and the ratio to A4 second. Lower ratios do not establish perceptual superiority.

System/control	BEAT	Trinity	TWH	A2P	Embody3D
A1 assembled	0.006 / 0.076	0.014 / 0.002	0.070 / 0.049	0.034 / 0.105	0.084 / 0.036
A3 static	1.000 / 13.386	1.000 / 0.146	1.000 / 0.706	1.000 / 3.061	1.000 / 0.425
A4 random	0.075 / 1.000	6.865 / 1.000	1.417 / 1.000	0.327 / 1.000	2.351 / 1.000
EMAGE	1.456 / 19.492	2.503 / 0.365	4.566 / 3.223	4.403 / 13.478	8.958 / 3.811
ZS-MSTM	2.022 / 27.071	1.680 / 0.245	5.440 / 3.839	4.006 / 12.261	7.889 / 3.356
HoloGest	1.728 / 23.136	1.671 / 0.243	3.987 / 2.814	3.064 / 9.377	4.496 / 1.912
DSG+	0.684 / 9.157	0.896 / 0.130	4.466 / 3.152	2.576 / 7.885	7.327 / 3.117
ConGRets M	0.274 / 3.666	6.164 / 0.898	1.629 / 1.150	0.456 / 1.396	2.708 / 1.152
ConGRets H	0.672 / 8.993	4.746 / 0.691	1.932 / 1.364	1.224 / 3.746	3.340 / 1.421
ConGRets D	1.564 / 20.942	0.669 / 0.097	4.741 / 3.346	4.499 / 13.772	6.910 / 2.939

F.6 Runtime, availability, and reproducibility

Table 45 records the baseline generation run log, not a controlled efficiency experiment. Compute is summed per-take generation time excluding model build/load; RTF divides that compute by generated duration. Jobs shared hardware and ran on a mixture of RTX A6000 48 GB and A100 80 GB GPUs. ConGRets enrollment, generation, configuration, and verification together took 61/51/40/37 seconds for Trinity/TWH/Audio2Photoreal/Embody3D on an RTX A6000, at approximately 0.6 GB peak VRAM. These combined wall times cannot be treated as warm generation latency or compared directly to the baseline compute column. The controlled canonical-input resource measurements remain the relevant efficiency experiment.

Table 45: Recorded external-generation compute and RTF from the protocol run log. Hardware, sharing, and output durations vary; these are not normalized head-to-head speed estimates.

Pack	Baseline	Compute (GPU-h)	RTF
Trinity	EMAGE	0.007	0.019
Trinity	ZS-MSTM	0.005	0.012
Trinity	HoloGest	0.125	0.334
Trinity	DSG+	0.360	0.959
TWH	EMAGE	0.051	0.024
TWH	ZS-MSTM	0.025	0.012

Continued on next page

Table 45 (continued)

Pack	Baseline	Compute (GPU-h)	RTF
TWH	HoloGest	0.757	0.354
TWH	DSG+	2.239	1.047
Audio2Photoreal	EMAGE	0.031	0.018
Audio2Photoreal	ZS-MSTM	0.019	0.011
Audio2Photoreal	HoloGest	0.529	0.320
Audio2Photoreal	DSG+	1.772	1.071
Embody3D	EMAGE	0.032	0.018
Embody3D	ZS-MSTM	0.022	0.012
Embody3D	HoloGest	0.616	0.336
Embody3D	DSG+	1.925	1.050

The append-only evidence contains 271 raw rows: 61 for Trinity and 70 for each other pack. Trinity contains two A1/jerk rows, giving 270 unique system/arm/metric combinations overall; the numerical tables use the retained latest value without deleting either ledger entry. Forty-eight unique combinations are not valid numerical results: 32 footskate errors because the metric is declared but unimplemented, eight ineligible footskate combinations for decoder arms without world trajectories, and eight ineligible A3 BC/jerk combinations because the output is constant. These unavailable quantities are not represented as zero or treated as quality outcomes. Trinity's ten Spk-Sim combinations were not requested. SRGR is ineligible without semantic annotations; Div-Sample is not measured with only one generation seed; the audio-to-motion beat-consistency variant was not run. All eligible primary metrics were computed successfully for every generated row.

Exact identifiers for reproduction: protocol, scoring, and table-building records `docs/eval_external_transfer.md`, `docs/eval_external_transfer_results.md`, and `docs/evidence_publication_map.md`; pack construction and validation `gesture-benchmark/harness/EvalSuite/suite/external/build_pack.py` and `check_pack.py`, whose directory also holds the enrollment, baseline-generation, scoring, coverage, and table-building scripts; per-dataset results under `gesture-benchmark/results/external/` (`evals.jsonl`, `table_G.csv`, `table_G_diagnostics.csv`, `by_speaker.csv`, `coverage.json`); the BEAT seven-arm subset `beat_task_G.csv`; the ZS-MSTM reference root `BEAT_STYLE_PROCESSED`; the scored joint set `congrets50`; the arm names `retrieve`, `latent-decode`, and `direct-decode` for M, H, and D; the BEAT-memory caveat flag `congrets_library_contains_test_takes`; and the unrun metric `bc_a2m`. The manifests record source-take mappings, audio offsets and checksums, speaker identities, enrollment/scoring assignment, and frozen selection rules. Dataset/checkpoint distribution and third-party permissions remain separate artifact-availability requirements.

The results use one generation seed, have no human evaluation, no confidence intervals over independent speakers or takes, and no matched-temporal-support sensitivity analysis. Adequate numbers of feature vectors do not establish sample independence or validate the BEAT-trained evaluator on these domains. The Trinity reversal, static and random-library controls, weak GT-Sim values, and unequal coverage are retained as substantive limitations of the observed transfer evidence.

G LATENT AND SPEAKER-REPRESENTATION DIAGNOSTICS

These analyses inspect the designated main predictor (seed 1, selected epoch 2), its frozen motion encoder, and the selected speaker encoder on BEAT test speakers 12, 17, 22, 25, and 28. The motion library contains 7,655 units: 1,645, 1,475, 1,442, 1,618, and 1,475 for these speakers, respectively. Re-encoding the stored unit cores with the designated motion encoder reproduces their saved latents with maximum absolute deviation zero. The analysis uses recorded checkpoints and inputs; it does not change the trained system or its checkpoint-selection rule. Results are descriptive and do not include variation across independently trained seeds or a perceptual study.

Table 46: Motion-space diagnostics on the test-speaker library. Silhouettes use cosine distance in the full 10-D space. Augmented-pair accuracy uses the 7,168 units in complete batches; other rows use all 7,655 units.

Diagnostic	Trained	Random initialization
Augmented-pair accuracy, mean ± SD	0.7081 ± 0.0141	0.2972 ± 0.0199
Speaker silhouette	0.01205	-0.12239
Emotion silhouette	-0.14292	-0.33587
Literal dominant semantic silhouette	-0.29252	-0.43262
Non-beat semantic-label silhouette	-0.08386	-0.20849
Variance represented by PCA components 1–2	46.58%	81.17%

Table 47: Prediction-to-target cosine distributions for the main checkpoint. SD and quantiles describe the 7,655 unit observations; overlapping temporal support means they are not independent replicates.

Condition	Mean	SD	Median	5th percentile	95th percentile
Original inputs	0.638716	0.221842	0.686807	0.205859	0.915714
Different same-speaker target	0.480360	0.252469	0.506969	0.031746	0.844894
Shuffled transcripts	0.628443	0.224843	0.676488	0.188587	0.911183
Different-speaker descriptor	0.625322	0.220147	0.665184	0.200463	0.910015
Zero previous latent	0.502662	0.220340	0.522331	0.104939	0.826305

G.1 Motion representation and an untrained control

We compare the frozen 10-D motion encoder with a randomly initialized encoder of the same architecture (initialization seed 0). Both encode identical test-speaker unit cores and use the implemented clean/noisy, temporally perturbed `aug3d` views for augmented-pair matching. The comparison scores 14 complete batches of 512 units (7,168 units); the remaining partial batch is not included. The reported standard deviation is across these 14 batches, not across training seeds or a confidence interval. The matching reference level is $1/512$. These test-unit measurements differ from the historical validation accuracy, which used random two-, three-, and four-second windows on validation speakers.

The trained encoder has higher augmented-pair matching accuracy than the untrained control, supporting learned robustness to the tested perturbations. The small speaker silhouette and negative emotion/semantic silhouettes do not establish separated classes in this space. They also do not prove the absence of information that another diagnostic could recover. Emotion labels are take-level: 6,118 units are neutral, and speaker 25 contributes no emotion contrast. Literal dominant interval labels assign 6,903 units to beat. For the non-beat visualization, a unit receives its dominant non-beat label only when such intervals cover at least 25% of the unit: 1,246 units meet this condition (511 iconic, 361 metaphoric, 331 deictic, 43 habit). Another 6,386 are beat-dominated and 23 unannotated. The non-beat silhouette includes this full labeling, not only the 1,246 highlighted units.

G.2 Fusion input controls under teacher forcing

We use all 7,655 stored unit texts, encoded independently as in the full-memory query diagnostic. Each prediction receives its recorded predecessor latent and the first stored ten-minute descriptor of its speaker. The three input interventions keep two inputs fixed while (i) permuting transcripts within each speaker (seed 0), (ii) substituting the next speaker's descriptor under the fixed cycle $12 \to 17 \to 22 \to 25 \to 28 \to 12$, or (iii) replacing the previous latent with zero, the sequence-start value. A separate target control keeps the original prediction and compares it with a uniformly sampled different unit from the same speaker (seed 1). It is not an input ablation. Tables 47 and 48 retain the cosine distributions and paired effects.

The previous-latent intervention has the largest measured mean cosine effect among the three input interventions. The smaller text and descriptor effects do not establish that these inputs contain

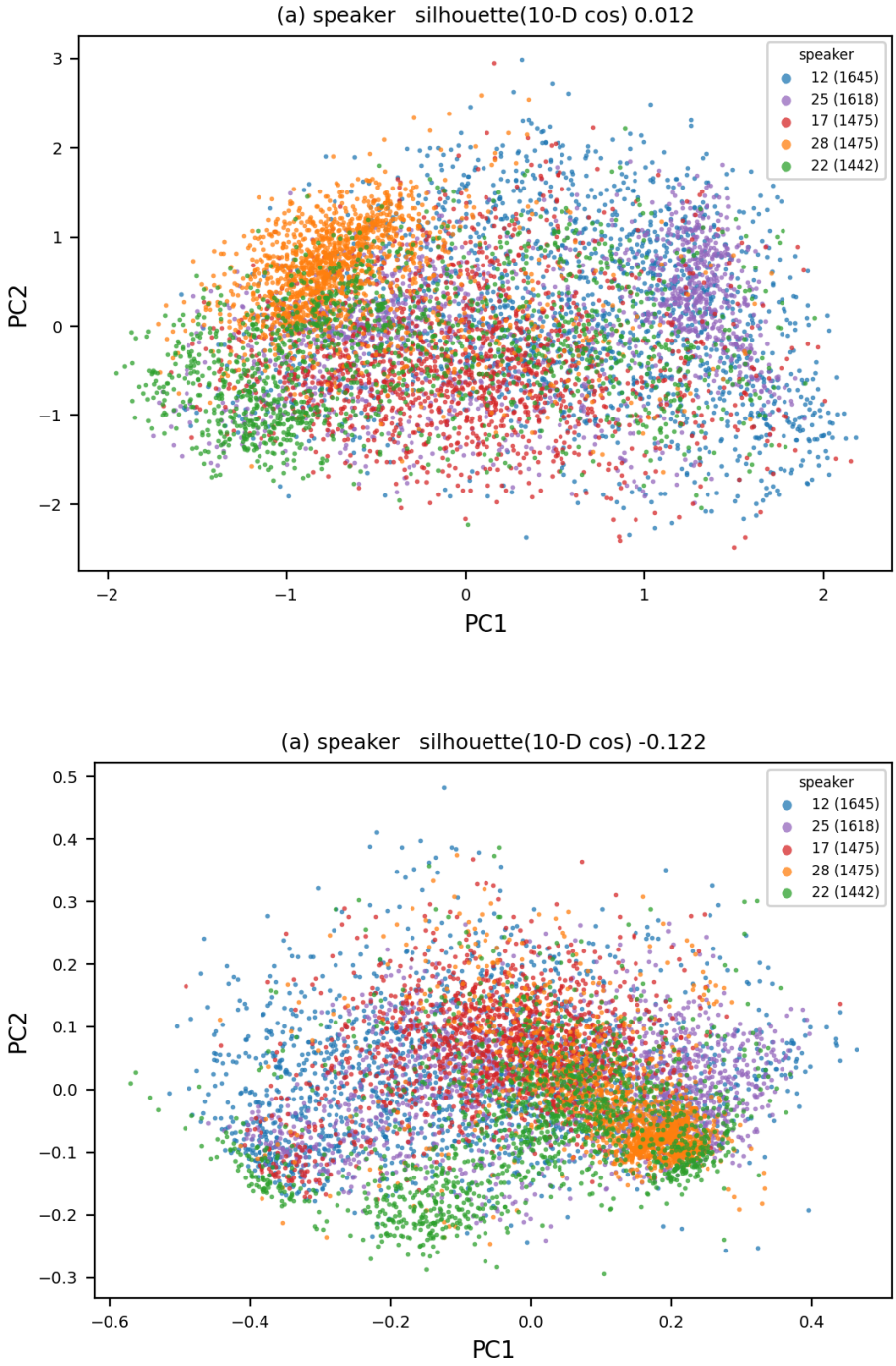


Figure 9: Motion-library PCA colored by speaker: trained encoder (top) and identical randomly initialized architecture (bottom). Each basis is fitted once by full SVD on its own pooled test-speaker latents; axes are not directly comparable between representations. All 7,655 units remain plotted. The untrained control also shows visible speaker structure; interpret both panels alongside the full-space values in Table 46.

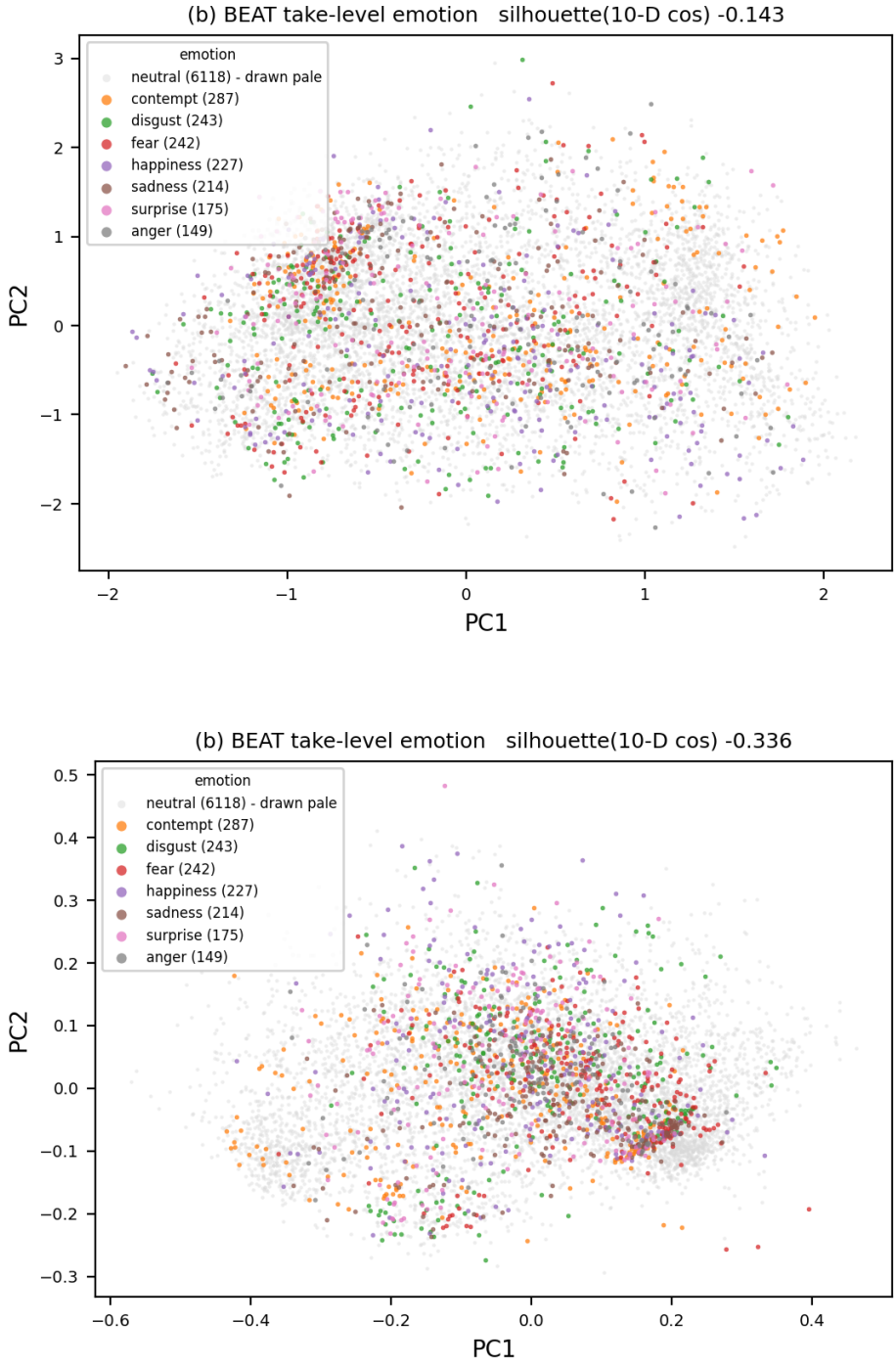


Figure 10: The same trained (top) and random-control (bottom) motion projections colored by take-level emotion. Neutral units are drawn pale, not omitted; 6,118 of the 7,655 units are neutral. The negative full-space silhouettes and coarse, imbalanced labels do not establish emotion-separated latent clusters.

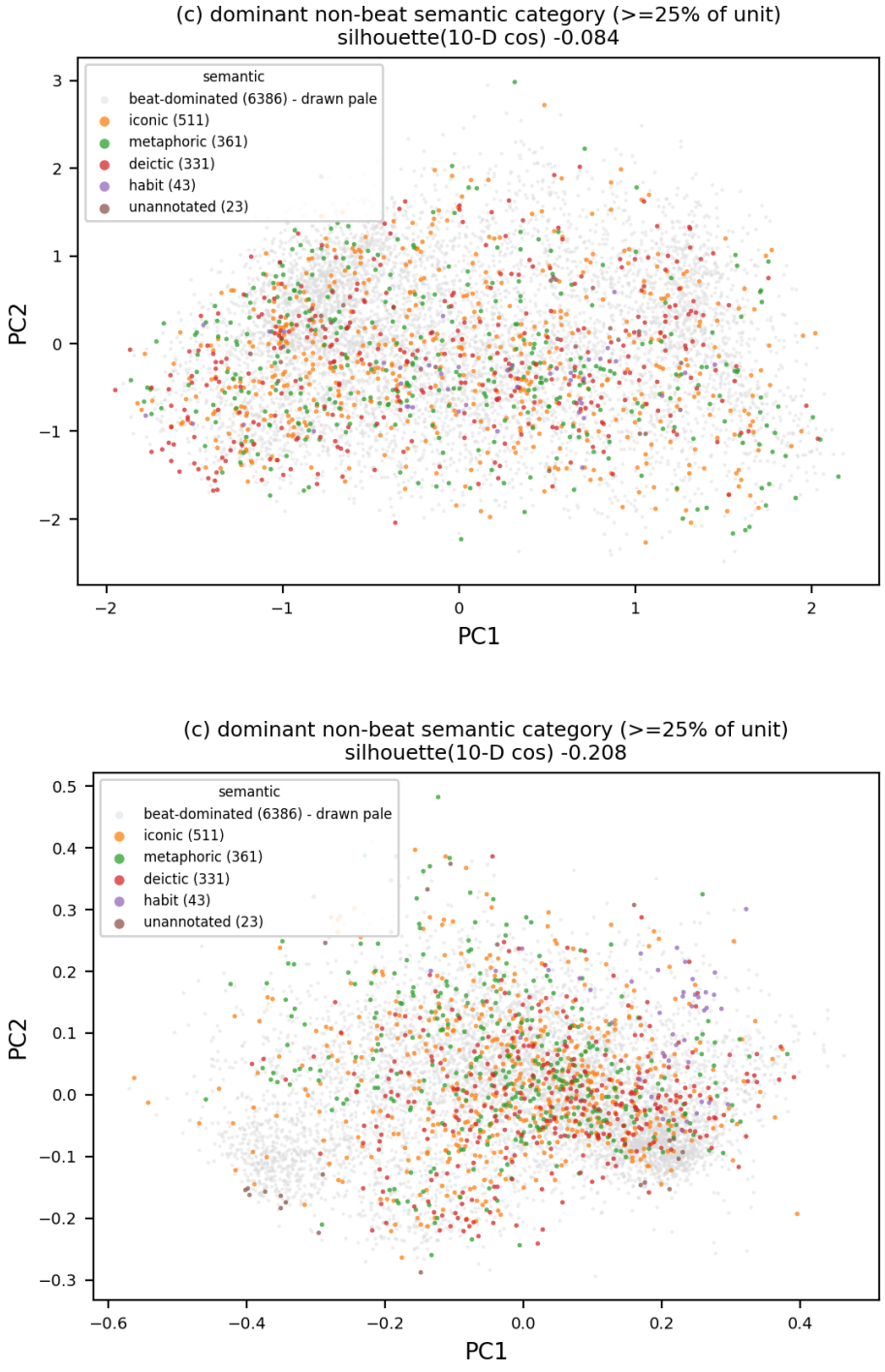


Figure 11: The trained (top) and random-control (bottom) projections colored by dominant non-beat interval label, with at least 25% non-beat coverage required for a highlighted semantic category. Beat-dominated units remain pale and unannotated units remain included. These panels retain the negative semantic-structure result; all labeling and scoring support is specified above.

Table 48: Paired cosine difference (original minus control) and fraction of units for which the original score exceeds the control. These single-input interventions are not a variance decomposition or a significance/equivalence test.

Control	Mean Δ	SD of Δ	5th percentile	95th percentile	Original wins (%)
Different same-speaker target	0.158356	0.313448	-0.338527	0.708921	69.4579
Shuffled transcripts	0.010272	0.207640	-0.332204	0.353015	52.5539
Different-speaker descriptor	0.013394	0.135993	-0.209482	0.236318	54.5003
Zero previous latent	0.136054	0.199130	-0.182959	0.476894	76.9170

Table 49: Retained-checkpoint query diagnostics, in fractions.

Checkpoint epoch	Mean cosine	SP	CC	CG
2 (reported)	0.638716	0.447076	0.044960	0.004794
11	0.606863	0.394918	0.033954	0.004811
12 / last	0.612060	0.391364	0.034306	0.002890

no information, are equivalent, or have no effect during free-running generation. No joint-input intervention, uncertainty test, or free-running conditioning intervention is reported here. A recorded predecessor already carries motion and speaker information, so the full-memory SP/CC/CG scores cannot independently attribute discrimination to text or to the enrollment descriptor.

The original query scores remain SP 0.447076, CC 0.044960, and CG 0.004794, averaged equally over speakers. The existing query table uses the uniform-unit cluster reference $\sum_c (n_{sc}/U_s)^2$ (mean 0.009734). The latent-analysis record additionally reports $1/K_s$ averaged over speakers (0.008816), a different, uniform-cluster reference. These two CC reference levels must not be interchanged. Neither is a semantic-appropriateness score.

G.3 Saved-checkpoint diagnostic

The same inputs, seeds, controls, and random target indices were evaluated at the other retained checkpoints of the same run. Epoch 2 remains the reported system. Epoch 12 and `last.ckpt` have identical recorded numerical results and are collapsed into one row below; these are not additional independent runs. The small measured text effects persist at the saved epochs inspected, which does not establish behavior at unsaved epochs or other training seeds.

G.4 One illustrative sequential trajectory

The trajectory example uses the first take of speaker 12 in the metadata order with at least six library units, `12_zhao_0_1_1`; no earlier take was skipped. It contains 30 consecutive chunks. Recorded latents, teacher-forced predictions, sequential predictions with retrieved-latent feedback, and retrieved-unit latents share the trained-library PCA projection. Mean consecutive distances in the full 10-D space are 1.359600, 1.413207, 1.205366, and 1.488211, respectively. Under retrieval

Table 50: Mean paired cosine gain and original-win percentage for each input control at saved checkpoints. Each cell is mean gain / win percentage. The target-only control is listed separately in the final column.

Epoch	Shuffled text	Other descriptor	Zero predecessor	Different target
2	0.010272 / 52.5539	0.013394 / 54.5003	0.136054 / 76.9170	0.158356 / 69.4579
11	0.009043 / 51.1300	0.023086 / 55.6107	0.119371 / 78.1450	0.129082 / 65.5519
12 / last	0.008668 / 51.0777	0.022281 / 55.5323	0.117323 / 78.0013	0.127510 / 65.9177

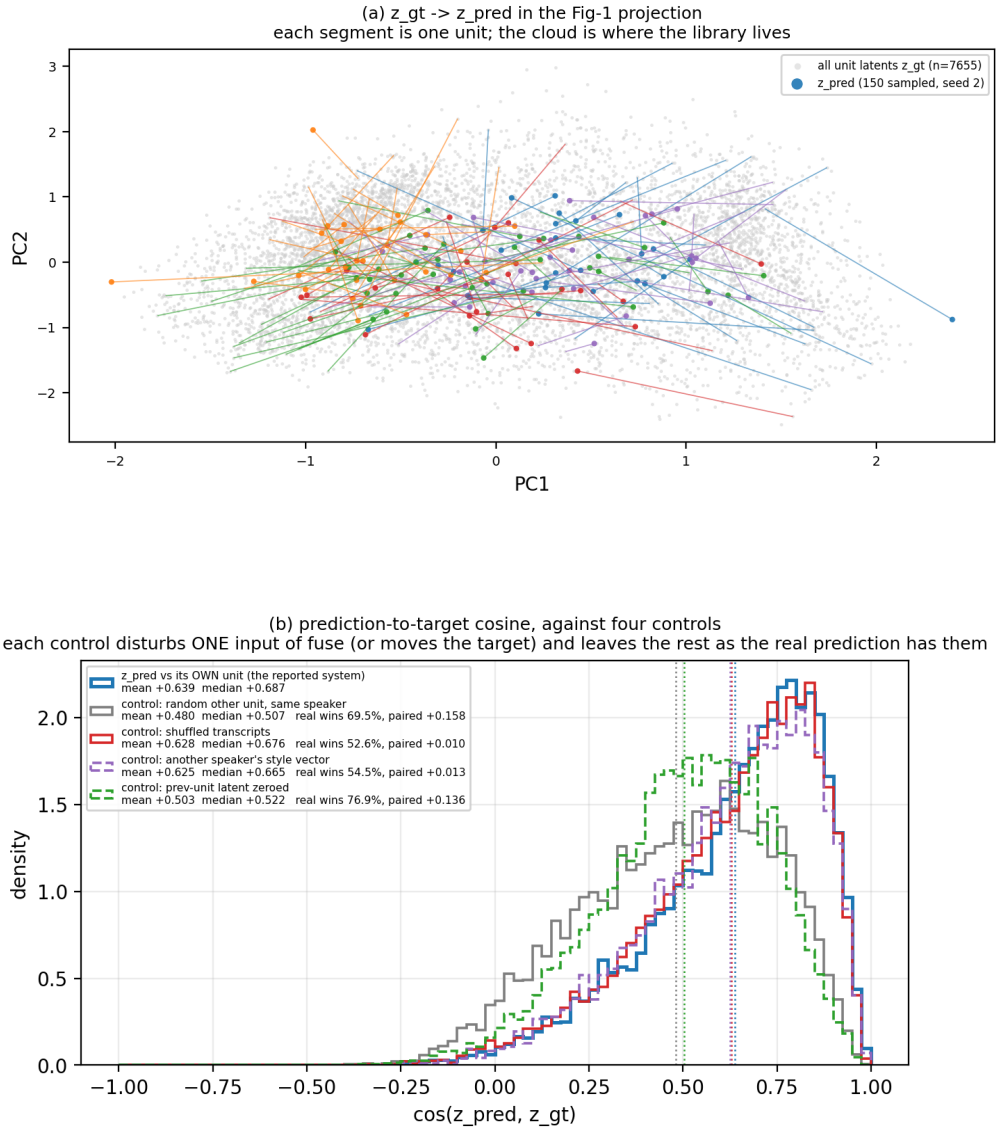


Figure 12: Teacher-forced latent predictions and controls. Top: 150 target-to-prediction segments sampled with seed 2; the panel's "Fig-1 projection" denotes the trained-library basis in Figure 9. Bottom: cosine distributions for all 7,655 units under the original inputs, three single-input interventions, and the different-target control. Each intervention holds the other inputs fixed.

seed 0, 28 of 30 selected units are distinct and none is the recorded target unit for its chunk. The stored teacher-forced prediction-to-recorded cosine mean is 0.712720; the retrieved-to-recorded mean is 0.625333. These are one-take measurements, not a distribution over takes or a test of semantic diversity. The analysis uses the loaded library and does not establish the take-excluded generation protocol for this illustrative path.

H Details moved from the main text

H.1 Method details

Units and prediction. Memory units keep ten lead-in and ten lead-out frames around each 60-frame core for transitions. The previous latent is teacher-forced from the recorded predecessor, with zeros at a take's start. The motion encoder, speaker encoder, precomputed descriptors, and decoder stay outside the predictor's training graph, and the main predictor uses no reconstruction loss. A training batch of 128 three-unit sequences is flattened to $N = 384$ predicted/target pairs, so each anchor competes with 766 negatives. Validation uses batches of 32 sequences, i.e. at most 96 target latents; the validation "retrieval accuracy" is therefore in-batch alignment accuracy over those targets, not retrieval from a speaker's full memory. Because M, H, and D feed back different latents, the modes can diverge after an identical first prediction.

Sentence-contextual text features. At generation, BERT-mini encodes each whole sentence S_t, up to 512 tokens, and pools its contextual token states over the token indices I_t of each consecutive five-word span, $e_t = \mathrm{Pool}_{I_t}(E_q(S_t))$. A final one-word tail joins the preceding span. A span wholly beyond the token limit is encoded on its own, and a partly truncated span pools its retained states. Each span is processed in order: compute Eq. 1, select a unit for M/H, realize it by Eq. 5, and feed the resulting context forward.

Transitions and output length. M inserts up to ten linearly cross-faded buffer frames between unmodified 60-frame cores. H and D overlap adjacent 60-frame decoded clips by ten frames, apply a decaying pose-offset alignment to the incoming clip, and linearly cross-fade the overlap. For T full-length chunks, the implementation therefore yields $60T + 10(T - 1)$ frames for M and $60T - 10(T - 1)$ frames for H/D.

H.2 Results details

Two objective families in Appendix C (runs E11 and E12) use the E2b motion space rather than the main model's latent space. Their changes relative to the main configuration therefore cannot be attributed solely to the prediction loss, and loss effects should be compared only within each family.

H.3 Latent query diagnostic

A full-memory, teacher-forced diagnostic over the 7,655 units of the five test speakers, with each target present, finds the prediction's nearest centroid in the correct speaker (SP) at rate 0.447 and in the paired unit's cluster (CC) at 0.0450, and the paired unit itself nearest (CG) at 0.00479, against chance levels of 0.200, 0.00973, and 0.000655 (Appendix D.2). Accuracy is above chance but low; because the recorded predecessor already carries speaker information, the diagnostic isolates neither text nor descriptor effects, nor shows that gestures suit the words (Appendix G).

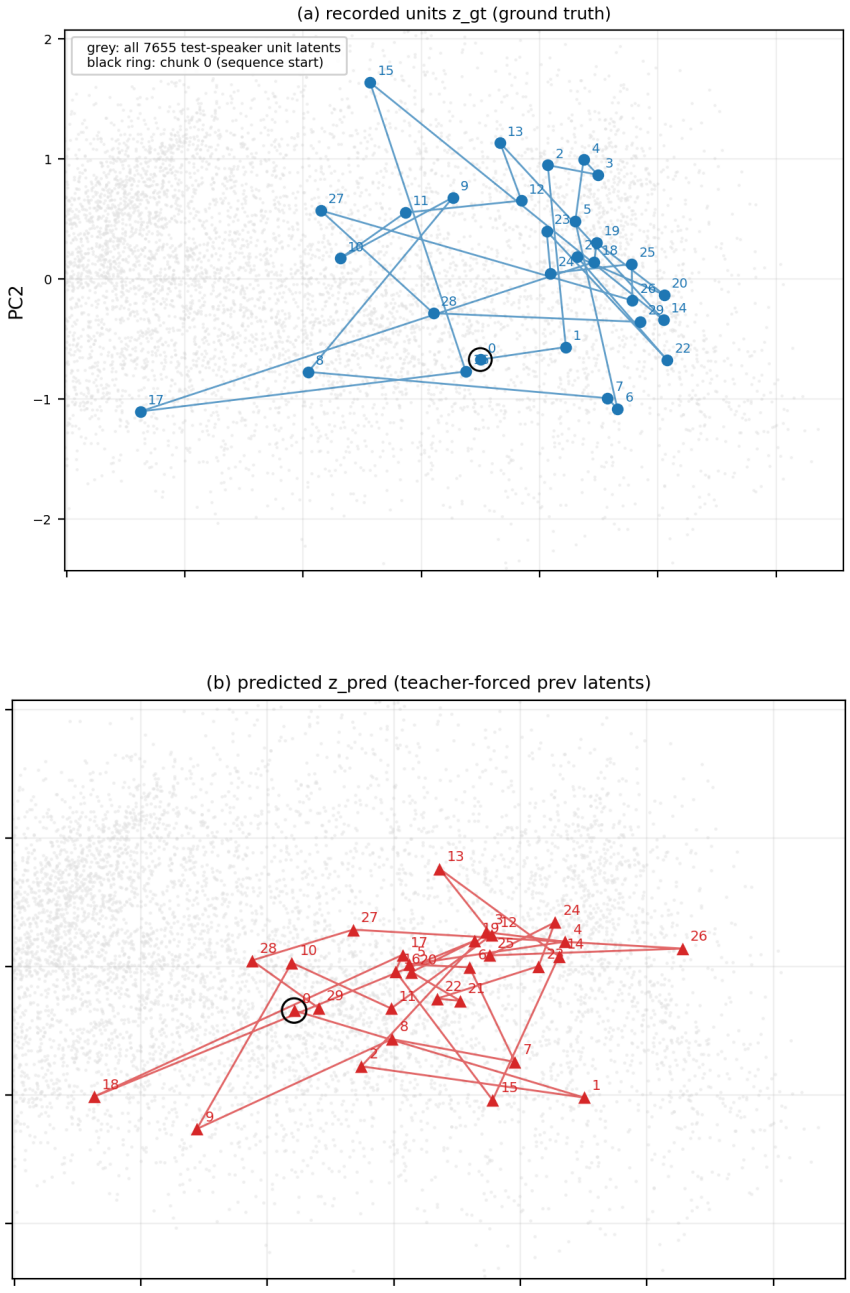


Figure 13: One 30-chunk test take: recorded unit latents (top) and teacher-forced predictions (bottom). These panels and Figure 14 share axes and the trained-library PCA basis. Point labels identify chunk order; the black ring denotes the start. This is one illustrative take, not a distribution over speakers, seeds, or utterances.

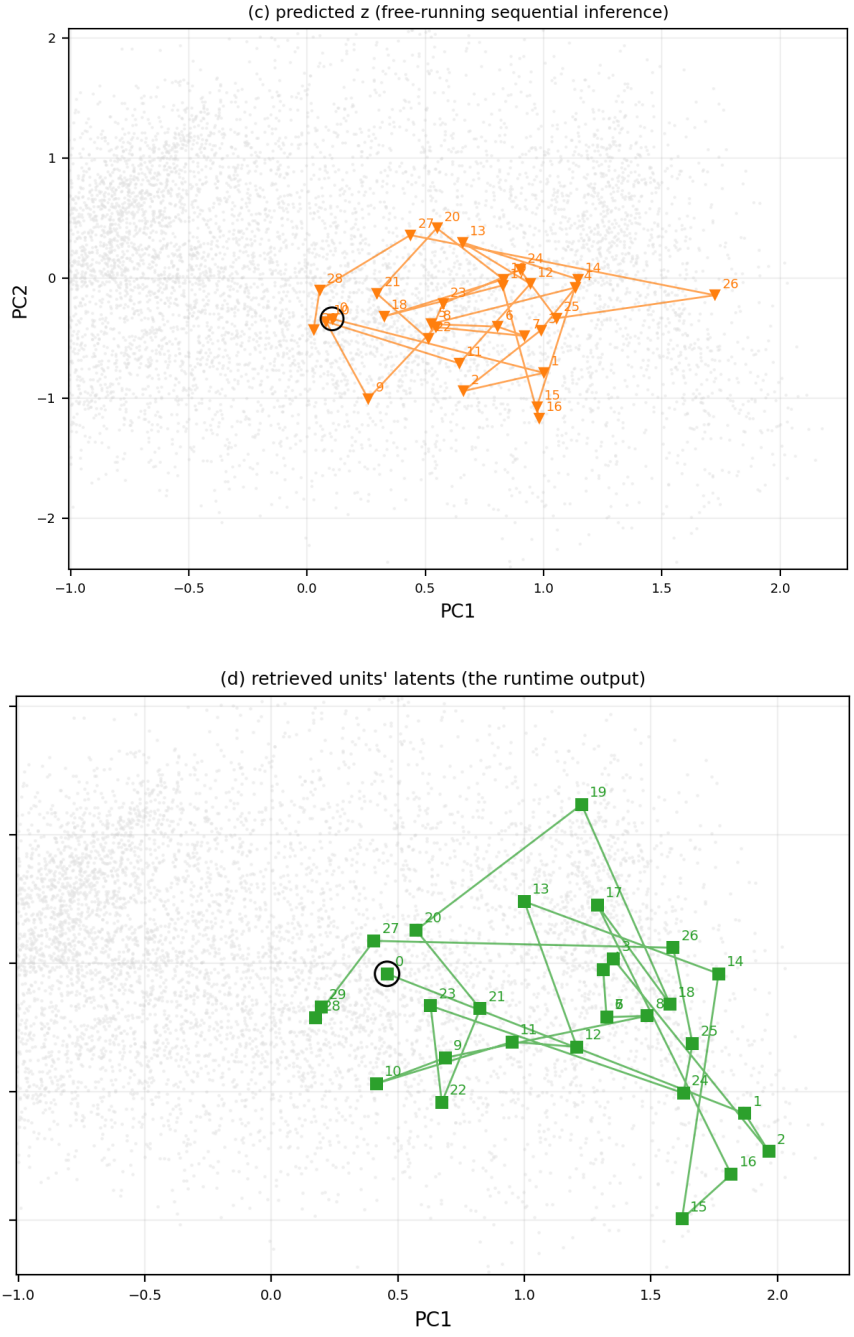


Figure 14: The same take under free-running sequential prediction with selected-latent feedback (top) and the latents of the selected units (bottom). Axes and PCA basis match Figure 13. Retrieval is stochastic (seed 0). The example describes different trajectories through a shared memory space; it does not demonstrate perceptual quality, semantic correspondence, or a general causal benefit of retrieval.

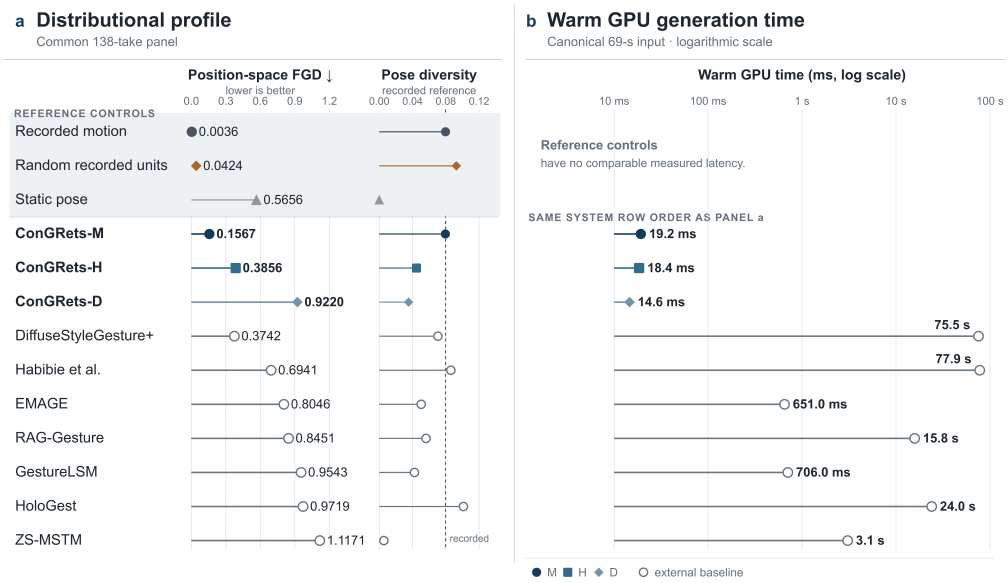


Figure 15: Distributional profile and measured generation time. (a) Position-space FGD and pose diversity on the common 138-take panel, including recorded-motion, random-unit, and static-pose controls; the dashed line marks recorded-motion diversity. (b) Warm GPU input-processing time for the canonical 69-second input, with logarithmic scaling and the same generator row order. Controls have no comparable timing measurements. A companion detail shows ConGRets warm GPU and CPU means with standard deviations over five and three repeats, respectively. Quality and timing are separate measurements; conditioning, preparation, and output-duration boundaries differ across systems (Appendices B and D).

SUPPLEMENTARY FIGURES

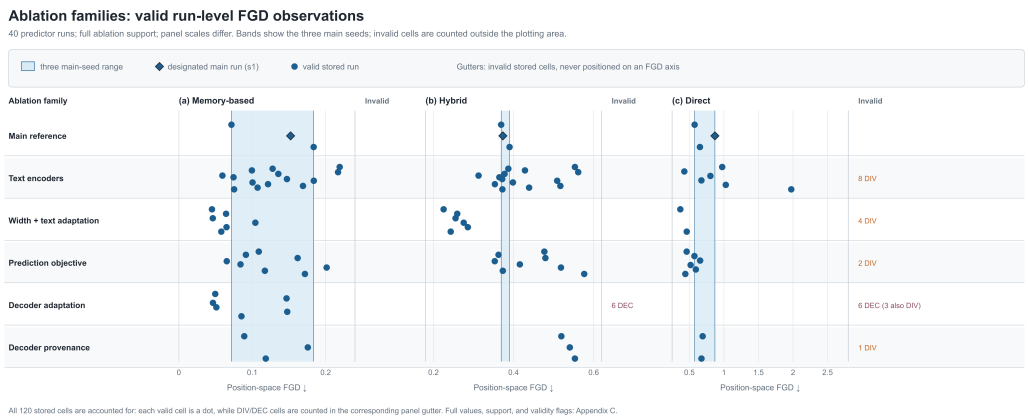


Figure 16: All final ablation families, with validity retained. Each dot represents one valid run–mode FGD measurement on the full ablation support; the shaded interval is the range of the three main seeds, not a confidence interval. Vertical offsets separate runs and have no quantitative meaning. Horizontal scales differ across modes. Invalid cells are counted outside the axes: DIV denotes a divergent direct trajectory and DEC an incorrect decoder load; overlapping flags count only once in the total. The six decoder-adaptation D cells include three with both flags. All 120 stored cells are accounted for by valid points or invalid counts. The families differ in their controls and seed coverage; this plot does not establish statistical significance. Appendix C gives every value, run identifier, actual support, and status.

COMPANION TIMING DETAIL

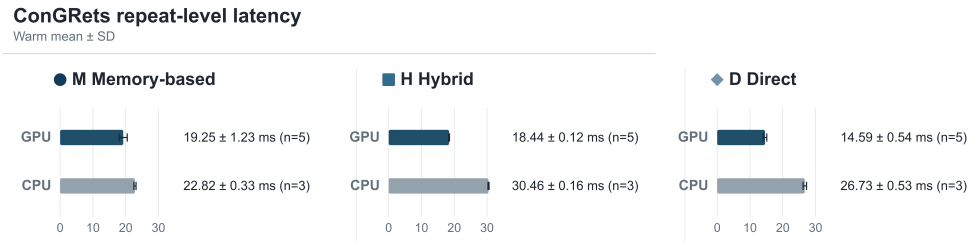


Figure 17: ConGRets warm GPU and CPU means with standard deviations over five and three repeats, respectively. Companion detail referenced in Figure 15; measurement boundaries follow Appendix D.