Do Motion Tokenizers for Co-Speech Gesture Generation Encode Gesture Semantics?

Varsha Suresh[1,2]*, Divij Jain [3]*†, Jia Liu[1], M. Hamza Mughal[2], Vera Demberg[1,2]

[1]Saarland University, [2]Max Planck Institute for Informatics, Saarland Informatics Campus, [3]IIT Delhi

vsuresh@mpi-inf.mpg.de, cs5240052@iitd.ac.in,

jili00008@stud.uni-saarland.de, mmughal@mpi-inf.mpg.de, vera@lst.uni-saarland.de

Abstract

Discrete motion tokenizers encode motion as atomic units and are widely used for co-speech gesture generation. It remains unclear which motion properties, especially those relevant to gesture semantics, are recoverable from these codebooks. We probe a reconstruction-trained codebook using 19 co-speech gesture descriptors spanning from raw motion to abstract communicative function. Results show that geometry and handedness are readily decodable from token embeddings, while motion category is only weakly decoded despite showing systematic differences in discrete code usage. This gap between reconstruction quality and descriptor decodability suggests that reconstruction objectives alone do not guarantee that gesture semantics are captured, and that evaluating codebooks on such properties can guide the design of more semantic motion tokenizers.

1 Introduction

Co-speech gesture generation increasingly relies on discrete motion codebooks, representing continuous motion as sequences of learned tokens (van den Oord et al., 2017; Zhang et al., 2023; Jiang et al., 2023; Liu et al., 2024). These codebooks are trained with reconstruction objectives that average error across all joints and frames, favoring motion that is natural and fluid rather than motion that is semantically distinctive. Fine-grained movement, such as finger configuration, contributes only marginally to this loss, since it is overshadowed by larger-magnitude body motion, even though it can be communicatively important (Brentari, 1998). Since higher-level properties like motion category and communicative intent are partly expressed through such fine-grained motion, it is unclear whether these atomic units of movement are preserved in the codebook, which is subsequently used for co-speech gesture generation.

Co-speech gesture spans multiple levels of abstraction, from raw motion coordinates to descriptors such as spatial form, handedness, handshape, and communicative function (McNeill, 1992; Kendon, 2004; Brentari, 1998). Many of these higher-level descriptors are essential for a gesture to be semantically meaningful (McNeill, 1992; Kendon, 2004), yet it is unclear whether a codebook trained purely for motion reconstruction preserves them. In this paper, we aim to examine which of these properties are present in a reconstruction-trained gesture codebook, and which are lost.

To study this, we use the semantic gesture dataset from Ram et al. (2025). Building on this dataset and prior gesture annotation frameworks (Kipp et al., 2007; Hensel et al., 2025), we define a panel of 19 descriptors spanning raw motion to abstract communicative function. Some descriptors, such as joint-level geometry, are computed automatically from skeleton coordinates, while others, including handshape, handedness, and communicative intent, are manually annotated to capture aspects of gesture that are not directly recoverable from raw coordinates. Following prior work on probing learned representations (Alain and Bengio, 2016; Belinkov, 2022; Hewitt and Liang, 2019; Pimentel et al., 2020), we use these descriptors to characterize which gesture properties are recoverable from the codebook at two complementary levels: we apply linear and shallow nonlinear probes to mean-pooled token embeddings, and we compare code-activation distributions across descriptor classes to examine whether properties are reflected in patterns of discrete code usage.

Geometry and handedness are strongly decodable from pooled token representations, whereas handshape, motion category, and communicative intent are only weakly decoded. Motion category,

*These authors contributed equally to this work.

†Work done during an internship at Max Planck Institute for Informatics.

Proceedings of the 1st Workshop on Multimodal Interaction in Face-to-Face Dialogue (MINT), pages 96–104
October 29, 2026 ©2026 Association for Computational Linguistics

however, shows systematic differences in discrete code usage. Handshape remains weakly recoverable under both analyses, suggesting that fine finger configurations, such as a cupped or pinched hand, are largely lost at the tokenization stage. We observe the same pattern on a second, architecturally distinct codebook, EMAGE (Liu et al., 2024), indicating that these results are not specific to a single tokenizer.

Our contributions are: (1) we introduce a systematic probing framework spanning 19 descriptors, from raw motion to communicative function, to characterize what reconstruction-trained gesture codebooks encode; and (2) Our results suggest that fine-grained, semantically important motion, such as handshape, may not be adequately preserved in a reconstruction-trained gesture codebook, even when the codebook achieves strong overall reconstruction accuracy.

2 Related Work

2.1 Gesture representations and descriptors

Gesture research has long analyzed movement form alongside communicative function (Efron, 1941; McNeill, 1992; Kendon, 2004; Goldin-Meadow, 2003). Computational approaches utilize raw trajectories (Kucherenko et al., 2019) to derive kinematic descriptors (Ferstl et al., 2020), or rely on manual annotations of location, orientation, handedness, and handshape (Kipp et al., 2007; Hensel et al., 2025). These methodologies parallel sign language phonology, which examines location, handshape, and movement (Stokoe, 1960; Brentari, 1998). At a higher level, gestures are categorized into semantic types (e.g., iconic, metaphoric) (McNeill, 1992; Kendon, 2004; Goldin-Meadow, 2003), which computational models often treat as distinct targets (Kucherenko et al., 2022). Recent models employ discrete tokenizers via vector quantization for general motion (Zhang et al., 2023; Jiang et al., 2023) and co-speech gesture (Liu et al., 2024). They are typically trained purely for reconstruction (van den Oord et al., 2017). While higher-level semantics are sometimes introduced externally (Zhang et al., 2024; Mughal et al., 2025), reconstruction performance alone does not clarify which interpretable properties the learned tokens inherently encode.

2.2 Probing learned representations

Probing tests whether specific properties are recoverable from learned representations (Alain and Bengio, 2016; Belinkov and Glass, 2019). We use linear probes (Hewitt and Liang, 2019; Pimentel et al., 2020) complemented by code-activation analysis to evaluate interpretable gesture features (Ghaleb et al., 2024), differentiating between simple accessibility and discrete structural encoding.

3 Data and Method

3.1 Descriptor Panel

We analyze the 19 gesture descriptors provided with the SEMANTIX dataset introduced by Suresh et al. (2026). We use the subset of 778 semantic gesture clips from TED, which originates from the gesture regions identified by Ram et al. (2025). Each clip is associated with descriptors covering physical gesture form and communicative function. These include automatically computed descriptors derived from the 3D skeleton motion, as well as manually annotated descriptors such as handshape, motion category class, and communicative function. In this paper, we retain the descriptor definitions and labels from the original study and organize the 19 descriptors into six groups to support our group-level probing and code-usage analyses. Table 2 summarizes the panel; full per-descriptor class labels are provided in Appendix A.

3.2 Codebook Architecture and Training

We analyze a two-stream vector-quantized autoencoder (VQ-VAE) (van den Oord et al., 2017) over a 38-joint 3D upper-body skeleton, partitioned into *body* and *hand* streams. Models are trained purely to minimize a position reconstruction and VQ commitment loss on a combined BEAT and TED gesture corpus. Our primary probing analyzes the best-reconstructing configuration: latent dimension $L = 256$, codebook size $K = 2048$, and temporal downsampling $d = 4$. Full architectural, configuration sweep, and training details are provided in Appendix A.1.2.

3.3 Probing Setup

We probe the codebook at two complementary levels: sequence-level probing of pooled token representations and code-level analysis of discrete code-usage patterns.

Descriptor group	N	Probing F1 Linear↑	Probing F1 Non-Linear ↑	Code usage MI ↑	Code usage Sig.	Reconstruction r(F1, MPJPE) ↓
Motion dynamics	2	$0.49_{\pm 0.07}$	$0.52_{\pm 0.08}$	0.35	1/2	−0.26
Spatial geometry	9	$\mathbf{0.77}_{\pm 0.01}$	$\mathbf{0.74}_{\pm 0.03}$	1.07	9/9	−0.69
Handedness	1	$0.73_{\pm 0.03}$	$0.66_{\pm 0.05}$	0.77	1/1	**−0.89**
Annotated spatial form	4	$0.30_{\pm 0.01}$	$0.31_{\pm 0.02}$	1.06	3/4	−0.48
Handshape	1	$0.16_{\pm 0.03}$	$0.20_{\pm 0.07}$	1.50	0/1	−0.67
Motion pattern and function	2	$0.11_{\pm 0.02}$	$0.12_{\pm 0.01}$	**3.62***	1/2	−0.45

Table 1: Group-level descriptor analysis. Linear and MLP probes use mean-pooled token embeddings; F1 is reported as mean ± SD across descriptors in each group. MI measures information in discrete code activations, and Sig. reports the number of descriptors with significant JS segregation. More negative correlations indicate that lower MPJPE is associated with higher probe performance. *This group-level MI is driven primarily by motion category (MI = 3.8, q = 0.006)

Group	Descriptor
Motion dynamics	Gesture size, Kinematics
Spatial geometry	Left/Right vertical, horizontal, depth, Left/Right elbow bend, Hand reach
Handedness	Handedness
Annotated spatial form	Hand orientation, Spatial radial, Spatial height, Spatial distance
Handshape	Handshape
Motion pattern and function	Motion category, Comm. intent

Table 2: Descriptor panel used in the main analysis.

Linear and nonlinear probing. Each gesture is represented using either mean-pooled codebook embeddings or normalized one-hot code histograms. For each descriptor, we train multinomial logistic regression with balanced class weights and report 5-fold cross-validated macro-F1 against majority-class and random-assignment baselines. We additionally train a shallow MLP with one 128-unit ReLU layer and L_2 regularization on the pooled embeddings, using identical folds and class balancing. This tests whether weak linear performance reflects nonlinear encoding while limiting probe capacity (Hewitt and Liang, 2019; Pimentel et al., 2020).

Code-level activation analysis. Each clip is represented as a normalized code-frequency histogram. We measure class separation using the mean pairwise Jensen–Shannon divergence between class-average usage profiles (Lin, 1991). Significance is assessed using label-permutation tests with Benjamini–Hochberg correction across descriptors within each stream. As a complementary measure, we sum the mutual information between each active code's presence and the descriptor label (Yang and Pedersen, 1997). Further details are provided in Appendix A.2.1.

Reconstruction fidelity. We test whether descriptor accessibility tracks reconstruction quality by correlating macro-F1 with mean per-joint position error (MPJPE) across models varying latent dimension, codebook size, and temporal downsampling. Further details are provided in Appendix A.1.2.

4 Results

4.1 Gesture Descriptor Probing

Table 1 summarizes descriptor accessibility from pooled token representations, structure in discrete code usage, and its relationship with reconstruction fidelity.

Accessibility from pooled representations. Probe performance shows a clear gradient from motion-grounded properties to higher-level gesture descriptors. Spatial geometry and handedness are the most accessible groups, reaching macro-F1 scores of 0.768 and 0.734, respectively. Performance decreases for annotated spatial form (0.299), hand configuration (0.155), and motion pattern and function (0.113). The descriptor-level comparison in Figure 1 shows the same pattern: geometric properties such as left and right depth exceed 0.89 macro-F1, whereas handshape, motion category, and communicative intent remain close to their chance baselines. A shallow MLP does not change this ordering; motion pattern and function remains effectively unchanged (0.113 → 0.112). Its weak linear accessibility is therefore unlikely to reflect shallow nonlinear encoding alone.

Structure in discrete code usage. Activation analysis asks whether descriptor classes use different subsets of codes. Higher MI indicates that code activations are more informative about the

descriptor, while $q < 0.05$ indicates reliable separation after multiple-testing correction. Spatial geometry is significant for all nine descriptors, and handedness is also reliably separated. Most notably, motion pattern and function has the highest group-level MI (3.62), driven by motion category (MI $= 3.8$, $q = 0.006$), despite its weak pooled probing performance. This suggests that motion category is encoded in *which* codes are used, but is obscured by pooling. This does not hold for all higher-level properties: communicative intent ($q = 0.06$) and handshape ($q = 0.70$) show no reliable overall separation. Thus, weak pooled probing reflects hidden code-level structure for motion category, but limited evidence for handshape and most communicative intents.

Relationship with reconstruction metrics Negative correlations indicate that lower reconstruction error is associated with higher probe performance. This relationship is strongest for handedness ($r = -0.88$) and spatial geometry ($r = -0.69$), suggesting that reconstruction quality mainly tracks motion-grounded properties.

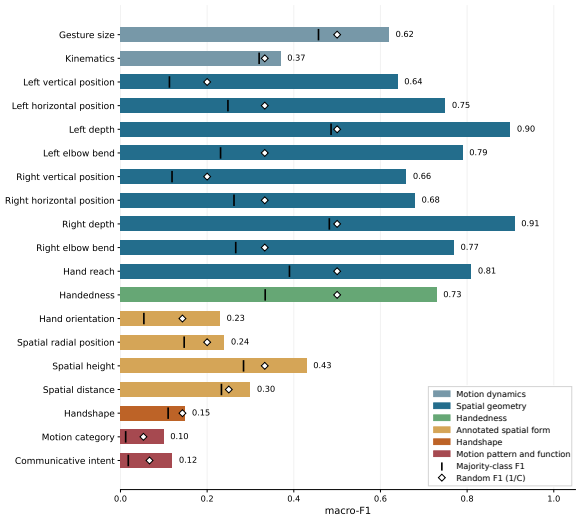

Figure 1: Probe macro-F1 for each descriptor

4.2 Generalization to Another Codebook

We repeat the analysis on EMAGE (Liu et al., 2024). As shown in Figure 2, both codebooks exhibit the same accessibility gradient: spatial geometry is strongest, while properties such as handshape remain weak. Although absolute scores differ slightly, the ordering of descriptor groups is stable across architectures. This consistency suggests that the accessibility gradient is not specific to a single tokenizer design.

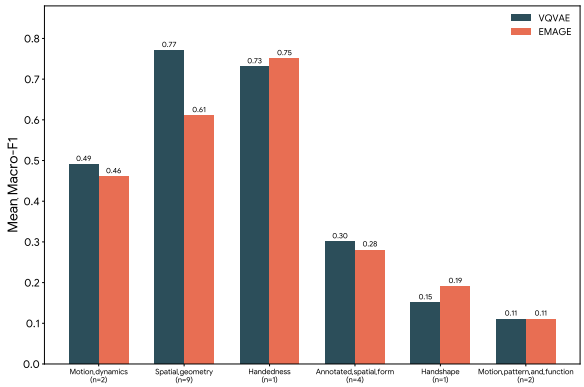

Figure 2: Probing macro-F1 by descriptor group, for EMAGE.

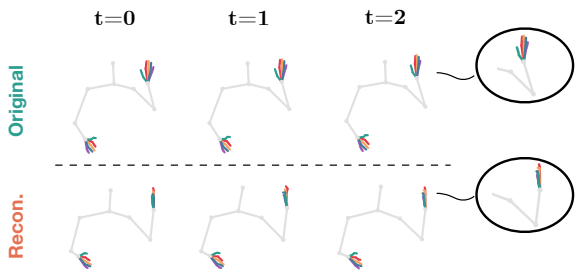

Figure 3: Qualitative comparison showing that reconstruction-only VQ-VAEs preserve upper-body geometry but average out fine-grained hand poses

5 Qualitative Case Study

Figure 3 illustrates a pattern that corroborates probing performance on communicative functions. VQ-VAE based tokenizers trained solely with a reconstruction loss, learn representations that decode into averaged-out motions. In particular, hands do not achieve same fidelity at reconstruction as compared to upper body joints, due to difference of scale between hands and upper body. Therefore, better performance on upper body results in strong accessibility of geometric descriptors. In contrast, poorly reconstructed hand poses depict different communicative functions, which eventually leads to subpar performance on gesture functions. These findings highlight that reconstruction losses do not incentivize learning of gesture semantics as part of the representation learning in codebooks.

6 Conclusion

Across two codebook architectures, geometric properties and handedness are readily recoverable, while handshape, motion category, and communicative intent remain weak. Better reconstruction and

finer temporal resolution mainly improve motion-grounded properties, suggesting that reconstruction quality alone is not a sufficient measure of semantic accessibility. Gesture codebooks should therefore be evaluated using both probing and code-level analyses, and future tokenizers should explicitly preserve communicatively relevant properties.

Limitations

Our analysis is limited to 778 semantic gestures from TED and two skeleton-based, reconstruction-trained tokenizers, so the observed accessibility gradient may not generalize to other domains, motion representations, or tokenizer objectives. In addition, we did not ablate the composition of the training corpus as the codebook was trained on BEAT and TED jointly, so the contribution of each dataset to descriptor accessibility remains untested. Manual descriptors, particularly handshape and communicative intent, are relatively sparse and may contain annotation subjectivity. Moreover, low probe performance indicates limited recoverability under the selected pooled representations and probe families; it does not establish that a property is entirely absent or inaccessible to a sequence-aware model. Future work should evaluate larger and more diverse datasets, use temporal probes, and test whether descriptor accessibility predicts the semantic quality of generated gestures.

Acknowledgments

This work is funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - GRK 2853/1 “Neuroexplicit Models of Language, Vision, and Action” - project number 471607914."

References

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *Preprint*, arXiv:1610.01644.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Diane Brentari. 1998. *A Prosodic Model of Sign Language Phonology*. MIT Press.

David Efron. 1941. *Gesture and Environment*. King’s Crown Press.

Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2020. Understanding the predictability of gesture parameters from speech and their perceptual importance. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA)*.

Esam Ghaleb, Bulat Khaertdinov, Wim Pouw, Marlou Rasenberg, Judith Holler, Asli Ozyurek, and Raquel Fernández. 2024. Learning co-speech gesture representations in dialogue through contrastive learning: An intrinsic evaluation. In *Proceedings of the 26th International Conference on Multimodal Interaction*, pages 274–283.

Susan Goldin-Meadow. 2003. *Hearing Gesture: How Our Hands Help Us Think*. Harvard University Press.

Laura B. Hensel, Stephanie Cheng, and Stacy Marsella. 2025. A richly annotated dataset of co-speech hand gestures across diverse speaker contexts. *Scientific Data*, 12:1748.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2733–2743.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079.

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

Michael Kipp, Michael Neff, and Irene Albrecht. 2007. An annotation scheme for conversational gestures: How to economically capture timing and form. *Language Resources and Evaluation*, 41(3):325–339.

Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents (IVA)*, pages 97–104.

Taras Kucherenko, Rajmund Nagy, Michael Neff, Hedvig Kjellström, and Gustav Eje Henter. 2022. Multimodal analysis of the predictability of hand-gesture properties. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. 2024.

Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1144–1154.

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

M Hamza Mughal, Rishabh Dabral, Merel CJ Scholman, Vera Demberg, and Christian Theobalt. 2025. Retrieving semantics from the deep: an rag solution for gesture synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16578–16588.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622.

Ashwin Ram, Varsha Suresh, Artin Saberpour Abadian, Vera Demberg, and Jürgen Steimle. 2025. Gesturecoach: Rehearsing for engaging talks with llm-driven gesture recommendations. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, pages 1–15.

William C. Stokoe. 1960. Sign language structure: An outline of the visual communication systems of the american deaf. Technical report, University of Buffalo.

Varsha Suresh, Mohammad Mahdi Abootorabi, Mohamed Salman, M Hamza Mughal, Christian Theobalt, Ashwin Ram, Jürgen Steimle, and Vera Demberg. 2026. Semantic motion anchors: Bridging motion and meaning in co-speech gestures. *arXiv preprint arXiv:2605.30608*.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML)*, pages 412–420. Morgan Kaufmann.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. 2023. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14730–14740.

Zeyi Zhang, Tenglong Ao, Yuyao Zhang, Qingzhe Gao, Chuan Lin, Baoquan Chen, and Libin Liu. 2024. Semantic gesticulator: Semantics-aware co-speech gesture synthesis. *ACM Transactions on Graphics (TOG)*, 43(4):1–17.

A Appendix

A.1 Additional Details: Data and Method

A.1.1 Descriptor Classes

Table 3 shows the detailed classes in each descriptor.

A.1.2 Codebook Architecture and Training

We analyse a two-stream vector-quantised autoencoder over a 38-joint 3D upper-body skeleton. The joints are partitioned into a *body* stream of 8 coarse joints and a *hand* stream of 30 finger joints. The two streams have separate encoders and codebooks and are quantised independently; a single shared decoder reconstructs the full pose. The hand joints are expressed relative to their wrist before encoding, meaning the hand stream encodes wrist-local finger configuration while the body stream encodes coarse-joint positions.

Each stream's encoder is a 1D temporal convolutional stack (kernel 4, stride 2) that downsamples time by a factor d. Each codebook is a vector-quantised bank trained with exponential-moving-average updates (decay 0.99) and dead-code reinitialisation. Tokenisation processes each clip in non-overlapping d-frame chunks, yielding one token per stream per chunk.

To map how representational granularity affects descriptor accessibility, we swept a $3 \times 3 \times 3$ grid of 27 models over $L \in \{64, 128, 256\}$, $K \in \{1024, 2048, 4096\}$, and $d \in \{4, 8, 16\}$. Every model is trained purely to reconstruct motion. The codebook is trained on a combined 3D gesture-motion corpus comprising the BEAT and TED datasets (processed at native frame rates of 30 FPS and 24–30 FPS respectively, balanced by duration). Models are optimised with the Adam optimiser (learning rate 1×10^{-3}, batch size 8) for up to 25 epochs (with early stopping), minimising a position reconstruction loss plus the VQ commitment loss weighted by $\beta = 1.0$.

A.1.3 Probing Setup

Table 4 reports descriptor-level linear probing results. The main pattern is consistent with the group-level results in the main paper. Automatic geometric descriptors and handedness are the strongest, while handshape, motion category, and communicative intent remain weak.

A notable trend across these descriptors is that the continuous codebook vector (Embedding) outperforms the code histogram (One-hot) for almost

Group	Descriptor	Annotation	C	Labels
Motion dynamics	Gesture size	Manual	2	normal; big
Motion dynamics	Kinematics	Manual	3	slow; normal; fast
Spatial geometry	Left vertical position	Automatic	5	below-waist; waist-level; torso-level; chest-level; above-head
Spatial geometry	Left horizontal position	Automatic	3	body-center; torso-side; extended-outward
Spatial geometry	Left depth	Automatic	2	at-torso; in-front-of-torso
Spatial geometry	Left elbow bend	Automatic	3	straight; slightly-bent; bent
Spatial geometry	Right vertical position	Automatic	5	below-waist; waist-level; torso-level; chest-level; above-head
Spatial geometry	Right horizontal position	Automatic	3	body-center; torso-side; extended-outward
Spatial geometry	Right depth	Automatic	2	at-torso; in-front-of-torso
Spatial geometry	Right elbow bend	Automatic	3	straight; slightly-bent; bent
Spatial geometry	Hand reach	Automatic	2	mid-reach; extended
Handedness	Handedness	Manual	2	one hand; both hands
Annotated spatial form	Hand orientation	Manual	7	inward; outward; toward center; up; down; side; forward
Annotated spatial form	Spatial radial position	Manual	5	inward; front; side; out; far out
Annotated spatial form	Spatial height	Manual	3	belt; chest; head
Annotated spatial form	Spatial distance	Manual	4	touch; close; normal; far
Handshape	Handshape	Manual	7	palm; cup; pinch; fist; bunch; fingers; index finger
Motion pattern and function	Motion category	Manual	19	arc-like trajectory; counting; diagonal; flick; glide; horizontal; long-tail; pointing; pull; push; rotate; shake; step; strike; sway; sweep; vertical down; vertical up; wave
Motion pattern and function	Communicative intent	Manual	15	contrast; discourse; emotion; emphasis; enumeration; listing; negation; other-reference; progression; quantification; relativity; self-reference; symbolise; temporal reference; uncertainty

Table 3: The 19-descriptor panel used for probing. Automatic descriptors are computed and discretized from skeleton coordinates, whereas manual descriptors are human annotated. C denotes the number of retained classes after removing labels represented by fewer than five clips.

all descriptors. The advantage is largest on graded spatial/positional descriptors (e.g., vertical reach bias, spatial distance), where a continuous vector successfully captures gradation that a discrete histogram naturally discards. One-hot representations only edge ahead on a handful of low-signal descriptors (like handshape), where the absolute F1 is already near the noise floor.

A.2 Additional Results

A.2.1 Code-Level Activation Procedure & Results

To rigorously evaluate whether a descriptor is reflected in the use of individual codebook entries, we employ an information-theoretic approach.

Jensen-Shannon divergence provides a symmetric and bounded metric for scoring how much the classes' profiles differ. The summed presence Mutual Information (MI) provides an additive measure of how much information the active vocabulary collectively carries about the label, avoiding the combinatorial infeasibility of estimating a full joint contingency table $I(Y; K)$ over all 2048 codes.

Table 5 reports the descriptor-level results. The data reveals that weak linear performance does not possess a single interpretation. Some descriptors leave strong code-level traces despite failing linear probing (e.g., motion category, $q = 0.006$), indicating the information lives in *which* discrete units are activated rather than their averaged vector geometry. Conversely, other abstract properties like communicative intent ($q = 0.063$) and handshape ($q = 0.697$) fail to reach broad code-level significance, indicating that these features are genuinely absent from the token representation.

A.2.2 Further examples for the code-level metrics.

To clarify the code-level metrics, we describe what they return in two extreme cases and in two real descriptors. In the first extreme, both classes of a binary descriptor use every code equally often. The two class-average profiles are then identical, so JS $= 0$, and because knowing which code fired tells us nothing about the class, MI $= 0$ as well. In the opposite extreme, each class uses only its own code and never the other's. The profiles share no codes, so JS reaches its maximum of 1 bit, and observing a single code reveals the class with certainty, so MI is maximal.

Real descriptors sit between these two extremes. For handedness, both classes share generic resting codes, but "one hand" clips concentrate on a few specific codes (e.g., hand code 746) while "both hands" clips use a separate set of bimanual codes. This partial overlap gives intermediate values (JS $= 0.31$, MI $= 0.766$), and the permutation test confirms the separation is far above what random labels would produce ($q = 0.006$). Left depth classes are heavily imbalanced (94% vs. 6%), so the minority profile is estimated from very few clips and looks spiky. JS weighs both profiles equally regardless of how many clips they come from, so it becomes inflated (JS $= 0.83$). MI instead accounts for how rare the minority class is, so it stays moderate (MI $= 0.493$). Reading JS together with

its permutation q-value and MI therefore prevents false discoveries caused by class imbalance.

Group	Descriptor	Cls.	Maj-F1	Rand-F1	Embed.	One-hot
Motion dynamics	Gesture size	2	0.457	0.500	**0.619**	0.562
Motion dynamics	Kinematics	3	0.320	0.333	**0.367**	0.340
Spatial geometry	Left vertical position	5	0.113	0.200	**0.639**	0.563
Spatial geometry	Left horizontal position	3	0.248	0.333	**0.749**	0.611
Spatial geometry	Left depth	2	0.486	0.500	**0.898**	0.884
Spatial geometry	Left elbow bend	3	0.231	0.333	**0.787**	0.669
Spatial geometry	Right vertical position	5	0.119	0.200	**0.663**	0.467
Spatial geometry	Right horizontal position	3	0.262	0.333	**0.679**	0.566
Spatial geometry	Right depth	2	0.482	0.500	**0.912**	0.858
Spatial geometry	Right elbow bend	3	0.266	0.333	**0.772**	0.487
Spatial geometry	Hand reach	2	0.390	0.500	**0.809**	0.751
Handedness	Handedness	2	0.334	0.500	**0.734**	0.664
Annotated spatial form	Hand orientation	7	0.054	0.143	**0.227**	0.201
Annotated spatial form	Spatial radial position	5	0.147	0.200	**0.236**	0.228
Annotated spatial form	Spatial height	3	0.284	0.333	0.434	**0.438**
Annotated spatial form	Spatial distance	4	0.233	0.250	**0.300**	0.297
Handshape	Handshape	7	0.110	0.143	0.155	**0.197**
Motion pattern and function	Motion category	19	0.012	0.053	**0.105**	0.083
Motion pattern and function	Communicative intent	15	0.018	0.067	**0.121**	0.101

Table 4: Descriptor-level linear probing macro-F1 with chance baselines. **Cls.** is the number of retained classes (≥ 5 samples). **Maj-F1** is the majority-class macro-F1 and **Rand-F1** is the random-assignment macro-F1 ($= 1/C$ in expectation). For near-balanced binary labels such as depth and handedness, Rand-F1 exceeds Maj-F1, so the higher of the two is the relevant chance level. The best of the two probe representations is shown in bold. Spatial geometry and handedness clear chance by large margins, while motion category, communicative intent, and handshape exceed chance only marginally.

Group	Descriptor	Cls.	Stream	MI	q
Motion dynamics	Gesture size	2	hand	0.397	**0.006**
Motion dynamics	Kinematics	3	hand	0.311	0.327
Spatial geometry	Left vertical position	5	hand	1.849	**0.006**
Spatial geometry	Left horizontal position	3	hand	1.056	**0.006**
Spatial geometry	Left depth	2	hand	0.493	**0.006**
Spatial geometry	Left elbow bend	3	hand	1.039	**0.006**
Spatial geometry	Right vertical position	5	hand	1.851	**0.006**
Spatial geometry	Right horizontal position	3	hand	1.024	**0.006**
Spatial geometry	Right depth	2	hand	0.575	**0.006**
Spatial geometry	Right elbow bend	3	hand	0.875	**0.006**
Spatial geometry	Hand reach	2	hand	0.865	**0.006**
Handedness	Handedness	2	hand	0.766	**0.006**
Annotated spatial form	Hand orientation	7	hand	1.802	**0.006**
Annotated spatial form	Spatial radial position	5	hand	1.191	**0.006**
Annotated spatial form	Spatial height	3	hand	0.534	0.208
Annotated spatial form	Spatial distance	4	hand	0.697	**0.021**
Handshape	Handshape	7	hand	1.499	0.697
Motion pattern and function	Motion category	19	hand	3.790	**0.006**
Motion pattern and function	Communicative intent	15	hand	3.450	0.063

Table 5: Descriptor-level code activation results using the stream with the highest mutual information for each descriptor. Bold q-values indicate FDR-corrected $q < 0.05$. Spatial geometry is consistently significant. Motion category is the main higher-level descriptor that shows a reliable code-level trace despite weak linear probing performance.