Evaluating Automatic Hand-Gesture Generation Using Multimodal Corpus Annotations: The Benefits of a Multidisciplinary Approach

Mickaëlla Grondin Verdon*
Domitille Caillat*
mickaella.grondin-verdon@loria.fr
domitille.caillat@univ-monpt3.fr
Multispeech, Lorraine University,
CNRS, Inria, LORIA
F-54000, Nancy, France
Praxiling, UMPV, CNRS
Montpellier, France

Louis Abel
Multispeech, Lorraine University,
CNRS, Inria, LORIA
F-54000, Nancy, France

Slim Ouni
slim.ouni@loria.fr
Multispeech, Lorraine University,
CNRS, Inria, LORIA
F-54000, Nancy, France

ABSTRACT

This exploratory study addresses the challenges of evaluating the quality of hand-gesture synthesis. It introduces an interdisciplinary methodology aimed at providing objective evaluation criteria. The study examines expert annotations applied on a small dataset combining both natural and synthetic gestures, showing how their comparison can reveal key indicators for assessing communicative efficiency and adequate movement dynamics. Communicative gestures are more frequent, shorter, and easier to interpret in natural data, while synthetic gestures are more ambiguous, with less precise annotations and less consistent velocity profiles. These findings support the idea that only an interdisciplinary approach —combining computational modeling with insights from gesture studies in the language sciences— can yield meaningful criteria for evaluating and ultimately improving the quality of synthesized gestures.

CCS CONCEPTS

• **General and reference → Evaluation**; • **Computing methodologies → Model verification and validation**; *Motion capture*; Neural networks; *Cross-validation.*

KEYWORDS

Deep learning architectures, Artifical co-speech gestures, Linguistic expertise, Corpus Enrichement, Model Assessment

ACM Reference Format:
Mickaëlla Grondin Verdon, Domitille Caillat, Louis Abel, and Slim Ouni. 2025. Evaluating Automatic Hand-Gesture Generation Using Multimodal Corpus Annotations: The Benefits of a Multidisciplinary Approach. In *Proceedings of the International Workshop on Generation and Evaluation of Non-verbal Behaviour for Embodied Agents (GENEA '25), October 27–28, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3746268.3759430

[*] Both authors contributed equally to this research.

GENEA '25, October 27–28, 2025, Dublin, Ireland

ACM ISBN 979-8-4007-2050-5/2025/10
https://doi.org/10.1145/3746268.3759430

1 INTRODUCTION

Designing a gesture synthesis model requires taking into account the complex relationship between gestures and speech. Indeed, foundational works [20, 21, 26, 27] have shown that gestures naturally generated by speakers have their own value, semantic, pragmatic, or syntactic, and thus facilitate the speech interpretation [14, 19, 29]. Understanding and modeling these complex multimodal components is a major challenge for developing credible human-machine communication, particularly through embodied conversational agents (ECAs) [9, 10]. In this context, research on co-speech gesture generation has focused on two main objectives: achieving physiological realism (human-likeness) and ensuring communicative efficiency (speech appropriateness).

While the automatic generation of gestures for ECAs is garnering increasing interest, as evidenced by the recent surge in research and challenge on this topic [5, 15, 18, 22, 23], current generation models still struggle to produce natural and effective gestures. For this reason, the Syncogest project [7] aims to explore an interdisciplinary approach that combines the development of computational models with findings from the gesture studies in language sciences. Led by computer scientists and linguists, Syncogest aims to measure the benefit of integrating expert multimodal annotations into the development of AI training designed for classifying and generating co-verbal manual gestures. This initiative is grounded in the premise that accurate replication of appropriate gestures can only be achieved by considering the different aspects of conversational gestures.

Gestures, defined as visible movements of body parts used to communicate [8, 20, 21, 26, 27] complement verbal discourse by conveying meaning and intention. In order to study how these two levels interact, gesture studies often rely on the annotation of both speech (e.g., global utterances, phonetic units, prosodic features) and various formal and functional aspects of gestures. These include gesture detection and segmentation (i.e., identifying meaningful units and their different phases such as preparation, stroke —the gesture's core—, retraction, and hold [20]), their form (eg.: body-parts involved), their link to the speech (eg.: lexical affiliates [31], prosodic affiliates [16]), and their function [11–13] —often categorized using standard typologies [20, 21, 26, 27], which identifies distinct communicative roles.

This paper presents the results of part of our exploratory study. It attempts to formalize objective types of criterion for evaluating gestures automatically generated through annotations made by specialists in multimodal linguistics. Our assumption is that expert annotation of multimodal data may help establish more objective evaluation methods, as recent studies have pointed out the inconsistency and limited comparability of traditional assessments [28]. Indeed, a comparative approach to the annotation of natural and generated data can help determine more precise criteria for evaluating the quality of generated gestures, in terms of formal aspects, communicative contribution, proportions and distributions.

To support this thesis, we conducted a detailed annotation (segmentation, functional classification, lexical affiliation) of the gestures naturally produced by a speaker and the gestures automatically generated from the same short verbal production. These annotations are then compared, allowing us to identify several formal and functional aspects that could serve as criteria for more objectively evaluating the quality of generated gestures and, consequently, provide insights for improving the effectiveness of the generation and recognition model.

2 METHODOLOGY

2.1 Dataset description

2.1.1 Original corpus. This study is based on The Body Expression Audio Text (BEAT) corpus [24], a large-scale, multimodal, and multilingual dataset containing approximately 60 hours of English recordings (among other languages). BEAT provides motion capture data for the body, hands, and face, recorded at 120 Hz with 16 synchronized cameras and Vicon suits equipped with 77 reflective markers, alongside corresponding audio. Annotations were also provided —including words and phonemes, gesture segmentation and classification, and lexical affiliations—, but they suffer from major inconsistencies that make them difficult to use [17]. As part of our exploratory project aimed at improving the accuracy and consistency of multimodal gesture corpora —within which the present study is situated— we reassessed the annotations of 14 randomly selected files, all taken from English rehearsed monologues by the Wayne speaker.

2.1.2 The STARGATE model. The synthesized data of this study were generated using the STARGATE model, a deep-learning based model developed using all of Speaker Wayne's files of the BEAT corpus [3, 4] and which leverages audio and text transcription to generate co-verbal manual gestures [2–4]. The model, built on a chunked-autoregressive architecture, generates multiple gesture frames per step, which are then used as input to produce the next segment. It consists of three encoders—audio, text, and motion. The motion encoder employs a spatio-temporal graph convolution network [32] to project input into a latent space while preserving graph structure, whereas the audio and text encoders use traditional convolutional networks. The latent spaces are concatenated and fed to the decoder to generate the next gesture frames. STARGATE processes sliding windows: 1 second of past motion and 2 seconds of audio and text (1 second past, 1 second future). This model was selected for this study for two main reasons. First, it was trained exclusively on the Wayne speaker, allowing for a reliable comparison with the natural data used in our exploratory study, without the risk of bias that could arise from individual variation. Secondly, it outperformed previous state-of-the-art approaches, producing not only rhythmic gestures but also more complex ones.

Table 1: Natural and synthetic data file description.

	Natural data	Synthesis data
File duration	1 minute 13 seconds	1 minute 11 seconds
Audio (english)	BEAT	BEAT
BVH origin	BEAT	STARGATE
Annotated By	Syncogest	Syncogest
FPS	120	60
Finger motion	Yes	No
Agent visual		

2.1.3 Data selection. The present analysis focused on File 2 from speaker Wayne (right-handed), which provided a basis for both gesture synthesis testing and natural–synthetic gesture comparison. Two version of this file were then compared (Table 1): natural (Nat.) data and synthesis (Synt.) data, both featuring the same audio and text from BEAT but differing in motion capture (BVH from BEAT in Nat. and BVH from STARGATE for Synt.). BVH files were used to reconstruct the movement on ECAs, as shown in Table 1. Their durations are 1min13s and 1min11s, respectively. Due to a 1s cut at the beginning of the Synt. file —caused by the model's architecture, which prevents it from generating the first and last seconds of gestures from a given input—, we added 64 empty video frames to synchronize both files for later comparisons.

Natural data kept BEAT corpus FPS, while synthesis was rendered at 60 FPS. Although this should be monitored long-term, this had no noticeable visual impact for the annotators. The joints in the *.bvh* files were recalculated with the hip as the reference point $(0, 0, 0)$ as shown in Fig.1, and all joint positions were expressed relative to this reference in centimeters for both data. The natural data includes detailed finger movements, whereas the synthesized data only provides fingertip positions. Given the importance of fine finger articulation for gesture analysis, the initial absence of finger joints in STARGATE was a limitation that affected certain aspects of the comparative analysis, as will be discussed later. This limitation should be addressed in the long term. The small size of the dataset is justified by the objectives of the present study: it is not intended to produce measurements generalizable to the diversity of speakers' gestural styles or to the outputs of different generation models, but rather to explore, through this brief comparison, the fundamental principles and benefits of an evaluation methodology for synthetic gestures based on expert annotations.

2.2 Annotation process

2.2.1 General protocol. To conduct this study, both Nat. and Synt. files were manually annotated, following a protocol designed to serve our objectives. The process involved using PRAAT [6] and ELAN [1] and had two main steps: first, verbal transcription and annotation (text, word segmentation, POS tagging), then gesture identification and annotation (functional classification), together forming an initial multidimensional, multimodal classification.

Both files were annotated by the same two experts in gestural analysis to ensure consistency. An initial synchronization phase was conducted, during which four of the 14 files used in our overall exploratory project were collaboratively annotated to establish a strong consensus on both formal and semantic criteria underlying the annotations. Then, each annotator independently annotated the remaining files. This annotation process was complemented by a joint review of all annotated data, providing a cross-validation step that ensured a level of accuracy and reliability rarely achieved — whether in linguistic corpora, where single-expert annotation remains common due to the labor-intensive nature of the task, or even more so in computational corpora, such as the BEAT corpus, where annotation quality is often significantly hindered by insufficient annotator training. During the joint revision, each functional categorization was assigned a score reflecting the degree of certainty of both annotators. Disagreements on segmentation or interpretation that could not be resolved during this phase were annotated as such. The annotation was not blind, as the natural and synthetic data were visually distinct due to differences in agent appearance and finger motion. However, even under ideal conditions distinguishing natural gestures from synthetic ones would remain highly perceptible, as will be demonstrated later, raising a persistent objectivity issue despite the use of the same blind annotation protocol.

2.2.2 *Annotation template.* The annotation template used in this study consists of multiple tiers, each serving a specific purpose in capturing various aspects of the data. The tiers include verbal transcriptions, gesture labels, and additional categorizations to further structure the data, among which those used in this study and their main aspects are listed in Table 2.

Verbal and prosodic annotations. Transcription and segmentation are carried out with an emphasis on accuracy and alignment with the spoken material, using the audio provided in the original corpus. This includes sentence-level segmentation, orthographic transcription of individual words using an automatic speech-recognition model (ASR), manually revised, and segmentation of words and phonemes using the Montreal Forced Aligner (MFA) [25], based on Kaldi [30]. This study does not examine gesture-speech coordination, planned for future Syncogest work. The exception is lexical affiliation (LA), where gestures were manually linked to at least one lexical item by gesture ID when relevant.

Gesture annotations. The annotations concern manual gestures. They follow a structured set of categories designed to capture the communicative roles of gestures after segmentation, such as semantic, syntactic and pragmatic functions. In addition to these categories, defined in Table 2, two special labels —*Unclear* and *Undefined*— were introduced to address ambiguous cases. The *Unclear* label is assigned when annotators disagree due to differing interpretations, regarding the segmentation, the communicative nature, or the specific function of an activity. In contrast, the *Undefined* label is used when both annotators express doubt about the communicative nature of an activity or its alignment with existing functional categories. This nuanced annotation system accommodates the complexity and ambiguity of gesture-based communication and enhances the ability to analyze gesture-speech coordination.

Table 2: Extract from the Annotation Template for Each File.

Tier	Labels	Definition
Sentences		Punctuated transcription per sentence.
Words		Orthographic transcription per word.
LA		Word(s) affiliated with a gesture.
Act. Phase	*Stroke*	Corresponding to a gesture's core.
	Unclear	Disagreement on interpretation.
Act. Type	*Communicative*	Movement interpreted as serving a communicative role. ***= Gesture.***
	Unclear	Disagreement on interpretation.
	Undefined	Shared doubts on interpretation.
Gest. Type	*Butterworth*	Hesitation support at the verbal level.
	Designation	Designates to a person or thing.
	Metaphoric	Illustrates metaphorically what is said.
	Modal	Illustrates a discursive modality.
	Parsing	Marks syntactic structure.
	Quasilinguistic	Comprehensible without speech.
	Spatial	Refers to a mentioned place.
	Temporal	Marks a temporal aspect.
	Unclear	Disagreement on interpretation.
	Undefined	Shared doubts on interpretation.
Manuality	*TH*	Use of both hands.
	LH	Use of the left hand.
	RH	Use of the right hand.

Certainty Score

0 = No certainty.	*3* = Moderate certainty.
1 = Very low certainty.	*4* = High certainty.
2 = Low certainty.	*5* = Absolute certainty.

LA = Lexical affiliation; *Act.* = Activity; *Gest.* = Gesture

2.3 Categories of gesture annotations

The gesture annotation categories include annotations related to the formal level and to the functional level, with each gesture uniquely identified by an identification number.

2.3.1 *Formal level.* The first two tiers focus on observing formal characteristics and dynamic properties, such as mobility and velocity. The **activity tier** identifies temporal sequences that are likely to correspond to gestures rather than mere movements (which were not annotated in this study). Based on Kendon's classification of gesture phases [20, 21], these activities are then categorized on the **activity phase tier** as corresponding to what could be the core phase of a gesture, then labeled as *Stroke*, or labeled *Unclear* when there is a disagreement between the annotators regarding this correspondence. Thus segmentation corresponds to the earliest start and the latest end timestamps among annotators. Each activity is also annotated on the manuality tier, specifying the hand(s) involved in the gesture execution: two hands (TH, either symmetrical or asymmetrical), right hand (RH), or left hand (LH).

2.3.2 Functional level. Several types of tiers provide insights into the functional status of the annotated activities. When labeled as *Communicative* on the **activity type tier**, activities are those that clearly and unequivocally serve a communicative purpose. They correspond to gesture, contributing semantic, pragmatic, or syntactic value to the interaction. Conversely, activities labeled as *Undefined* or *Unclear* on this tier are those whose communicative status remains uncertain for both annotators or for which no consensus could be reached between them. Each activity labeled as *Communicative* is then categorized on at least one **gesture type tier** as gestures may have more than one functional role. Gesture types are then annotated for the first functional dimension (*d1*) and when present for the second functional dimension (*d2*), primarly based on [20, 21, 26, 27]. They include several categories that highlight the specific roles gestures play in speech, namely in the two files compared: *Butterworth*, *Designation*, *Metaphoric*, *Modal*, *Parsing*, *Quasi-linguistic*, *Spatial* and *Temporal* gestures, as defined in Table 2. Note that only the functional categories observed in the study corpus are detailed — beat gestures, for example, are absent. The labels *Unclear* and *Undefined* were used when the annotators either disagreed or were unable to precisely determine the role of a gesture. Apart from these two cases, all functional labels were also annotated in the corresponding tier for each dimension with a certainty score ranging from 0 (no confidence) to 5 (absolute confidence) as described in Table 2.

2.4 Analysis methodologies

The annotations were analyzed to derive quantitative and qualitative insights highlighting the main aspects in which, in our dataset, the synthetic gestures differ from the natural gestures — thereby shedding light on the various potential types of limitations of the sample model.

2.4.1 Quantitative analysis. Data management and descriptive statistics were performed in RStudio, focusing on occurrences (Occ. or N), average duration (AD, in seconds), and median duration (MD, in seconds).

2.4.2 Inter-annotator agreement analysis on annotation timestamps. The time difference between our annotations was analyzed for gestures with common annotation identification. Absolute differences were calculated for both the start and end times of annotations, avoiding compensatory effects of positive and negative values. This approach, however, does not preserve the directionality of the differences (i.e., whether annotator (Ann.) 1's timing is earlier or later than annotator 2's).

2.4.3 Activity trajectory and position analysis. Gesture trajectories for each annotated activity were reconstructed in an animated 3D space based on the positions of the right and left index fingertips, as this choice provides a more expansive and thus more comprehensive visualization. Only hand(s) involved were annotated in the manuality tier. Color coding was applied based on their vertical distance (cm) from the reference point (hip = 0, in blue) to the top of the head (red), with points below the hip colored dark blue for enhanced interpretability of movement dynamics. Position plots visualize the motion of the right index fingertip, with its position relative to the hip (in cm) represented along three dimensions: laterality (right>0>left), verticality (down>0>up), and depth (backward>0>forward), where values indicates the position relative to the hip (= 0). The line plot represents the fingertip's position at each frame.

Table 3: Activity Phase and Activity Type Annotation Details.

	Synthesis			Natural		
	N	*AD* *(seconds)*	*MD*	*N*	*AD* *(seconds)*	*MD*
Stroke	25	0.584	0.519	39	0.352	0.32
Unclear	16	0.706	0.669	-	-	-
Communicative	31	0.660	0.625	38	0.354	0.322
Unclear	10	0.545	0.605	-	-	-
Undefined	-	-	-	1	-	-
Activity	**41**	**0.632**	**0.625**	**39**	**0.352**	**0.32**

2.4.4 Velocity computation. Velocities were derived from the successive 3D positional data $[x, y, z]$ of the right index fingertip which is involved in nearly all of the annotated activities in both datasets. Assuming a constant time interval $\Delta t = 1/fps$, velocity vectors were calculated as the difference between successive positions, with a zero vector $[0, 0, 0]$ added for the first frame. The Euclidean norm of the velocity vectors was computed to convert them into scalar values (cm/s).

3 RESULTS

3.1 Annotation analysis

The analysis of the annotation categories, as summarized in Table 3, reveals a quite similar number of annotated activities in both the synthesized (41) and natural (39) datasets. However, significant differences emerge between the two files.

3.1.1 Activity phases. In the synthesized dataset, the activity phase annotations reveal that the majority of annotated activities are identified as *Strokes*, accounting for 25 occurrences (61% of the annotations). The remaining 16 activities (39%) were marked as *Unclear* phases due to disagreements among annotators regarding this correspondence. In contrast, all 39 activities annotated in the natural dataset are consistently identified as *Stroke* phases (100%). Synthesis annotated activities are longer overall, with mean and median durations (mean = 0.632s, median = 0.625s) nearly double those of natural activities (mean = 0.352s, median = 0.32s), with a slightly shorter duration for activities identified as corresponding to a *Stroke* phase (mean = 0.584s, median = 0.519s) compare to activities labeled as *Unclear* (mean = 0.706s, median = 0.669s). Both the disparities in activity phase categorization and activity duration point to a clearer identification of the gesture's core in the natural dataset, while highlighting ambiguities in gesture identification within the synthesized dataset.

3.1.2 Activity types. Table 3 also highlights differences in activity types between synthesis and natural data. In the synthesis dataset, 31 of the 41 activities are interpreted as gesture and then labeled *Communicative* (76% of annotations); the other 10 occurrences were all subject to disagreement between the annotators regarding their

Table 4: Gesture Type Annotation Details.

	Synthesis data			**Natural data**					
	d1			**d1**			**d2**		
	N	*AD* (*seconds*)	*MD*	*N*	*AD* (*seconds*)	*MD*	*N*	*AD* (*seconds*)	*MD*
Undefined	16	0.671	0.635	14	0.343	0.374	-	-	-
Unclear	10	0.612	0.598	-	-	-	-	-	-
Parsing	3	0.441	0.473	8	0.293	0.319	1	-	-
Metaphoric	2	1.135	1.135	4	0.691	0.642	1	-	-
Designation	-	-	-	5	0.240	0.217	2	0.646	0.646
Modal	-	-	-	3	0.348	0.403	1	-	-
Spatial	-	-	-	1	-	-	2	0.302	0.302
Quasi-ling.	-	-	-	1	-	-	1	-	-
Butterworth	-	-	-	1	-	-	-	-	-
Temporal	-	-	-	1	-	-	-	-	-
Gesture type	**31**	**0.660**	**0.625**	**38**	**0.354**	**0.322**	**8**	**0.408**	**0.336**

communicative status and labeled as *Unclear* (24%). No activities were annotated as *Undefined* in the synthesis data. In the natural dataset, activities interpreted as *Communicative* dominate with 38 occurrences (97% of annotations). No communicative status was subject to disagreement (the *Unclear* category is absent in the natural data), but one of the activity is labeled *Undefined* as its communicative status remains uncertain for both annotators.

The higher frequency of *Communicative* activity types and the absence of *Unclear* activity types in the natural dataset suggest easier interpretability. In contrast, the synthesized dataset, characterized by a significant proportion of *Unclear* annotations activity types, reflects greater ambiguity. Both activities labeled as *Communicative* (mean = 0.659s, median = 0.625s) and *Unclear* (mean = 0.545s, median = 0.605s) in the synthesis data are significantly longer than the communicative activities in the natural data (mean = 0.354s, median = 0.322s). Durations exhibit homogeneity, with means close to medians, suggesting a relatively uniform distribution. This may support the notion that, even when gestures were interpreted as communicative, annotators faced challenges in precisely identifying their boundaries.

3.1.3 Gesture types. Gesture type annotations for the first (*d1*) and second (*d2*) dimensions are compared by Table 4. All activities labeled as *Communicative* have at least one functional label. Note that no *d2* dimension was found in synthesis data. In synthesis data, the gesture type of the 32 communicative activities are labeled in *d1* as follow: *Undefined* (16 occ.), *Unclear* (11 occ.), *Parsing* (3 occ.) and *Metaphoric* (2 occ.). In the natural data, where 38 activities are interpreted as *Communicative*, *Undefined* gestures (15 occ.) remain the most prominent gesture type in *d1*, followed by *Parsing* (8 occ.), *Metaphoric* (4 occ.), *Designation* (5 occ.), and *Modal* gestures (3 occ.). Fewer isolated gestures were annotated as *Spatial*, *Quasi-linguistic* (Quasi-ling.), *Butterworth*, and *Temporal*. In natural data *d2*, only 8 gesture types were annotated as a second functional dimension. *Spatial* gestures (2 occ.) were noted, along with single instances of *Quasi-linguistic*, *Parsing*, *Metaphoric*, and *Modal* gestures. While synthesized data (*d1*) and natural data (*d1*) show similar occurrences of *Undefined* gestures (difficulty of interpretation), only the synthesized data includes the *Unclear* gesture type (disagreement on interpretation), and to a significant extent. This suggests a lack of interpretability in the synthesized data, in contrast to the natural data which also displays a greater diversity of specific gesture types. As previously noted, synthetic gestures are generally longer than natural gesture. *Metaphoric* gestures are consistently longer than all other types in both datasets, with durations exceeding the overall average. In synthetic data, *Parsing* gestures are notably shorter than the average, suggesting more precise identification or concise execution. Conversely, *Undefined* gestures in synthetic data are much longer, with average durations nearly double those in natural data. The duration could serve as an indicator of the difficulty in interpreting gestures, especially when compared to natural data.

Table 5: Annotations of Activity types per sentences.

Sent. id	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
↪*Words*	*47*	*24*	*16*	*12*	*19*	*17*	*36*	*26*	*9*
Natural (LAc = 24)									
Activities	**8**	**4**	**5**	**2**	**2**	**2**	**8**	**6**	**2**
↪Und.	-	-	-	-	-	-	1	-	-
↪CA	8	4	5	2	2	2	7	6	2
↪LAc	**8**	**2**	**5**	**2**	**1**	**1**	**4**	**2**	**1**
Synthesis (LAc = 4)									
Activities	**9**	**3**	**4**	**3**	**3**	**3**	**9**	**5**	**2**
↪Unc.	2	1	-	-	-	1	5	1	-
↪CA	7	2	4	3	3	2	4	4	2
↪LAc	**1**	**-**	**1**	**-**	**-**	**2**	**-**	**-**	**-**

3.1.4 Verbal distribution. Table 5 highlight various aspects of the distribution of annotations in relation to the text by providing detailed statistics on sentences (Sent.), including the number of words, annotated activities, communicative activities (CA), and lexical affiliate count (LAc). Across both datasets, which include 9 sentences totaling 206 words, longer sentences, such as Sent. 1 and Sent. 7, generally contain more activities (*Communicative*, *Unclear* (Unc.), or *Undefined* (Und.)), reflecting a logical correlation between sentence length and the number of annotations. However, when focusing solely on gestures (communicative activities), this correlation is more evident in the natural data than in the synthesized data. In addition, many gesture in the synthesized dataset lack lexical affiliations throughout the entire file (only 4 lexical affiliations; LAc =4), unlike those in the natural dataset (LAc =24), as detailed also in Table 5. This observation aligns with the difference in gesture types annotated in both files. Gestures labeled as *Undefined* and *Unclear*, which are highly prevalent in the synthesized data, are annotated as such due to their difficult interpretability, and inherently lack a clear connection to the speech. In contrast, gestures identified with functional types (*d1* and *d2*), which are more frequent in the natural data, typically have corresponding lexical affiliations that are frequently established in both datasets.

3.2 Annotation Agreement analysis

3.2.1 Interannotator agreement on annotation timestamps. The Table 6 compares the time differences in the annotation timestamps of activities between annotators. For the synthesized data, the median difference is 0.117s, with a mean of 0.195s, whereas the natural

Table 6: Annotation Duration and Timestamp Variability.

File	Annotation time		Revision time	Timestamps diff.	
	Ann. 1	*Ann. 2*		*AD(s)*	*MD(s)*
Synthesis	0:37:58	1:20:00	12:00:00	0.117	0.195
Natural	1:24:40	2:13:00	02:00:00	0.050	0.084

Ann. = Annotator; *Diff.* = differences (s).

dataset shows much smaller differences, with a median of 0.050s and a mean of 0.084s. These duration differences highlight a higher consistency in the natural data annotations, where the identified activities, aligning reliably with the gesture's stroke, were annotated at nearly identical positions. In contrast, annotations in the synthesized data exhibit greater variability, reflecting a lower level of agreement on he precise boundaries of gestures.

3.2.2 Certainty scores of gesture type annotations. The certainty scores for gesture type annotations across natural and synthesis files are presented by Table 7. For the natural data, dimension *d1* has a mean score of 4.05 and a median of 4, while dimension *d2* shows slightly higher scores (mean = 4.13, median = 4). These values indicate strong confidence in gesture type interpretation, as scores of 4/5 reflect high certainty as observed for overall gesture types identified (N=32). Conversely, the synthesis data demonstrates significantly lower scores, with dimension *d1* (N=5) reporting a mean of 1.4 and a median of 1, indicating very low certainty in gesture type interpretation. This disparity underscores the challenges associated with determining gesture functions in the synthesized dataset. This finding is consistent with the lower number of gesture types identified in the synthesized files compared to the natural data, with both observations reinforcing the gap in interpretability between the two datasets.

3.2.3 Annotation and revision time. Table 6 compares the times spent by the two annotators on the annotation and revision process across both datasets. For both the synthesized and natural datasets, annotator 2 required significantly more time to complete the annotations (natural: 2h13; synthesized: 1h20) compared to annotator 1 (natural: 1h24; synthesized: 38min), a trend that is consistent across all 14 re-annotated files in the main study. Additionally, both annotators spent substantially less time annotating the synthesized data than the natural data. This indicates that a more complete and detailed annotation process was possible for the natural data than for the synthesized data, as activity type and gesture type labels that cannot be determined (*Undefined* and *Unclear*), which occur more frequently in the synthesized data, do not require further annotation details. Table 6 also highlights a notable difference in the revision time between the annotation of synthesized and natural data. The annotators spent only 2 hours revising the natural data annotations, compared to 12 hours revising the synthesized data annotations. This time disparity reflects the more ambiguous or less consistent nature of the synthesized data. Indeed, it demonstrates the challenge for annotators in relying on the criteria usually used to segment and interpret gestures properly, leading to a longer revision period for this dataset. In contrast, the natural dataset allowed for a more straightforward revision process.

Table 7: Certainty Scores of Gesture Type Annotations.

	Natural			Synthesis		
	N	*Mean*	*Median*	**N**	*Mean*	*Median*
d1	24	4.05	4	5	1.4	1
d2	8	4.13	4	-	-	-
Gesture type	**32**	**4.06**	**4**	**5**	**1.4**	**1**

3.3 Kinematic and Spatial Properties

3.3.1 Manuality. Information on the distribution of manuality types observed across both datasets is provided by Table 8. It specifies which hand(s) are involved in performing the activities: either both hands move in the same way, only the left hand is used, or only the right hand. The table reveals that in both datasets, the left hand is rarely used on its own. However, in the natural data, the speaker uses both hands almost as frequently as he uses only his right hand, while in the synthetic data, gestures are predominantly performed with bimanual movements. The limited number of activities involving only the left hand in the synthetic data may be explained by the fact that, as Wayne is right-handed, this configuration is generally rare in the data used for training. But the prevalence of TH gestures in the synthetic data could reflect certain formal complexities of gestures. Indeed, even when a specific hand is clearly engaged, the other hand is rarely completely static, as it is often at least influenced by the overall movement of the body.

3.3.2 Trajectories and positions. In Fig. 1, we observe the superimposed trajectories of the engaged hands across all activity annotations. For the natural data Fig.1A, there is a noticeable gradient in color, with activities reaching higher into the space, evident in the presence of orange and red, and a broader distribution of movements both in height and along the sides. This suggests that the speaker utilizes a large and dynamic space for gestures. In contrast, the synthesis data Fig.1B shows several limitations. The color range is more restricted, mostly around light blue and green, indicating movements slightly above the hip. Gestures are placed noticeably lower overall, and show less variation in height or horizontal range, with trajectories generally concentrated in the same spatial region. Additionally, the hands are often positioned similarly, performing movements with minimal diversity, with an excessive symmetry and more constrained motion profiles. Complementarily, Fig.2 displays positions of the right index fingertip in space. Movement peaks of the index fingertip closely follow the velocity peaks, reflecting the dynamics of the motion and highlights a significant difference in the movement profiles of the index finger between natural and synthesized activities. The synthesized gestures exhibit

Table 8: Manualities for Annotated Activities and Gestures.

	Natural		Synthesis	
	Activity	*Gesture*	*Activity*	*Gesture*
Two Hands	19	19	31	28
Left Hand	3	3	1	-
Right Hand	17	16	9	3

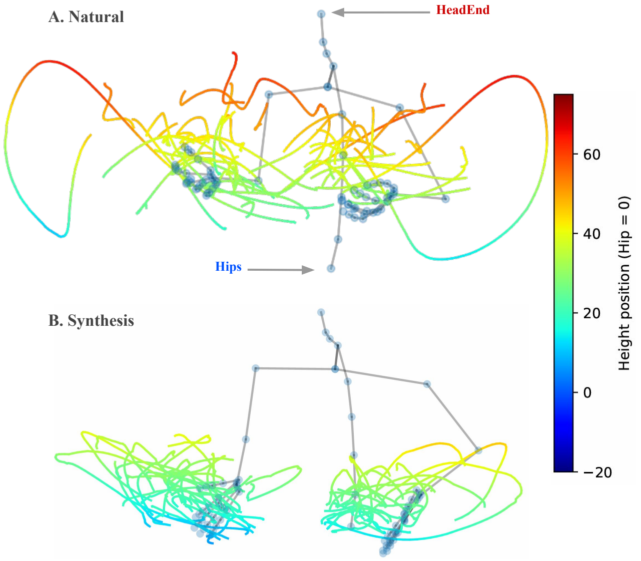

Figure 1: Motion during annotated activites.

a lower profile, with less contrast in positional changes compared to the natural gestures. Both the differences in trajectory patterns and in distance variation highlight the expansive and dynamic nature of natural gesture space compared to the confined and rigid movement profile observed in the synthesis data.

3.3.3 Velocity. We can also observe in Fig.2, the scalar velocity values of the right index fingertip motion over time in frames for both natural and synthesis data, expressed in cm/s. For the natural data, the velocity ranges from 0 to a maximum of around 300 cm/s (mean = 38.84 cm/s, median = 22.88 cm/s). This indicates that the majority of the movement is relatively slow, with some clear peaks of higher speed over time. They generally correspond to peaks in distance, reflecting the dynamics of the motion, and then appear to globally correspond with temporal annotations of strokes, suggesting a clear pattern of position changes in the natural data. The relatively low average and median values further support the notion of subtle, slow movements with occasional bursts of faster motion. The upward deviation of the mean is likely influenced by these velocity peaks that align with the movement, which temporarily increase the speed during some of them. In contrast, for the synthesis data, the velocity shows a narrower range, with values from 0 to 100 cm/s (mean = 26.68 cm/s, median = 22.98 cm/s). The profile appears more erratic in the synthesis data, with frequent and less pronounced changes in speed. While movements are generally more continuous, this constant activity reduces the prominence of peaks in motion. Additionally, these peaks tend to be less pronounced, with lower values compared to natural data.

4 DISCUSSION

Striking differences emerged between natural and synthetic gestures when examined through the lens of expert annotations conducted using the same protocol. These differences relate to gesture identifications (formal aspects) and the determination of discourse contributions (communicative aspects). In the natural data, potential gestures were easily identified, segmented and largely recognized as effectively communicative. Gesture roles were determined in majority of cases (63%) with high certainty scores and mostly linked to specific discourse elements. In contrast, interpretability issues emerged for synthetic data, as early as the formal level, with segmentation discrepancies among the annotators. One-third of the annotations showed disagreement, with annotators perceiving different gesture's core position. These movements were also questioned on their genuine communicative role (25%) and when communicative, on their role in discourse (1/2 undefined, 1/3 unclear). Synthesized data produced four movements that were annotated as gestures with link to discourse elements, but interpretation had a very low certainty score.

While some interpretability issues in the synthesis data can be attributed to the absence of finger motion—since it is a crucial feature for disambiguating functional gesture categories—this limitation mainly affected the ability to determine the specific communicative contribution of the generated gestures. This may therefore partly explain the higher proportion of "undefined" gestures in synthesis data, but it had less impact on detecting potential gestures per se. Differences in gesture interpretability are more closely linked to kinematic and spatial factors. Synthetic movements suffer from speed-position variation peaks and dynamism, making it harder to identify salient elements within the constant and relatively subtle gesticulations produced by the model. Overall, in synthesis data, it remains difficult to determine with certainty whether the movements interpreted as gestures are not merely coincidental. Indeed, annotators naturally sought to connect movements to the accompanying discourse but the communicative status attributed to a synthetic movement could either reflect their primary influence by the verbal context, or genuinely result from the system producing movements in response to this verbal input. For instance, annotators might be tempted to interpret movements as illustrating a verbal enumeration —a *Parsing*— given their potential coincidence with listed elements in discourse.

As expected, expert annotation comparisons of natural and synthetic data provided valuable insights into the salient characteristics absent or poorly reproduced by the sample gesture generative model. By examining differences between the properties of gestures produced in both contexts, it became possible to identify major elements that characterize human co-verbal gestures. Aspects such as pattern diversity, velocity increase, spatial variation, finger micro-movements, and connection to discourse are thus revealed to be essential. While these results are based on the study of a single generation model and a single speaker —reflecting the exploratory aim of this study, which is not to generalize the findings but to offer avenues for reflection on how to assess the effectiveness and credibility of synthetic gestures—, they nonetheless confirm that expert annotations can serve as an effective means of identifying aspects where synthetic gestures may show limitations. As such, they can serve as key criteria in establishing a reference framework for both a more objective evaluation of a model's output quality and cross-model comparison, as both currently suffer from issues of objectivity and a lack of standardization [28].

Indeed, while the comparison conducted in this small exploratory study allowed us to identify some initial concrete limitations of

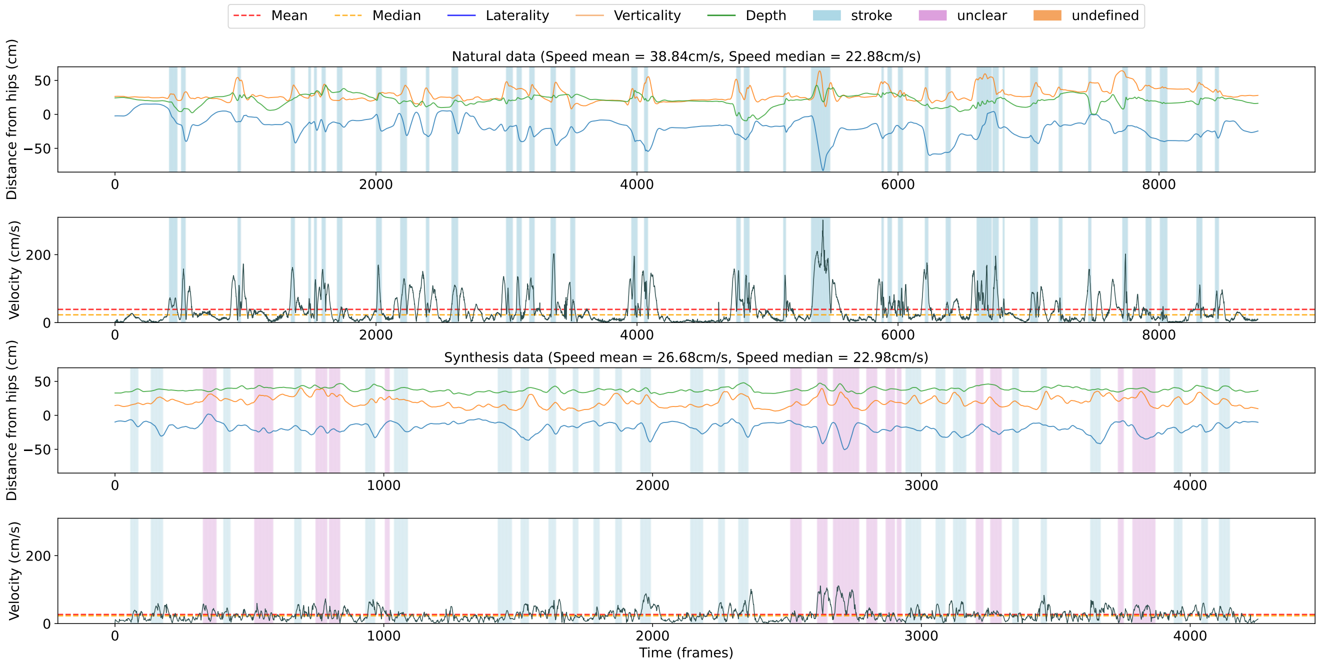

Figure 2: Analysis of the velocities (cm/s) and distance from hips (cm) of right index fingertip motion.

synthetic gestures, the features that characterize high-quality gestures should be further refined and expanded through a larger-scale study. Such a study would, at minimum, involve expert annotation of a broader set of natural motion-capture data, including analysis of the dynamic and communicative properties of gestures and their alignment with discourse at both lexical and prosodic levels. The aspects identified could then serve as reference standards, providing objective benchmarks for evaluating synthetic gestures. Then, fundamental aspects aspects identified by a broader study—such as increases in velocity or spatial variations, as highlighted in this paper—could form distinct categories within objective metrics for evaluating human-likeness. Complementarily, the speech appropriateness of gestures could be assessed through the ability of trained annotators to recognize a communicative contribution in the generated movements, to measure their diversity, and to identify their anchoring in discourse.

In this sense, the study lays the groundwork for developing new, objective means of evaluation that may eventually replace the currently used subjective assessments—still recently regarded as the gold standard [28] —but which, as their name suggests, remain inherently limited by their subjective nature.

5 CONCLUSION

Multimodal annotations are time-consuming task, but they are an essential resource for understanding the mechanisms of speech-gesture articulation. While a model may replicate movements resembling human gestures, this does not guarantee appropriate placement or motion, which could hinder meaning conveyance. To address these limitations, the Syncogest project [7] aims to guide the model's learning towards annotated gesture segments and their associated features. Incorporating salient communicative motion should enable the model to better identify patterns, distinguish them from surrounding noise, and determine their functional relationship with discourse, thereby improving generation.

Our analysis, the first of its kind to attempt to identify objective evaluation criteria for generated gestures through a comparison of natural and synthetic gestures annotated by expert, indeed reveals specific limitations of synthetic gestures. Its primary aim is to highlight the crucial importance of interdisciplinary collaboration in gesture research —at the intersection of gesture studies, computational modeling, and annotation practices. By emphasizing this necessity, our work advocates for a more fine-grained and in-depth evaluation of gesture synthesis that extends beyond surface-level perceptual assessments. Only by taking into account the specific characteristics of human gesturality can we objectively assess 'to what extent gestures visually look like something a human might produce' (*human-likeness*) and 'quantify the link between the gestures and the speech' (*speech appropriatness*). Our findings thus lay the groundwork for an interdisciplinary methodology aimed at providing a foundation for future advances in gesture research, establishing a benchmark for model cross-evaluation, and guiding future improvements in computational modeling by highlighting fundamental features and unmet targets.

ACKNOWLEDGMENTS

This project was financially supported by the CNRS through the MITI interdisciplinary programs as part of its exploratory research initiatives, and by the French National Research Agency (ANR) under project ANR-24-CE23-2926-03.

REFERENCES

[1] 2024. *ELAN (Version 6.8) [Computer software].* Nijmegen. https://archive.mpi.nl/tla/elan

[2] L. Abel. defended 6th February 2025. *Expressive audio-visual speech synthesis in an interaction context.* Ph.D. dissertation. University of Lorraine, Loria, Multispeech INRIA.

[3] Louis Abel, Vincent Colotte, and Slim Ouni. 2024. Towards interpretable co-speech gestures synthesis using STARGATE. In *International Conference on Multimodal Interaction (ICMI Companion'24: GENEA Workshop).*

[4] Louis Abel, Vincent Colotte, and Slim Ouni. 2024. Towards realtime co-speech gestures synthesis using STARGATE. In *25th Interspeech Conference (INTERSPEECH 2024).*

[5] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. *Computer Graphics Forum* 39, 2 (May 2020), 487–496. https://doi.org/10.1111/cgf.13946

[6] Paul Boersma and David Weenink. 2024. *Praat: doing phonetics by computer.* Computer program. http://www.praat.org/.

[7] Domitille Caillat, Ludovic Marin, Christelle Dodane, Fabrice Hirsch, Slim Ouni, Pierre Slangen, Patrice Guyot, Vincent Colotte, Aliyah Morgenstern, Louis Abel, Mickaëlla Grondin-Verdon, and Juliette Lozano Goupil. 2022. Synchronization of speech and gestures in an interactional context (SyncoGest Project). In *ISGS 2022 - 9th Conference of the International Society for Gesture Studies.* Chicago, United States. https://imt-mines-ales.hal.science/hal-03875218

[8] Geneviève Calbris. 2011. *Elements of Meaning in Gesture.* John Benjamins Publishing Company. https://doi.org/10.1075/gs.5

[9] Justine Cassell. 2000. Embodied conversational interface agents. *Commun. ACM* 43, 4 (April 2000), 70–78. https://doi.org/10.1145/332051.332075

[10] Justine Cassell. 2001. Embodied Conversational Agents: Representation and Intelligence in User Interfaces. *AI Magazine* 22, 4 (Dec. 2001), 67. https://doi.org/10.1609/aimag.v22i4.1593

[11] Jean-Marc Colletta, Olga Capirci, Carla Cristilli, Susan Goldin-Meadow, Michèle Guidetti, et al. 2011. *Manuel de codage : Transcription et annotation de données multimodales sous ELAN.* Technical Report. Université Stendhal-Grenoble III. https://hal.archives-ouvertes.fr/hal-04397224 ⟨hal-04397224⟩.

[12] Jacques Cosnier, Pierre Coulon, Claude Berrendonner, and Henri Orecchioni. 1982. Communications et langages gestuels. In *Les voies du langage, communications verbales, gestuelles et animales.* Dunod, Paris, 255–304.

[13] Jacques Cosnier and Jacqueline Vaysse. 1997. S'emiotique des gestes communicatifs. *Nouveaux actes s'emiotiques* 52 (1997), 7–28.

[14] Nicole Dargue, Naomi Sweller, and Michael P. Jones. 2019. When our hands help us understand: A meta-analysis into the effects of gesture on comprehension. *Psychological Bulletin* 145, 8 (Aug. 2019), 765–784. https://doi.org/10.1037/bul0000202

[15] Anna Deichler, Shivam Mehta, Simon Alexanderson, and Jonas Beskow. 2023. Diffusion-Based Co-Speech Gesture Generation Using Joint Text and Audio Representation. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23).* ACM. https://doi.org/10.1145/3577190.3616117

[16] Gaelle Ferre. 2011. Annotation multimodale du français parlé. Le cas des pointages. In *Proceedings of TALN - Atelier Degels.* Montpellier, France, 29–43. https://hal.science/hal-00609128

[17] Mickaëlla Grondin-Verdon, Domitille Caillat, and Slim Ouni. 2024. Qualitative study of gesture annotation corpus : Challenges and perspectives. In *ICMI Companion '24: Companion Proceedings of the 26th International Conference on Multimodal Interaction.* ACM, San Jose, Costa Rica, 147–155. https://doi.org/10.1145/3686215.3688820

[18] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of Speech-to-Gesture Generation Using Bi-Directional LSTM Network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents (IVA '18).* ACM. https://doi.org/10.1145/3267851.3267878

[19] Spencer D. Kelly, Aslı Özyürek, and Eric Maris. 2009. Two Sides of the Same Coin: Speech and Gesture Mutually Interact to Enhance Comprehension. *Psychological Science* 21, 2 (Dec. 2009), 260–267. https://doi.org/10.1177/0956797609357327

[20] Adam Kendon. 1980. *Gesture and Speech.* Cambridge University Press, Cambridge.

[21] Adam Kendon. 2004. *Gesture: Visible Action as Utterance.* Cambridge University Press, Cambridge ; New York.

[22] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing Input and Output Representations for Speech-Driven Gesture Generation. (2019). https://doi.org/10.48550/ARXIV.1903.03369

[23] Taras Kucherenko, Rajmund Nagy, Youngwoo Yoon, Jieyeon Woo, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2023. 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* (Paris, France) *(ICMI '23).* Association for Computing Machinery, New York, NY, USA, 792–801. https://doi.org/10.1145/3577190.3616120

[24] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022. BEAT: A Large-Scale Semantic and Emotional Multi-Modal Dataset for Conversational Gestures Synthesis. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII.* Springer, 612–630.

[25] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi.. In *Interspeech*, Vol. 2017. 498–502.

[26] David McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought.* The University of Chicago Press, Chicago and London.

[27] David McNeill. 2005. *Gesture and Thought.* The University of Chicago Press, Chicago and London.

[28] Rajmund Nagy, Hendric Voss, Youngwoo Yoon, Taras Kucherenko, Teodor Nikolov, Thanh Hoang-Minh, Rachel McDonnell, Stefan Kopp, Michael Neff, and Gustav Eje Henter. 2024. Towards a GENEA Leaderboard – an Extended, Living Benchmark for Evaluating and Advancing Conversational Motion Synthesis. https://doi.org/10.48550/ARXIV.2410.06327

[29] Asli Özyürek. 2018. Role of Gesture in Language Processing: Toward a unified account for production and comprehension. , 591–607 pages. https://doi.org/10.1093/oxfordhb/9780198786825.013.25

[30] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding.* IEEE Signal Processing Society.

[31] Emanuel A. Schegloff. 1985. *On some gestures' relation to talk.* Cambridge University Press, 266–296.

[32] Kanglei Zhou, Zhiyuan Cheng, Hubert PH Shum, Frederick WB Li, and Xiaohui Liang. 2021. Stgae: Spatial-temporal graph auto-encoder for hand motion denoising. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR).* IEEE, 41–49.

Received 11 July 2025; revised August, 1th; accepted August, 15th 2025