

# A Benchmark and Challenge for Scene-Aware Referential Gesture Generation

Anna Deichler  
KTH Royal Institute of Technology  
Sweden

Rishabh Dabral  
Max Planck Institute for Informatics  
Germany

Fethiye Irmak Dogan  
University of Cambridge  
United Kingdom

Anindita Ghosh  
Max Planck Institute for Informatics  
Germany

Jonas Beskow  
KTH Royal Institute of Technology  
Sweden

## Abstract

Referential gestures, pointing, indicating, and orienting the body toward objects in shared space, are fundamental to embodied communication. For virtual agents and physical robots operating in human environments, the ability to generate spatially grounded gestures is essential for disambiguation, instruction, and collaborative interaction. Yet, research on communicative gesture generation has largely focused on co-speech beat and iconic gestures, trained on corpora in which spatial grounding is absent or incidental. This lack of active research on referential gestures can be attributed to three key factors: datasets that pair gestures with 3D scene context are scarce, referential gesture generation lacks task formulation, and metrics for evaluating spatial grounding do not exist. In this work, we address all three gaps by introducing the MM-Conv Referential Gesture Generation Challenge. Specifically, the benchmark consists of three components: (i) a paired data release of 3,000 pointing-annotated clips from MM-Conv and SGS-HSI, with pointing-quality annotations and scene-disjoint splits; (ii) a task formulation that requires systems to produce spatially grounded reference gestures aligned with speech, without oracle apex timing or motion templates; (iii) a spatio-temporal evaluation protocol decomposing referential gesture quality into temporal alignment, spatial accuracy, and referent recall. We present a modular baseline based on OmniControl and position the benchmark as the foundation for the scene-aware gesture generation challenge at the 1st Workshop on Human–Scene Interaction at ECCV 2026. We envision this challenge as a testbed for the next generation of referential gesture synthesis works.

## CCS Concepts

• **Computing methodologies** → **Animation**; *Computer vision*.

## Keywords

gesture generation, human–scene interaction, referential grounding, co-speech motion, benchmark

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*HSI '26, Malmö, Sweden*

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-XXXX-X/2026/10

## 1 Introduction

Embodied agents operating in shared environments, virtual avatars, social robots or AR assistants must produce communicative behavior that is grounded in physical space. When a person says “put the cup on *that* one” while pointing at a table, the meaning of the utterance lives jointly in the speech, the speaker’s body and the physical space. A gesture generation model that reproduces realistic motion but lacks spatial grounding to produce referential gestures is missing a capability central for embodied communication: indicating something about the scene.

Current co-speech gesture generation mostly focuses on the distribution-matching problem in motion synthesis: models are trained to reproduce the statistical properties of human motion, and evaluated accordingly with metrics such as Fréchet Gesture Distance [24] and diversity, as established by the GENE Challenge [13] for community-wide benchmarking of co-speech gesture generation. For beat and iconic gestures, where the goal is natural-looking motion synchronized with speech, this is appropriate. Referential gestures pose a fundamentally different problem. A pointing gesture directed at the wrong object can score well under distributional metrics while being a complete communicative failure. This mirrors a broader challenge in motion generation: in reinforcement learning approaches like Adversarial Motion Priors, a discriminator enforces distributional realism while task-specific rewards impose grounding constraints; gesture generation lacks the latter. Evaluating referential gestures therefore requires complementing distributional fidelity with task success: did the gesture indicate the intended object?

The shift has been blocked by three missing pieces: datasets pairing human gesture with 3D scene context, a task formulation for scene-aware referential gesture generation, and metrics that measure spatial grounding in gestures. This paper addresses all three. We build on MM-Conv [8], whose referring expressions are each annotated with a ground-truth referent in a 3D scene graph, thereby turning “does the motion look plausible” into “would a listener find the right object?”

The benchmark comprises three core components:

- (1) A **paired data release**: approximately 2K pointing-annotated clips from MM-Conv’s naturalistic dialogue, and 1K clean single-target pointing clips from SGS-HSI with shared actors (§3).
- (2) A **task formulation** treating referential gesture as an end-to-end problem: systems receive speech, text, and scene context, and must determine when to gesture, where to point,

and how to move, without oracle timing or spatial targets (§4.1).

- (3) A **spatio-temporal evaluation protocol** decomposing referential gesture quality into three axes (§4.3):
- *Axis 1a, temporal alignment.* Did the gesture peak at the communicatively appropriate moment?
  - *Axis 1b, spatial accuracy.* At the system’s chosen moment, did the pose point at the right place?
  - *Axis 2, referent recall.* Did the pose uniquely indicate the intended object among scene-graph distractors?

The product of Axes 1a and 1b yields a combined spatio-temporal grounding (STG) score that prevents accuracy inflation: a gesture that accidentally passes through the correct pointing vector at the wrong time scores near zero. Furthermore, submissions are evaluated through both objective metrics (eg. FGD) and perceptual studies rating motion quality and semantic appropriateness.

We present a reference baseline (§4.4): a modular OmniControl-based pipeline that uses a learned IK module and phrase-anchored timing. The benchmark is designed to extend to more challenging scene-grounding conditions in future iterations (§5).

## 2 Related Work

*Co-speech gesture generation.* Recent work has driven rapid progress on speech-conditioned motion synthesis, with diffusion [2, 7] and flow-matching models [15] trained on corpora such as BEAT2 [14] producing increasingly natural co-speech gestures. Progress has focused on semantic and prosodic alignment, producing gestures that match the rhythm and content of speech, while spatial grounding in the surrounding environment remains largely unaddressed.

*Scene-aware human motion.* Human–scene interaction (HSI) and human–object interaction (HOI) have received significant attention [19, 20, 23, 25], with a focus on contact, affordance, and physically plausible motion. Wang et al. [21] use a dense skeleton-to-scene distance field to guide language-conditioned motion generation. Ghosh et al. [10] synthesize scene-aware motion through geometry-grounded token planning, demonstrating effective spatial reasoning for 3D motion generation in indoor environments. These approaches ground motion in physical contact and affordance but do not address the communicative function of gestures in scenes.

*Referential grounding.* Referring expression [1, 3, 4] and instruction following [18] benchmarks have established the task of linking language to scene objects. MM-Conv [8] extended this to spontaneous multimodal dialogue with synchronized motion and gaze. Doğan et al. [9] address referential ambiguity in human-robot interaction through semantic disambiguation strategies. 3D Embodied Reference Understanding tasks [16, 17] combine pointing gestures with language for object identification in ScanNet [5] scenes, and show that gesture substantially improves grounding accuracy. We address the inverse problem: rather than recovering the referent from an observed gesture, we evaluate whether generated pointing gestures indicate the intended object.

## 3 Corpus and Data Sources

The benchmark release comprises two paired data sources that share actors and capture conditions but differ in scene structure.

*MM-Conv* [8] is the primary evaluation data: approximately 2,000 pointing gesture ranges drawn from 6.7 hours of dyadic VR interaction across five AI2-THOR [12] indoor rooms, each centered on a referring expression accompanied by a detectable pointing gesture. Each clip is annotated with word-level speech transcripts, full-body SMPL-X motion capture at 30 fps, the ground-truth referent, the dominant pointing hand, and the full scene graph (object identities, 3D bounding boxes, and positions; typically ~50 candidate objects per scene). Expressions span exact noun phrases, partitive/attribute noun phrases, and pronominals (“that one,” “this one”), reflecting the natural distribution in spontaneous dialogue.

*SGS-HSI* is a paired synthetic resource: 1,138 clips recorded with pointing motion directed at 3D targets, retargeted to the MM-Conv actors, with referential expressions synthesized with the actors’ voices. Each clip has a single annotated target coordinate and apex frame, but no surrounding scene graph. SGS-HSI provides clean supervision for the geometric primitive of pointing at an arbitrary 3D coordinate, uncluttered by distractors or referential ambiguity.

All sources are available in SMPL-x format. Participants may use any subset, though the declaration of training data is mandatory.

*Spatial coverage.* MM-Conv contributes room-scale geometry: targets reach up to 5.86 m from the speaker (mean 3.02 m, 95th percentile 5.19 m). SGS-HSI concentrates in the close-to-mid-range volume (mean distance 2.20 m, max 2.85 m).

*Phrase metadata.* Referring-expression phrase intervals from the original MM-Conv corpus are released alongside the training data. These are not provided for the test set and are not consumed by the scorer.

## 4 Benchmark

### 4.1 Task Formulation

We define Track A in this release.

*Track A, target-aimed pointing.* Given (i) a spoken utterance with word-level timing, (ii) a 3D target coordinate, and (iii) the dominant pointing hand, generate SMPL-X motion containing a pointing gesture directed at the target. No apex frame is provided; the system must determine when the gesture peak occurs. Evaluable on both MM-Conv and SGS-HSI.

Track A serves as a basis for further releases in referential gesture generation.

### 4.2 Scene-Disjoint Splits

Evaluation uses a scene-disjoint split to test generalization to unseen environments. One room is held out as the test set; the remaining four rooms are released as training data. Participants may further subdivide the training rooms into train/validation splits as needed for their methods.

Table 1 summarizes the split statistics. The test room contains 339 MM-Conv referring expressions, with spatial and linguistic diversity comparable to the training rooms. By holding out an entire environment rather than random samples, the benchmark isolates the challenge of spatial generalization: models must ground references in novel room layouts with unseen object arrangements.

**Table 1: Scene-disjoint split statistics on MM-Conv and SGS-HSI.**

Split	Rooms	MM-Conv	SGS-HSI	Total
Train	4 rooms	1,503	1,022	2,525
Test	1 room	339	116	455

### 4.3 Evaluation Protocol: Spatio-Temporal Grounding

The protocol decomposes referential gesture quality into three independently measurable axes. All scores are bounded in  $[0, 1]$ . The released scorer computes all axes automatically from submitted SMPL-X motion; no apex frame or other metadata needs to be declared by participants.

**4.3.1 Axis 1a: Temporal alignment.** Axis 1a measures whether the system’s gesture peak coincides with the ground-truth (GT) gesture. The GT *hold region* is defined kinematically from the forearm extension profile of the original motion capture. A kinematic apex detector (logistic regression over pooled per-frame body features) identifies the gesture peak from the submitted motion. If the detected peak falls within the GT hold, the system receives full credit; peaks outside the hold are penalized with a Gaussian decay proportional to temporal distance. On GT motion, the detector achieves a mean  $\text{Score}_{1a}$  of 0.92, establishing the metric ceiling.

**4.3.2 Axis 1b: Spatial accuracy.** At the detected apex frame, we measure how accurately the forearm pointing ray is directed at the target. The scorer computes an extent-aware angular error that accounts for object size via the target’s bounding box, and converts it to a score via Gaussian decay ( $\sigma_{\text{angle}} = 20^\circ$ ). On GT motion, human pointing achieves  $\text{Score}_{1b} \approx 0.81$  with a median angular error of  $\sim 10^\circ$ .

**4.3.3 Axis 2: Referent recall.** Axis 2 measures whether the generated pose uniquely indicates the intended referent among scene-graph distractors ( $\sim 50$  objects per scene). At the detected apex, pose-derived angular features are used to rank all scene objects; we report top- $K$  recall. Challenge submissions are scored with a geometric Heuristic ranker (ray angle, head direction, distance; no learned parameters) to ensure full transparency. A learned ranker is reported alongside for analysis. Axis 2 applies to MM-Conv clips only; SGS-HSI lacks distractors.

**Combined Spatio-Temporal Grounding (STG).**  $\text{STG} = \text{Score}_{1a} \times \text{Score}_{1b}$ . This multiplicative combination prevents accuracy inflation: a gesture that accidentally passes through the correct pointing vector at the wrong time scores near zero, as does a gesture with perfect timing but poor spatial accuracy. Together with Axis 2, the three-axis decomposition exposes *where* a system fails, timing, spatial precision, or referent disambiguation, rather than collapsing performance into a single scalar.

### 4.4 Reference Baseline

We present OmniControl-PT, a fully modular baseline where timing and spatial placement are handled by separate components with no joint learning.

**Table 2: Spatio-temporal grounding on the test split (455 clips). Scores reflect kinematically detected apex frames for all methods.  $\text{STG} = \text{Score}_{1a} \times \text{Score}_{1b}$ . FGD is computed against EMAGE VAE.**

Condition	temporal $\uparrow$	spatial $\uparrow$	STG $\uparrow$	Med. $\theta_{\text{eff}} \downarrow$	FGD $\downarrow$
GT MoCap	0.909	0.766	0.718	$11.8^\circ$	0.961
OmniControl-PT	0.823	0.515	0.431	$23.4^\circ$	5.612

**Temporal: phrase-anchored heuristic.** The apex frame is predicted at a fixed offset before the end of the referring expression, calibrated on training data ( $\delta = 11$  frames,  $\approx 367$  ms at 30 fps) [11]. No learned parameters.

**Spatial: learned IK module.** A small MLP maps the 3D target coordinate and body shape parameters to predicted wrist and elbow positions at the apex frame, providing spatial control hints to the motion generator.

**Motion generation.** OmniControl [22], a diffusion-based model with joint-level spatial guidance, is fine-tuned on the benchmark training set following [6]. At inference, it receives the predicted control hints and generates full SMPL-X motion without consuming ground-truth data.

### 4.5 Results

Tables 2 and 3 report baseline results on the test split (455 clips). GT MoCap establishes the metric ceiling, scored by the same kinematic apex detector applied to all submissions; GT scores are below 1.0 because the detector does not perfectly recover annotated apex frames from kinematics alone. OmniControl-PT consumes no ground-truth motion or timing at inference and is scored identically to all submissions.

The baseline achieves  $\text{STG} = 0.431$  against a GT ceiling of 0.718. The gap decomposes across both axes: temporal alignment (0.823 vs. 0.909) reflects the phrase-anchored heuristic, which lands in the GT hold on roughly 63% of clips; spatial accuracy (0.515 vs. 0.766) reflects learned-*IK* imprecision compounded by OmniControl’s soft enforcement of spatial hints, yielding a median angular error of  $23.4^\circ$  compared to  $11.8^\circ$  for human pointing.

On referent recall (Table 3), human pointing recovers the intended referent in the top-10 candidates roughly two-thirds of the time against  $\sim 50$  scene objects—a  $3.4\times$  improvement over random. The baseline preserves about half this signal at top-5 and two-thirds at top-10, consistent with its  $\sim 2\times$  larger angular error. Both the temporal and spatial gaps are open targets for challenge participants. A second baseline will be reported at the workshop.

## 5 Challenge

The benchmark serves as the basis for the scene-aware gesture generation challenge at the 1st Workshop on Human–Scene Interaction at ECCV 2026.

**Phase 1: A12-THOR scenes (this release).** Participants receive MM-Conv training data from four rooms with full annotations, plus SGS-HSI as auxiliary data. One room is held out as the test set.

**Table 3: Axis 2: referent recall on the test split. Top- $K$  candidate recall over scene-graph distractors. Heuristic is a geometric ranker; Learned is a 73K-parameter MLP. All methods are evaluated at the kinematic-detector apex.**

Condition	Scorer	Top-5 $\uparrow$	Top-10 $\uparrow$
GT MoCap	Heuristic	44.1%	63.6%
	Learned	51.0%	68.7%
OmniControl-PT	Heuristic	21.6%	41.1%
	Learned	25.4%	45.2%

Submissions consist of generated SMPL-X motion, scored under FGD, Axis 1a, 1b, STG and scene-based Axis 2 recall. Axis 2 scoring uses the geometric Heuristic ranker (Table 3) to ensure full transparency and reproducibility. Participants may use any external data or pretrained models; all training sources must be declared with the submission. Submission format and detailed instructions are provided alongside the data release at <https://huggingface.co/mm-conv-scene-gesture>.

*Future extensions.* The current benchmark uses controlled AI2-THOR environments with canonical object arrangements. Future iterations may extend to more complex real-world scene layouts and integrate with referring expression grounding baselines for future track evaluation, testing generalization to cluttered environments, and automatic target inference.

*No single-scalar ranking.* The results will be reported for all axes separately. Pareto-dominant submissions are highlighted. Perceptual evaluation will be conducted during the challenge to contextualize objective scores and identify cases where motion scores well on metrics but lacks communicative appropriateness.

*Code and model release.* The OmniControl-PT baseline will be released at <https://huggingface.co/mm-conv-scene-gesture> alongside the data release to ensure reproducibility and lower barriers to participation. Submissions are scored server-side; the scorer is not publicly released.

## 6 Conclusion

Gesture generation evaluation has been dominated by distributional realism metrics because, until recently, the data required for other metrics has been unavailable. MM-Conv’s referential annotations close that gap. The spatio-temporal decomposition proposed here, separating *when* the system gestures (Axis 1a), *where* it points (Axis 1b), and *what* it indicates (Axis 2), turns those annotations into complementary scoring functions whose divergence is itself diagnostic of generator failure modes. By withholding oracle timing from submissions, the benchmark evaluates referential gesture generation rather than conditioned motion synthesis.

The OmniControl-PT baseline establishes a functional lower bound: STG = 0.431 (at  $\sigma_{\text{angle}} = 20^\circ$ ) against a GT ceiling of 0.718, with the gap split between timing (phrase-heuristic misses on naturalistic MM-Conv dialogue) and spatial accuracy (soft constraint enforcement and learned-IK imprecision). This decomposition is the benchmark’s primary contribution to participants: knowing *where* the gap is tells them where to direct their modeling effort.

The benchmark establishes a foundation for scene-aware gesture evaluation on controlled environments with ground-truth scene graphs. Future extensions to real-world scans will test transfer to unstructured geometry, connecting the gesture generation community with the broader 3D referential grounding literature.

*Limitations.* The extent-aware angle uses axis-aligned bounding boxes, which overestimate extent for elongated objects at oblique angles. Phrase-gesture association in MM-Conv is structurally noisier than in single-target setups. Phrase metadata is available for training only and may not align cleanly with clip-level gesture boundaries. The current evaluation focuses on controlled environments; extension to more complex real-world layouts will test robustness to cluttered scenes and challenging spatial configurations.

## References

- [1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. 2020. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European conference on computer vision*. Springer, 422–440.
- [2] Tenglong Ao, Zeyi Zhang, and Libin Liu. 2023. Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–18.
- [3] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. 2020. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*. Springer, 202–221.
- [4] Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, Song-Chun Zhu, Tao Gao, Yixin Zhu, and Siyuan Huang. 2021. Yourefit: Embodied reference understanding with language and gesture. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1385–1395.
- [5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5828–5839.
- [6] Anna Deichler et al. 2025. Grounded Gesture Generation: Language, Motion, and Space. *arXiv preprint arXiv:2507.04522* (2025). Presented at the Humanoid Agents Workshop, CVPR 2025.
- [7] Anna Deichler, Shivam Mehta, Simon Alexanderson, and Jonas Beskow. 2023. Diffusion-based co-speech gesture generation using joint text and audio representation. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 755–762.
- [8] Anna Deichler, Jim O’Regan, Fethiye Irmak Dogan, Lubos Marcinek, Anna Klezovich, Iolanda Leite, and Jonas Beskow. 2026. MM-Conv: A Multimodal Dataset and Benchmark for Context-Aware Grounding in 3D Dialogue. In *Proceedings of the Language Resources and Evaluation Conference*. LREC 2026.
- [9] Fethiye Irmak Dogan et al. 2025. A model-agnostic approach for semantically driven disambiguation in human-robot interaction. In *IEEE RO-MAN*. 649–656.
- [10] Anindita Ghosh et al. 2026. SceMoS: Scene-Aware 3D Human Motion Synthesis by Planning with Geometry-Grounded Tokens. In *CVPR*.
- [11] Adam Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press.
- [12] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Aniruddha Kembhavi, Abhinav Gupta, and Ali Farhadi. 2017. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv preprint arXiv:1712.05474* (2017).
- [13] Taras Kucherenko et al. 2023. The GENE Challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In *ICMI*. 792–801.
- [14] 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 (CVPR)*. 1144–1154.
- [15] Pinxin Liu, Luchuan Song, Junhua Huang, Haiyang Liu, and Chenliang Xu. 2025. Gesturelm: Latent shortcut based co-speech gesture generation with spatial-temporal modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10929–10939.
- [16] Ziyang Lu, Yunqiang Pei, Guoqing Wang, Peiwei Li, Yang Yang, Yinjie Lei, and Heng Tao Shen. 2024. Scanner: Interactive 3d visual grounding based on embodied reference understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 3936–3944.

- [17] Atharv Mahesh Mane, Dulanga Weerakoon, Vigneshwaran Subbaraju, Sougata Sen, Sanjay E Sarma, and Archan Misra. 2025. Ges3ViG: Incorporating Pointing Gestures into Language-Based 3D Visual Grounding for Embodied Reference Understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 9017–9026.
- [18] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 2017–2025.
- [19] Liang Pan, Zeshi Yang, Zhiyang Dou, Wenjia Wang, Buzhen Huang, Bo Dai, Taku Komura, and Jingbo Wang. 2025. Tokensi: Unified synthesis of physical human-scene interactions through task tokenization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 5379–5391.
- [20] Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. 2025. Tridi: Trilateral diffusion of 3d humans, objects, and interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5523–5535.
- [21] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. 2024. Move as you say interact as you can: Language-guided human motion generation with scene affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 433–444.
- [22] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. [n. d.]. OmniControl: Control Any Joint at Any Time for Human Motion Generation. In *The Twelfth International Conference on Learning Representations*.
- [23] Pradyumna Yalandur Muralidhar, Yuxuan Xue, Xianghui Xie, Margaret Kostyrko, and Gerard Pons-Moll. 2025. Physic: Physically plausible 3d human-scene interaction and contact from a single image. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*. 1–12.
- [24] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity. *ACM Transactions on Graphics* (2020). doi:10.1145/3414685.3417838
- [25] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Ilya A Petrov, Vladimir Guzov, Helisa Dharmo, Eduardo Pérez Pellitero, and Gerard Pons-Moll. 2025. Force: Dataset and method for intuitive physics guided human-object interaction. In *12th International Conference on 3D Vision*. IEEE.