Prompt-to-Gesture: Measuring the Capabilities of Image-to-Video Deictic Gesture Generation

FG 2026 KYOTO 京
May 25-29, 2026, Kyoto, Japan

Hassan Ali, Doreen Jirak, Luca Müller, Stefan Wermter
Knowledge Technology, Dept. Informatics, University of Hamburg, Germany
Behavioral Lab, Dept. Product Development, University of Antwerp, Belgium

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG
KNOWLEDGE TECHNOLOGY

Motivation

→ **Data scarcity** in gesture research.
→ **High cost** of real data collection
→ Emergence of video **Generative AI** models.
→ Can Generative AI models **augment** and **complement human-generated gestures?**

Contributions

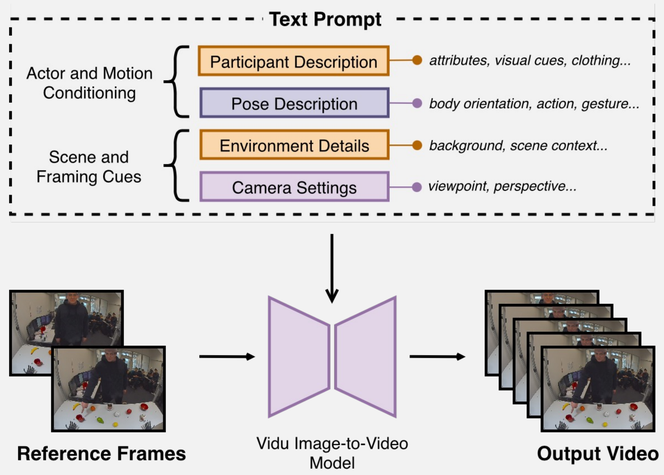
→ E2E pipeline to generate **photorealistic gestures** on a scale from minimal real data.
→ A novel **synthetic** deictic gesture **dataset**.
→ Rigorous evaluation across **visual fidelity**, **semantic alignment**, **motion plausibility**, and **downstream ML tasks**.

Key Takeaways

→ AI-Generated gestures show **high visual and semantic similarity** to the real data.
→ Synthetic data introduces **meaningful variability** beyond lab settings.
→ **Downstream task performance improves** when using hybrid data (synthetic + real).

Prompt-to-Gesture Pipeline

→ **Image-to-video pipeline** for generating deictic (pointing) gestures.
→ **Zero-shot** video synthesis (**Vidu** Diffusion Model).
→ **Reference frames** and structured **text prompts** to guide the generation.

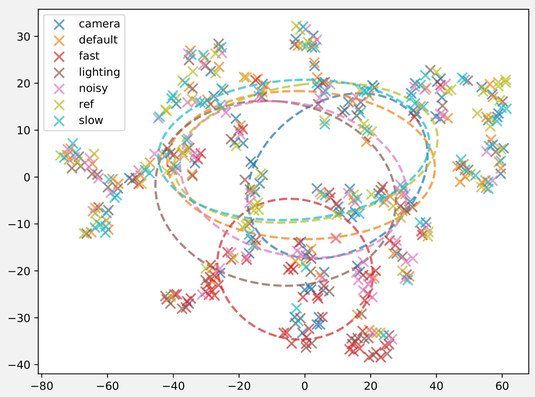


SCAN ME

Link to our paper, dataset and scripts!

Dataset Quality & Diversity

→ **24x expansion** in data size, easily **scalable**.
→ Synthetic gestures align with real data (↓ FID/FVD, ↑ CLIP similarity, akin motion derivatives, hand confidence, joint angles.)
→ Available to researchers without extensive ML experience.



Downstream Performance

→ **Baseline**: trained on real data only.
→ **Pre-training**: trained on synthetic data only.
→ **Fine-tuning**: synthetic data pre-training and real data refinement.

→ Hybrid approach with **mixed data** → best accuracy and F1-score across all models.
→ Synthetic-to-real knowledge transfer → **real-world (OOD) generalization** capability.

	Experiment	Training Data	Test Data	Accuracy	F-1 Score
CNNLSTM	**Baseline**	Real	Real	0.909	0.799
	Pre-training	Synthetic	Real	0.887	0.848
	Fine-tuning	Synthetic+Real	Real	**0.937**	**0.918**
MM-ITF	**Baseline**	Real	Real	0.891	0.892
	Pre-training	Synthetic	Real	0.872	0.874
	Fine-tuning	Synthetic+Real	Real	**0.897**	**0.898**
VideoMAE	**Baseline**	Real	Real	0.850	0.780
	Pre-training	Synthetic	Real	0.923	0.866
	Fine-tuning	Synthetic+Real	Real	**0.950**	**0.944**