CoherentHand: Temporally Consistent 3D Hand Trajectory Synthesis with Semantic Motion Priors –Supplementary Material

Bikram Boote[1] Junho Kim[1] Ozgur Kara[1] Sangmin Lee[2†] James M. Rehg[1†]
[1]University of Illinois Urbana-Champaign [2]Korea University

{boote, arkimjh, ozgurk2, jrehg}@illinois.edu sangmin-lee@korea.ac.kr

1. Additional Implementation Details

1.1. Network Details

For completeness, we provide a consolidated description of the architectural components used throughout CoherentHand, expanding upon the high-level overview presented in the main paper. Our system combines pre-extracted visual and linguistic features with intermediate VLM representations, which are subsequently processed by lightweight transformer modules and a DiT-based flow decoder. The visual features are extracted from DeiT [18] (dimension 768), and action prompt text features are obtained via the CLIP text encoder (dimension 512). For VLM-based guidance, we obtain intermediate representations from Qwen2.5-VL-7B [2], whose hidden dimension (3584) is projected to 512 using a single MLP projector. The resulting tokens (with optional zero-padding for fixed sequence lengths) are passed into a Perceiver Resampler [1], which compresses them into a fixed set of 30 latent tokens through a learnable latent querying mechanism.

The VLM-guided Residual Vector Quantization (RVQ) codebook utilizes a compact encoder-decoder structure to ensure stable codebook learning. Both the encoder and decoder are constructed from a single, 1-layer Transformer block featuring followed by a single attention head. The output features of the decoder are routed through two distinct MLP-based heads to recover the final reconstructed MANO parameters and its corresponding contact map. This entire RVQ module contains approximately 24.6M trainable parameters. Our flow-based codebook utilizes a DiT backbone [15]. This backbone is constructed from four sequential 1D DiT blocks. Within each DiT block, the Multi-Head Attention module is configured with 4 attention heads, and the block maintains a consistent hidden feature dimension of 512. This module contains roughly 36M trainable parameters and forms the core of our continuous trajectory generation pipeline.

†Corresponding author.

1.2. Training Details

We optimize the VLM-guided RVQ codebook using a combination of reconstruction objectives to supervise both hand articulation and contact dynamics. Specifically, we train the VLM-guided codebook with a Smooth L1 loss over the MANO hand pose and shape parameters, L1 loss to predict the contact centroid, and another L1 loss on the relative rotation and translation between the initial and future frames. For the contact map we adopt a standard binary cross entropy loss for the predictions. In addition, for our flow-based decoder, we trained the model following the rectified-flow formulation, where the objective is an L2 loss between the model's predicted velocity field and the ideal straight-line velocity. All models are trained on a single Nvidia RTX 4090 24GB GPU. The end-to-end training for the both codebook learning and the flow-based decoder optimization takes approximately 16 GPU hours, respectively.

2. Expansion of Related Works

To complement the Related Work section in the main manuscript, we provide an expanded discussion that situates CoherentHand within a broader landscape of recent developments in 3D hand–object interaction modeling. Although the manuscript focuses on works most directly related to image-conditioned trajectory generation and vector-quantized motion priors, several additional research directions have emerged that further contextualize our approach. These include human pose modeling works using LM-guided priors [3, 4, 8, 11] or flow-based pose sequence generators [6, 19]. In addition, to facilitate the development of highly robust and accurate models, multiple large-scale datasets have been proposed, from lab setting [7, 12–14] to much more real-world scenarios [5, 9, 10]. The most closely related to our approach is concurrent work [19] that shares the goal of utilizing a VLM to reason about hand interaction trajectory generation, their implementation differs in ways that preclude a direct comparison. Their method replaces the input visual feature encoders to explicitly fuse

Table 1: Performance comparison on generalization setup with LatentAct diffusion variant (LatentActDiff [16]). We report MPJPE (cm), MPJPE-PA (cm) & MPJPE-FA (cm) for the following setup: novel tasks, objects, actions & scene, to evaluate the accuracies of the predicted trajectories.

Method	**Task-level**		**Object-level**		**Action-level**		**Scene-level**	
	M-PE↓	M-PA↓	M-PE↓	M-PA↓	M-PE↓	M-PA↓	M-PE↓	M-PA↓
Hand Visible								
Forecasting								
LatentActDiff [16]	7.90	3.03	8.49	3.20	8.29	3.34	8.32	3.10
CoherentHand	6.76	2.72	7.09	2.75	6.88	2.82	7.61	2.88
Interpolation								
LatentActDiff [16]	6.53	2.62	7.23	2.81	7.11	3.10	6.70	2.84
CoherentHand	5.91	2.61	6.03	2.64	5.67	2.75	6.55	2.75
No Hand								
Forecasting								
LatentActDiff [16]	7.82	2.89	8.63	3.06	8.19	3.15	7.93	2.96
CoherentHand	7.23	2.60	7.73	2.78	7.53	2.85	8.11	2.85
Interpolation								
LatentActDiff [16]	6.60	2.65	7.73	2.78	7.42	3.04	6.89	2.71
CoherentHand	6.48	2.68	6.57	2.63	6.39	2.73	7.03	2.69

depth and RGB streams. This fundamental change mandates a data and compute-intensive training schedule involving additional finetuning of these encoders, unlike our approach which relies solely on querying the VLM with an image and text prompt for intermediate features. Furthermore, their model is trained on a substantially larger, distinct dataset than HoloAssist and predicts a larger state space (109 MANO parameters vs. our 61). Consequently, due to these major differences and the lack of officially released code or training setup, a meaningful, apples-to-apples comparison between our performance and its results cannot be established.

3. Additional Results

3.1. Comparison with LatentAct-Diffusion

Prakash *et al.* [16] also has trained a diffusion variant of their final hand trajectory predictor, namely LatentActDiff. LatentActDiff performs denoising in the codebook latent space, using an iterative denoising process model from MDM [17], which is followed by a decoder to generate the trajectories. As summarized in Tab. 1, we compare how our method performs against LatentActDiff in a more detailed manner. We consistently outperform LatentActDiff on both MPJPE and MPJPE-PA across all of the generalization setting, hence providing further validation on both the use of a VLM-guided codebook and a flow-based decoder that can directly denoise in the hand trajectory space, with the codebook latents as a conditioning signal.

3.2. VLM-guidance Increases Codebook Perplexity

We further investigate the effectiveness of our VLM-guided RVQ training in increasing codebook utilization, which in turn suggests the capture of fine-grained hand articulation details. To do so, we have utilized perplexity score as a metric, defined as the exponentiated Shannon entropy of the codebook distribution, quantifying the *effective* number of discrete codes currently being used by the model. A value close to the total codebook size indicates the discrete representation is under optimal utilization to embed motion priors. We verify clear advantage of using additional VLM guidance related to affordance and hand articulation cues during the RVQ codebook training improves utilization of the learned codebook as represented by a 30 point increase (from 340 without guidance to 370 with VLM-guidance) in the perplexity metric.

3.3. Additional Visualization Results

We qualitatively compare CoherentHand against LatentAct by visualizing their predicted trajectories overlaid on the underlying video frames, as shown in Fig. 1. CoherentHand demonstrates significantly superior temporal consis-

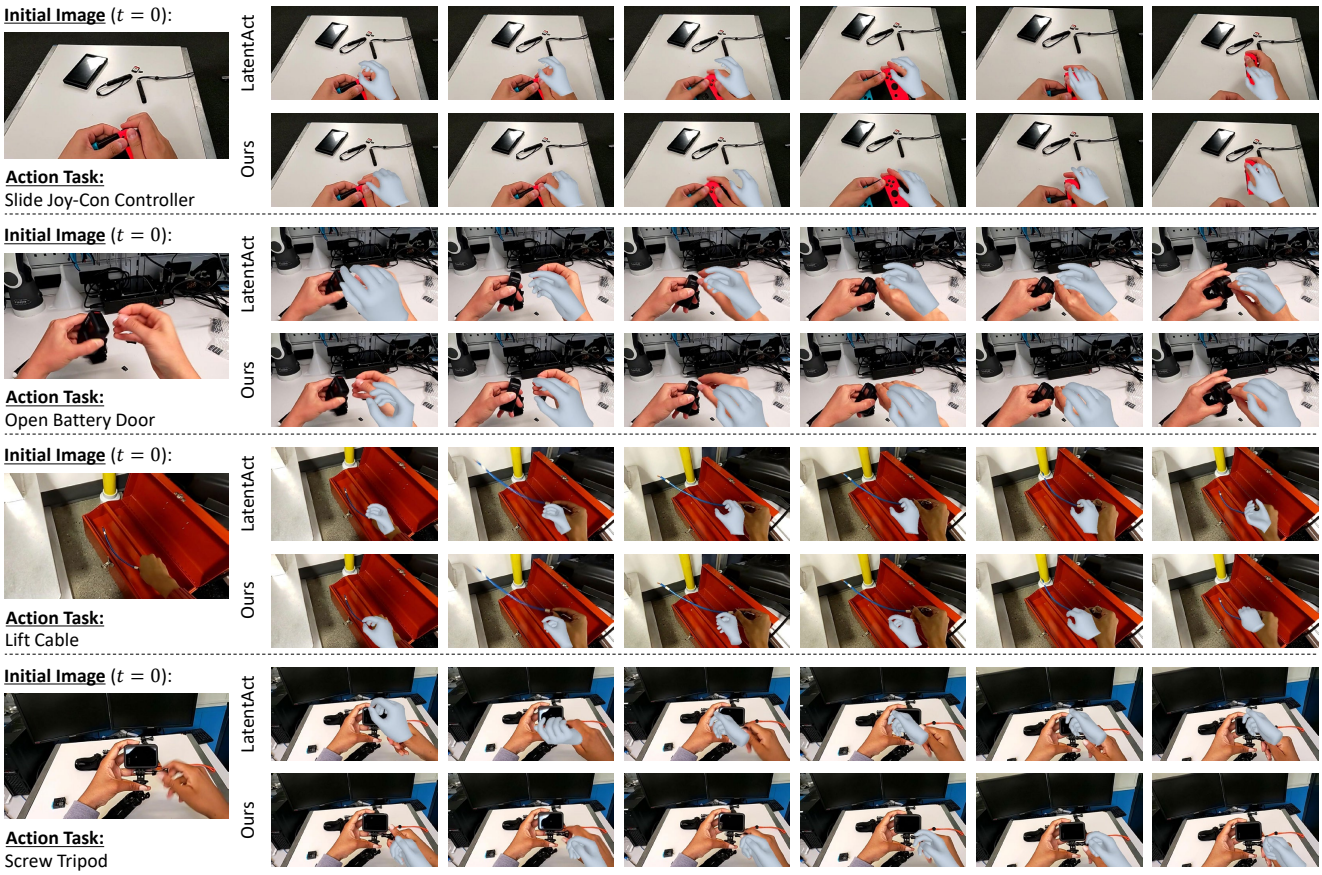

Figure 1: **Qualitative results:** We compare hand pose generations of CoherentHand with LatentAct [16] from different generalization settings on the forecasting task, at over the sequence. We show hand poses overlayed on the underlying video frames. Our method shows better fidelity in the hand poses compared to the baseline.

tency and fluidity, producing trajectories free from the erratic rotations and jerky movements characteristic of LatentAct's transformer-based decoder. Furthermore, the impact of our VLM-guided RVQ codebook—which facilitates the learning of more fine-grained hand articulation details—is evident in the "Screw Tripod" action (last row). While LatentAct only generates generic hand rotational motion, our method captures the subtle, screw-like manipulation involving the index finger and thumb.

4. Discussion and Limitation

Our experiments and ablation studies show how to effectively utilize intermediate VLM representation to provide high-level affordance and articulation cues to improve the fidelity of generated hand pose sequences. However, this does not overcome the geometric guidance having ground truth 3D object models can provide. Future work can focus on exploring avenues to incorporate additional 3D object-level cues to enhance this task. We also predict trajectories only for the right-hand and extending our setting to bimanual interaction trajectories can be an obvious next step.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[3] Chen Bao, Jiarui Xu, Xiaolong Wang, Abhinav Gupta, and Homanga Bharadhwaj. Handsonvlm: Vision-language models for hand-object interaction prediction. *arXiv preprint arXiv:2412.13187*, 2024. 1

[4] Junuk Cha and Jihyeon Kim. Cot-pose: Chain-of-thought reasoning for 3d pose generation from abstract prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4301–4309, 2025. 1

[5] Kefan Chen, Chaerin Min, Linguang Zhang, Shreyas Hampali, Cem Keskin, and Srinath Sridhar. Foundhand: Large-scale domain-specific learning for controllable hand image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17448–17460, 2025. 1

[6] Jungbin Cho, Junwan Kim, Jisoo Kim, Minseo Kim, Mingu Kang, Sungeun Hong, Tae-Hyun Oh, and Youngjae Yu. Discord: Discrete tokens to continuous motion via rectified flow decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14602–14612, 2025. 1

[7] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12943–12954, 2023. 1

[8] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. Chatpose: Chatting about 3d human pose. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2093–2103, 2024. 1

[9] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022. 1

[10] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 1

[11] Mingzhen Huang, Fu-Jen Chu, Bugra Tekin, Kevin J Liang, Haoyu Ma, Weiyao Wang, Xingyu Chen, Pierre Gleize, Hongfei Xue, Siwei Lyu, et al. Hoigpt: Learning long-sequence hand-object interaction with language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7136–7146, 2025. 1

[12] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10138–10148, 2021. 1

[13] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022.

[14] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21740–21751, 2024. 1

[15] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 1

[16] Aditya Prakash, Benjamin Lundell, Dmitry Andreychuk, David Forsyth, Saurabh Gupta, and Harpreet Sawhney. How do i do that? synthesizing 3d hand motion and contacts for everyday interactions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7026–7036, 2025. 2, 3

[17] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 2

[18] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 1

[19] Bohan Zhou, Yi Zhan, Zhongbin Zhang, and Zongqing Lu. Megohand: Multimodal egocentric hand-object interaction motion generation. *arXiv preprint arXiv:2505.16602*, 2025. 1