TRACEFORCING: LEARNING FROM PERSISTENT INTERACTION TRACES FOR FINE-GRAINED HAND-OBJECT INTERACTION VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained hand-object interaction (HOI) understanding is not merely about recognizing hands, objects, and actions in individual frames. It requires preserving a persistent interaction trace: who interacts with what, where contact occurs, and how these relations evolve into manipulation processes and state changes over time. However, existing approaches typically model entities, trajectories, or semantic relations as isolated structured cues, lacking a unified mechanism to preserve localized contact evidence and persistent hand-object identities throughout videos. In this work, we propose TraceForce-HOI, a unified framework that models HOI videos through Persistent Interaction Traces, a structured representation capturing hand entities, object instances, contact evidence, and interaction state evolution. To construct reliable traces, we introduce Fast-Slow Trace Identification, which combines dense interaction-aware perception with sparse temporal propagation to capture short-lived contact dynamics while maintaining consistent instance identities over time. The resulting traces are encoded in both pixel and latent spaces through HOI mask videos and instance-aware trace marks, exposing entity identities, active-object states, and localized interaction evidence to the model. Furthermore, we introduce causal identity and contact-attribution constraints to regularize intermediate visual representations, ensuring that interaction evidence remains anchored to the correct entities throughout the video. Extensive experiments demonstrate that TraceForce-HOI achieves significant improvements in fine-grained HOI video understanding, particularly on tasks requiring persistent instance association and interaction evolution modeling, including procedural understanding, state transition recognition, and object part-level comprehension.

1 INTRODUCTION

Hand-object interaction (HOI) is a fundamental way for humans to interact with the physical world and accomplish goal-oriented tasks Pei et al. (2025); Sun et al. (2026); Tateno et al. (2026a); Zhang et al. (2026d); Huang et al. (2025b); Dang et al. (2026a). Understanding fine-grained HOI processes from videos supports manipulation reasoning in embodied agents and intelligent assistance systems, while enabling automatic annotation and quality control of manipulation video datasets Sun et al. (2026); Xiao et al. (2026); Choi et al. (2026); Nagrani et al. (2026); Dang et al. (2026b); Chu et al. (2026); Xu et al. (2026). Unlike traditional object recognition or coarse-grained action recognition, fine-grained HOI video understanding requires not only identifying involved entities, but also localizing specific hand and object instances, capturing contact regions, modeling interaction evolution, and understanding part-level or state-level changes caused by manipulation Tateno et al. (2026a); Li et al. (2026b); Qi & Wu (2026); Ye et al. (2026); Zhu et al. (2026); Banerjee et al. (2025).

HOI understanding requires modeling dynamic interactions across different temporal scales Pei et al. (2025); Plizzari et al. (2025); Feng et al. (2026); Chen et al. (2025). Critical interaction cues are often concentrated in local parts of objects and small hand-object contact regions, where interaction establishment, maintenance, and release may occur within only a few consecutive frames Zhu et al. (2026); Tateno et al. (2026a); Li et al. (2026b). Meanwhile, complete manipulation processes usually span longer durations with fast motions, occlusions, and viewpoint changes Ye et al. (2026); Yoon et al. (2026); Lee et al. (2026a); Zhang et al. (2025); Banerjee et al. (2025). Therefore, models must

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

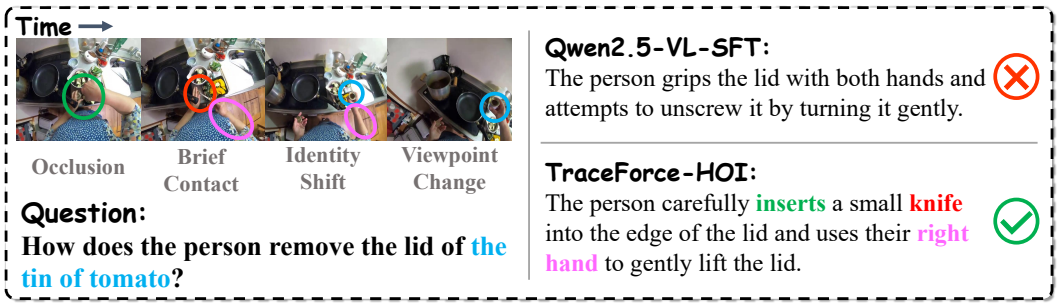


Figure 1: Motivation of TraceForce-HOI. Existing MLLMs often fail to bind transient contact cues with persistent hand-object identities. TraceForce-HOI explicitly models dynamic HOI, enabling accurate fine-grained interaction understanding under occlusion and viewpoint changes.

capture short-lived contact dynamics while continuously associating them with the correct hand and object instances. The key challenge is not simply detecting hands, objects, or contact regions, but establishing stable associations between evidence of transient interaction and persistent instance identities over time Feng et al. (2025a); Ye et al. (2026); Zhu et al. (2026); Xiao et al. (2026).

Recently, multimodal large language models (MLLMs) have achieved significant progress through large-scale pre-training, providing a strong foundation for open-world HOI understanding Gou et al. (2026); Tang et al. (2025); Zhu et al. (2025); Bai et al. (2025b;a); Zhang et al. (2026a). Existing methods further incorporate region localization Nagrani et al. (2026); Feng et al. (2025b); Wang et al. (2026a), hand geometry Choi et al. (2026); Ye et al. (2026); Tse et al. (2025), object trajectories Feng et al. (2025a); Ye et al. (2026), and local interaction dynamics Pei et al. (2025); Tateno et al. (2026a) to enhance entity awareness. However, these approaches usually model instance identity, motion trajectories, and interaction relationships separately, lacking a unified spatio-temporal representation that organizes contact evidence and interaction state evolution based on persistent instance identities Pei et al. (2025); Tateno et al. (2026b); Ye et al. (2026). Moreover, instance association and contact attribution from RGB features still largely rely on implicit reasoning Feng et al. (2025a); Nagrani et al. (2026); Tateno et al. (2026a). Fine-grained interaction structures may also be weakened during visual encoding, token compression, and cross-modal fusion, limiting fine-grained HOI understanding Du et al. (2026); Zhang et al. (2026b); Feng et al. (2026); Gou et al. (2026).

Based on these observations, we propose **TraceForce-HOI**, a unified framework that models HOI videos through **Persistent Interaction Traces**. Each trace organizes hand identities, persistent object instances, localized contact evidence, and interaction-state evolution. We encode this structure in both pixel and latent spaces: an HOI mask video exposes instance regions, trajectories, and contact dynamics, while instance-aware trace marks inject entity identities, active-object states, and localized interaction evidence into aligned RGB and mask dual-stream visual tokens.

To construct a reliable interaction trace with both short-term contact sensitivity and long-term identity consistency, we propose **Fast-Slow Trace Identification**. The Fast branch captures rapidly changing interaction and contact cues, whereas the Slow branch propagates instance masks and consolidates trajectories across sparse frames to maintain identity under motion and occlusion.

Building upon these constructed representations, we further propose **Contact-Aware Trace Forcing** to preserve instance associations and interaction structures in the model's high-level visual representations. It leverages generated instance associations and interaction constraints as structured supervision for intermediate visual representations. Specifically, Causal Track Alignment constrains the representations of the same object instance to remain consistent across time while improving separability from other instances. Contact Alignment further encourages the model to identify interaction-associated object instances and align their visual representations accordingly. These constraints guide the model toward learning stable instance identities and discriminative interaction semantics.

Extensive experiments demonstrate that TraceForce-HOI significantly improves fine-grained HOI understanding, especially on tasks requiring persistent interaction state evolution modeling, including procedural understanding, state transition recognition, and object part-level comprehension.

Our contributions are summarized as follows:

- We identify Persistent Interaction Traces as a missing representation for fine-grained HOI video understanding. Unlike object trajectories that mainly describe entity motions, interaction traces organize persistent hand-object identities, localized contact evidence, and evolving interaction states into a unified structured representation.
- We design Fast-Slow Trace Identification to automatically construct reliable traces. It combines dense semantic perception with sparse temporal propagation, capturing short-lived interaction dynamics while preserving identity consistency under motion and occlusion.
- We introduce Contact-Aware Trace Forcing, which uses automatically constructed trace structures as structured supervision to regularize high-level MLLM representations. Temporal identity and contact-attribution constraints guide the model to learn stable instance representations and discriminative interaction relationships.

2 Related Work

Hand-Object interaction understanding has been extensively studied in egocentric vision, including interaction-aware parsing, hand-object tracking, pose estimation, and 3D reconstruction SU et al. (2026); Tse et al. (2025); Zhang et al. (2025); Banerjee et al. (2025); Ye et al. (2026); Zhu et al. (2026). These approaches provide rich spatial and geometric descriptions of interacting hands and objects, supporting manipulation analysis. Recent works further explore fine-grained interaction dynamics and their consequences, including localized hand-object dynamics modeling for representation learning, video question answering for action and object-state understanding, and studies on mistake attribution and hand-object complementary cues Pei et al. (2025); Tateno et al. (2026a); Li et al. (2026b); Qi & Wu (2026); Tateno et al. (2026b). However, HOI reasoning requires associating transient contact evidence and evolving interaction states with persistent hand and object identities under motion, occlusion, and viewpoint changes. We explicitly model such contact-aware dynamic associations throughout manipulation processes.

Multimodal large language models have substantially advanced open-ended video understanding Cheng et al. (2024); Lin et al. (2024); Bai et al. (2025b;a); Zhu et al. (2025). Recent efforts extend these models to egocentric scenarios with localized observations, spatio-temporal cues, personalized grounding, gestures, and cross-view knowledge Feng et al. (2025b); Nagrani et al. (2026); Xiao et al. (2026); Choi et al. (2026); Wang et al. (2026a); Zhang et al. (2026a); Plizzari et al. (2025); Gou et al. (2026); Tang et al. (2025); Du et al. (2026); Zhang et al. (2026b); Feng et al. (2026); Guo et al. (2026). However, existing approaches typically treat interaction cues as separate hints or auxiliary inputs, leaving fine-grained contact-to-instance association largely reliant on implicit RGB reasoning. Recent forcing-based approaches further regularize video representations using explicit geometric targets or make intermediate representation prediction part of the modeling process Wu et al. (2026); Li et al. (2026a); Zhang et al. (2026c); Wang et al. (2026b). Our approach instead derives supervision from persistent HOI structures, explicitly preserving temporal identities and contact-to-object associations in intermediate MLLM representations.

Object-Centric temporal modeling represents videos and embodied scenes through entities and trajectories, enabling reasoning about motion, identity, and state evolution. Recent works tokenize object dynamics, recover persistent object tracks and tubes, or represent manipulation dynamics in compact 3D trace spaces for video understanding and embodied world modeling Feng et al. (2025a); Huang et al. (2025a); Banerjee et al. (2025); Ye et al. (2026); Zhu et al. (2026); Liu et al. (2026b); Lee et al. (2026c;b); Liu et al. (2026a). While these representations effectively capture object motion and geometric evolution, they do not explicitly organize persistent hand-object identities, localized contact evidence, and evolving interaction states for fine-grained video reasoning.

3 Method

3.1 Problem Formulation and Framework Overview

Given an HOI video $\mathcal{V} = \{\mathbf{I}_t\}_{t=1}^{T}$ and a natural-language question Q, the goal of fine-grained HOI video understanding is to generate an answer $Y = (y_1, \ldots, y_N)$ that correctly describes

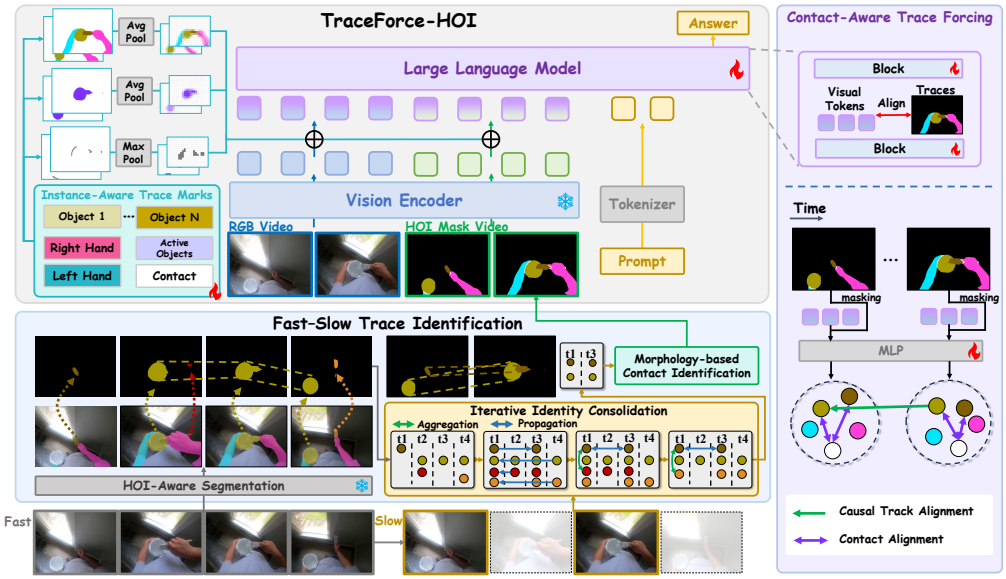


Figure 2: Overview of TraceForce-HOI. We introduce a Persistent Interaction Trace to explicitly model HOI evidence evolution. Fast-Slow Trace Identification constructs this representation through dense semantic extraction and sparse temporal identity propagation. Contact-Aware Trace Forcing further guides intermediate representations with causal track and contact alignment.

the interaction process depicted in the video. The model estimates the answer autoregressively as $p_\theta(Y \mid \mathcal{V}, Q) = \prod_{n=1}^{N} p_\theta\left(y_n \mid y_{<n}, \mathcal{V}, Q\right)$. As illustrated in Fig. 2, TraceForce-HOI centers on Persistent Interaction Traces, which explicitly encode persistent hand and object instances, active-object states, and localized contact evidence in both pixel and latent spaces. Fast-Slow Trace Identification automatically constructs this representation by combining dense interaction-aware perception with sparse temporal propagation. Contact-Aware Trace Forcing then uses the resulting interaction traces to regularize intermediate representations during training.

3.2 Persistent Interaction Traces

To explicitly associate transient interaction cues with temporally persistent hand and object instances throughout a video, we construct a unified representation $\mathcal{M} = \{\mathbf{H}_t^L, \mathbf{H}_t^R, \{\mathbf{O}_t^k\}_{k=1}^K, \mathbf{A}_t, \mathbf{C}_t\}_{t=1}^T$, where $\mathbf{H}_t^L$ and $\mathbf{H}_t^R$ denote the left- and right-hand masks, $\mathbf{O}_t^k$ denotes the k-th persistent object-instance mask, and $\mathbf{A}_t$ and $\mathbf{C}_t$ denote the active-object and contact masks, respectively.

3.2.1 HOI Mask Video

To explicitly expose instance trajectories and local interaction dynamics while reusing the strong visual priors of the MLLMs, we convert the representation $\mathcal{M}$ into an auxiliary pixel-space video stream, denoted by $\mathcal{V}^{\text{hoi}} = \{\mathbf{P}_t\}_{t=1}^T$ and termed the HOI Mask Video. Each frame $\mathbf{P}_t$ is rendered on a black background: the left and right hands are assigned predefined semantic colors, each object instance is assigned a distinct identity color that remains fixed across frames, and contact regions are highlighted using a dedicated color. We do not render the active-object mask $\mathbf{A}_t$, since changing the appearance of the same object between active and inactive states may weaken its identity consistency; instead, the active state is encoded by the latent interaction marks introduced below. The HOI Mask Video shares the same frame indices and spatial resolution as the original video $\mathcal{V}$ and is processed by the same encoder, allowing the structured HOI cues to directly benefit from its priors.

3.2.2 Instance-Aware Trace Marks

Although the HOI Mask Video explicitly presents interaction structures in pixel space, fine-grained instance identities and localized contact cues may still be weakened by visual patchification and

token compression. Moreover, the active-object state is not rendered in the HOI mask video to preserve object identity.

We introduce a shared learnable Mark Bank $\mathcal{E} = \{\mathbf{e}_L, \mathbf{e}_R, \mathbf{e}_{\text{obj}}, \{\mathbf{e}^k_{\text{slot}}\}_{k=1}^K, \mathbf{e}_{\text{active}}, \mathbf{e}_{\text{contact}}\}$ to encode hand identities, object instances, active-object states, and contact evidence. The mark of the k-th object instance is defined as $\mathbf{e}^k_{\text{obj}} = \mathbf{e}_{\text{obj}} + \mathbf{e}^k_{\text{slot}}$, where the shared object mark captures common object semantics and the slot mark distinguishes individual instances. Object tracks are randomly assigned to the available slots to avoid encoding a fixed track order.

We derive token-level mark weights from the masks in $\mathcal{M}$. At each pixel, the weights of overlapping hand, object, active-object, and contact masks are normalized to sum to one, preventing the injection magnitude from increasing with the number of activated masks. The normalized weights are then pooled onto the visual-token grid, using average pooling for hand, object, and active-object masks and max pooling for the sparse contact mask. Let w^L_i, w^R_i, $\{w^k_i\}_{k=1}^K$, w^{active}_i, and w^{contact}_i denote the resulting mask weights for the i-th visual token. Its mark embedding is computed as $\mathbf{m}_i = w^L_i \mathbf{e}_L + w^R_i \mathbf{e}_R + \sum_{k=1}^K w^k_i \mathbf{e}^k_{\text{obj}} + w^{\text{active}}_i \mathbf{e}_{\text{active}} + w^{\text{contact}}_i \mathbf{e}_{\text{contact}}$.

The RGB video $\mathcal{V}$ and HOI Mask Video $\mathcal{V}^{\text{hoi}}$ are separately encoded by a shared frozen visual encoder, yielding $\mathbf{Z}^{\text{rgb}} = E_{\text{vis}}(\mathcal{V}) \in \mathbb{R}^{N_v \times D}$ and $\mathbf{Z}^{\text{hoi}} = E_{\text{vis}}(\mathcal{V}^{\text{hoi}}) \in \mathbb{R}^{N_v \times D}$, where N_v and D denote the number and dimension of visual tokens, respectively. We then inject the same projected mark into the aligned RGB and HOI-mask tokens as $\widetilde{\mathbf{z}}^s_i = \mathbf{z}^s_i + \phi_{\text{mark}}(\mathbf{m}_i)$, where $s \in \{\text{rgb}, \text{hoi}\}$ and ϕ_{mark} is a shared linear projection.

3.3 Fast-Slow Trace Identification

Fine-grained HOI understanding requires capturing short-lived interaction cues while maintaining consistent instance identities over time. On a densely sampled timeline, InterFormer SU et al. (2026) predicts frame-wise semantic masks for the hands and objects interacting with either or both hands. The connected object regions are treated as frame-local candidates and associated across neighboring frames into coarse tracklets using lightweight overlap, motion, and appearance cues. This dense branch preserves transient contact and active-object observations without performing expensive mask propagation over the entire dense sequence.

For each coarse object tracklet, we select its earliest reliable observation as a seed and combine it with the MLLM-aligned frames to form a sparse propagation sequence. SAM3 Carion et al. (2026) then performs bidirectional mask propagation using the available InterFormer masks as prompts, and only the propagated masks at the MLLM-aligned frames are retained. The left and right hands are jointly propagated as two fixed identities. Duplicate object tubes are subsequently merged according to their spatio-temporal overlap and observation consistency, after which the most interaction-relevant K tracks are retained.

Finally, the active-object mask $\mathbf{A}_t$ is obtained by matching the frame-level interacted-object regions to the persistent object tubes through spatial overlap. The contact mask is independently computed at each frame as $\mathbf{C}_t = \bigcup_{k=1}^K \left[\text{Dilate}(\mathbf{H}^L_t \cup \mathbf{H}^R_t) \cap \text{Dilate}(\mathbf{O}^k_t)\right]$. It represents local contact evidence within a small spatial neighborhood rather than an exact physical contact point. Together, these outputs instantiate the trace representation $\mathcal{M}$. More details are provided in **Appendix** A.

3.4 Contact-Aware Trace Forcing

Although the Persistent Interaction Trace explicitly provides instance and interaction cues, these structures may still be weakened in the high-level visual representations of the MLLMs. We introduce Contact-Aware Trace Forcing, which uses the automatically constructed instance associations and contact relations in $\mathcal{M}$ to regularize intermediate visual representations during training.

Specifically, we extract the hidden states corresponding to the RGB visual tokens from the l-th layer of the MLLM, where l denotes a selected intermediate layer. The extracted features are reshaped as $\mathbf{X}^l = \{\mathbf{x}^l_{t,s}\} \in \mathbb{R}^{T \times S \times D}$, where T and S denote the temporal and spatial dimensions of the visual-token grid, and D is the hidden dimension. Using the token-level mask weights, we aggregate object, contact, and background representations from $\mathbf{X}^l$. To avoid conflicting instance supervision caused by patch-level mask overlap, only tokens with unambiguous object ownership are used for instance pooling. The pooled representations are projected and normalized before similarity computation.

3.4.1 CAUSAL TRACK ALIGNMENT

Causal Track Alignment encourages each persistent object instance to maintain a stable identity over time while remaining distinguishable from other instances. Let $\mathbf{h}_{t,k}$ denote the projected instance representation obtained by weighted pooling over object k at temporal slice t. For each valid object-time pair (t, k), we use the most recent historical representation of the same track as the positive target, while historical representations of other objects and the background form the negative set $\mathcal{N}_{t,k}$. All historical targets are detached from gradient computation, such that the current representation is constrained only by previously observed instances.

Let Ω_{tr} denote the set of valid object-time pairs with an available historical observation. The margins m_+ and m_- specify when positive pairs are sufficiently aligned, and negative pairs are sufficiently separated, respectively. For each valid object-time pair (t, k), we define the positive alignment term $\ell^+_{t,k}$ and the negative separation term $\ell^-_{t,k}$. The complete track-alignment objective is formulated as

$$\begin{cases} \ell^+_{t,k} = \Big[m_+ - s\Big(\mathbf{h}_{t,k}, \mathrm{sg}(\mathbf{h}^+_{t,k})\Big)\Big]_+ , \\ \ell^-_{t,k} = \dfrac{1}{|\mathcal{N}_{t,k}|} \displaystyle\sum_{\mathbf{n}\in\mathcal{N}_{t,k}} [s(\mathbf{h}_{t,k}, \mathrm{sg}(\mathbf{n})) - m_-]_+ , \\ \mathcal{L}_{\mathrm{track}} = \dfrac{1}{|\Omega_{\mathrm{tr}}|} \displaystyle\sum_{(t,k)\in\Omega_{\mathrm{tr}}} \Big(\ell^+_{t,k} + \ell^-_{t,k}\Big) . \end{cases} \tag{1}$$

where $s(\cdot, \cdot)$ denotes cosine similarity, $\mathbf{h}^+_{t,k}$ is the most recent historical representation of the same instance, $\mathrm{sg}(\cdot)$ denotes stop-gradient, and $[x]_+ = \max(x, 0)$. The bounded margins preserve cross-frame identity consistency while allowing reasonable representation variations caused by viewpoint and object-state changes.

3.4.2 CONTACT ALIGNMENT

Contact alignment associates localized contact evidence with the active object instance. Let Ω_{ct} denote the set of temporal slices containing valid contact evidence. For each temporal slice containing valid contact evidence, we obtain a contact representation $\mathbf{q}_t$ by weighted pooling over the contact region and reuse $\{\mathbf{h}_{t,k}\}$ as the representations of object instances. Objects selected by the active-object mask $\mathbf{A}_t$ form the positive set $\mathcal{P}_t$, while other visible objects and the contact-free background form the negative set $\mathcal{N}_t$. Objects with ambiguous spatial overlap with the contact region are excluded from the negative set. We optimize a multi-positive contrastive objective with temperature τ:

$$\mathcal{L}_{\mathrm{contact}} = -\frac{1}{|\Omega_{\mathrm{ct}}|} \sum_{t\in\Omega_{\mathrm{ct}}} \log \frac{\sum_{\mathbf{h}\in\mathcal{P}_t} e^{(s(\mathbf{q}_t,\mathbf{h})/\tau)}}{\sum_{\mathbf{h}\in\mathcal{P}_t\cup\mathcal{N}_t} e^{(s(\mathbf{q}_t,\mathbf{h})/\tau)}}. \tag{2}$$

3.5 OPTIMIZATION OBJECTIVE

The video $\mathcal{V}$ and mask video $\mathcal{V}^{\mathrm{hoi}}$ are encoded as two visual streams, augmented with the learnable marks, and jointly fed into the MLLMs together with the prompt Q to autoregressively generate the response Y. We jointly optimize the autoregressive LM objective and the two forcing constraints: $\mathcal{L} = \mathcal{L}_{\mathrm{LM}} + \lambda_{\mathrm{track}}\mathcal{L}_{\mathrm{track}} + \lambda_{\mathrm{contact}}\mathcal{L}_{\mathrm{contact}}$, where $\mathcal{L}_{\mathrm{LM}}$ is the cross-entropy loss applied only to the answer tokens. During inference, the forcing heads and auxiliary losses are removed.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

4.1.1 DATASET

We conduct experiments on HanDyVQA Tateno et al. (2026a), a benchmark designed to evaluate fine-grained spatio-temporal understanding of hand-object interaction dynamics. It is constructed from five-second egocentric video clips collected from Ego4D Grauman et al. (2022) and contains

Table 1: Quantitative comparison on HanDyVQA. TraceForce-HOI achieves the best overall performance. *Reproduced on Qwen2.5-VL-3B under the same fine-tuning setting, using the same HOI segmentation and tracking results for fair comparison.

Models	HOI Modeling	Resolution	#Frames	Objects (AP)	Action (Acc)	Process (Acc)	Location (Acc)	State (Acc)	Parts (Acc)	Avg.
Random	–	–	–	28.6	19.3	18.9	20.4	19.8	19.4	19.5
Open-source video-language models										
LaViLa (TSF-L)	Generic video	224×224	16	68.5	61.6	40.1	36.9	38.9	35.6	42.6
InternVideo2-Stage2	Generic video	224×224	16	37.0	41.1	30.2	29.7	34.9	30.5	33.3
Open-source MLLMs										
VideoLLaMA2.1-7B	Generic video	384×384	16	52.9	41.4	47.3	34.6	47.0	40.5	42.1
LLaVA-Video-7B	Generic video	384×384	16	60.4	56.9	53.7	50.5	58.5	54.6	54.8
Qwen2.5-VL-3B	Generic video	480×854	2 fps (10 avg.)	43.7	34.2	25.9	24.6	30.2	30.9	29.2
Qwen2.5-VL-7B	Generic video	480×854	2 fps (10 avg.)	48.6	59.1	55.6	47.6	56.8	48.8	53.6
Qwen3-VL-8B	Generic video	480×854	2 fps (10 avg.)	59.2	71.6	64.9	58.3	65.3	58.8	63.8
Proprietary MLLMs										
GPT-4o (vision)	Generic video	480×854	16	64.1	61.3	64.4	51.5	59.0	58.5	58.9
Gemini-2.5-Pro	Generic video	480×854	16	**78.8**	79.1	73.3	67.6	73.9	69.3	72.6
SFT										
EgoVideo-L	HOI caption pretraining	336×336	16	63.9	70.5	69.5	54.0	51.3	56.6	60.4
Qwen2.5-VL-3B-SFT	RGB-only reasoning	480×854	2 fps (10 avg.)	64.2	70.7	66.1	66.3	68.9	57.3	65.9
VideoMindPalace*	Semantic Graph	480×854	2 fps (10 avg.)	62.3	80.4	78.6	73.0	77.9	65.2	75.0
VideoOrion*	Object Dynamics tokens	480×854	2 fps (10 avg.)	64.7	82.7	82.3	76.2	79.6	70.6	78.3
TraceForce-HOI-3B (Ours)	**Interaction trace forcing**	480×854	2 fps (10 avg.)	66.6	**84.6**	**86.0**	**76.8**	**80.0**	**73.3**	**80.1**

11,092 human-verified question-answer pairs. The benchmark covers six complementary question categories. *Action*, *Process*, and *Objects* evaluate the manipulation aspect of HOI, including the performed action, fine-grained manipulation process, and objects involved in the interaction. *Location*, *State Change*, and *Object Parts* evaluate the resulting effects, including spatial displacement, object-state transitions, and part-level changes. We follow the official data split and use 1,116 question-answer pairs for training, 545 pairs for validation, and the remaining 9,431 pairs for testing.

4.1.2 Evaluation Metrics and Baselines

Following the official evaluation protocol of HanDyVQA, we report top-1 accuracy for the *Action*, *Process*, *Location*, *State Change*, and *Object Parts* categories. Since an *Objects* question may contain multiple correct answers, we use Average Precision (AP) for this category. The overall score, denoted as *Avg.*, is the arithmetic mean over the five accuracy-based categories and excludes the *Objects* score because AP and classification accuracy follow different evaluation protocols.

We compare TraceForce-HOI with representative video-language models across different architectures and model scales. The open-source baselines include LaViLa (TSF-L) Zhao et al. (2023), InternVideo2-Stage2 Wang et al. (2024), VideoLLaMA2.1-7B Cheng et al. (2024), LLaVA-Video-7B Lin et al. (2024), Qwen2.5-VL-3B/7B Bai et al. (2025b), and Qwen3-VL-8B Bai et al. (2025a). We further compare with the proprietary GPT-4o Hurst et al. (2024) and Gemini-2.5-Pro Comanici et al. (2025). We additionally include EgoVideo-L Pei et al. (2025) and two closest structured baselines: VideoMindPalace Huang et al. (2025a), which organizes video evidence as semantic graphs, and VideoOrion Feng et al. (2025a), which tokenizes object dynamics.

4.1.3 Implementation Details

We implement TraceForce-HOI based on Qwen2.5-VL-3B-Instruct Bai et al. (2025b). We default to sampling the input videos at 2 FPS (Avg. 10 frames/video) as the inputs to the MLLM. Notably, our method can be adapted to different backbones. The Fast-Slow Trace Identification module adopts a 5× denser temporal sampling strategy in the Fast branch, using 10 FPS dense observations to capture transient interaction cues. The learnable interaction marks have a dimension of 256, with an object-slot bank of size 8. The marks are residually added to both visual streams. Contact-Aware Trace Forcing is applied to the RGB visual tokens from the 24th LLM layer, with positive and negative margins of 0.6 and 0.2 for Causal Track Alignment and loss weights $\lambda_{\text{track}} = 0.005$ and $\lambda_{\text{contact}} = 0.01$. We fine-tune the model with LoRA Hu et al. (2022) adapters using a rank of 32 and a scaling factor of 64. All experiments are performed on a single NVIDIA A800 80GB GPU using the AdamW optimizer, with an initial learning rate of 1×10^{-4}, cosine learning rate scheduling, a weight decay of 0.01, and a global batch size of 32. The forcing heads are only used during training and removed during inference.

4.2 Quantitative Evaluation

Table 1 compares TraceForce-HOI with existing video-language models on HanDyVQA. TraceForce-HOI achieves the best overall performance despite using a 3B backbone, surpassing Qwen3-VL-8B and improving the same-backbone Qwen2.5-VL-3B-SFT by 14.2 Avg. points. More importantly, it outperforms the two closest structured baselines across every category, exceeding VideoMindPalace and VideoOrion by 5.1 and 1.8 Avg. points, respectively. Against VideoOrion, the largest gains occur on Process and Object Parts, where answers require linking transient contact cues to persistent entities and evolving states. Unlike semantic graphs or object-dynamics tokens, our traces jointly encode identities, localized contact, active-object state, and interaction evolution, and explicitly preserve contact-to-instance continuity in intermediate representations. On Objects AP, TraceForce-HOI performs slightly below LaViLa and Gemini-2.5-Pro. Unlike the accuracy-based categories, Objects evaluates multi-answer object retrieval and therefore favors exhaustive object coverage. Our interaction-centric pipeline prioritizes interaction-relevant tracks and emphasizes active-object and contact attribution, which benefits fine-grained manipulation reasoning but may under-represent peripheral or weakly interacted objects.

4.3 Efficiency Analysis

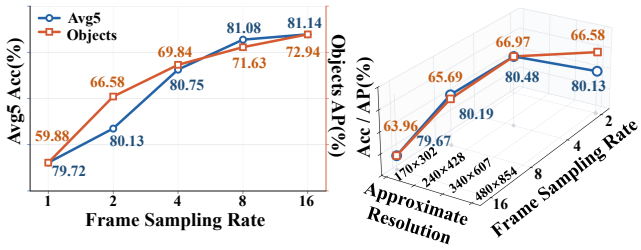


Figure 3: Efficiency analysis under different input sampling configurations. (a) Performance variation with different frame sampling rates under fixed spatial resolution. (b) Performance trade-off between temporal sampling rate and spatial resolution under a fixed visual input budget.

We further investigate the influence of visual input sampling on HOI understanding during inference. In real-world HOI applications, the visual input budget is inherently limited, making it impractical to indefinitely increase video resolution or frame sampling rate. Therefore, a key challenge is balancing temporal coverage and spatial details under a fixed visual budget while preserving interaction-relevant information from limited observations. As shown in Fig. 3, we evaluate different frame sampling rates under two settings: fixed spatial resolution and fixed overall visual token budget.

Under fixed spatial resolution, increasing the sampling rate improves performance by providing denser temporal observations, but the gain gradually saturates after sufficient temporal coverage is reached. This indicates that simply increasing the number of frames cannot fully solve the instance association problem in HOI videos, since additional observations may contain redundant visual information without reliable hand-object correspondence.

Under a fixed visual budget, we observe that maximizing either spatial resolution or temporal sampling rate does not necessarily lead to better HOI understanding. The best performance is achieved with a balanced configuration of 4 FPS sampling and 240×427 resolution, outperforming settings with higher spatial or temporal inputs. This suggests that HOI understanding requires a careful balance between spatial details and temporal coverage rather than simply increasing visual inputs. Excessive resolution may introduce redundant background tokens beyond interaction-relevant regions, while excessive frame sampling may provide highly correlated observations without improving action association. Additional computational cost analysis is provided in **Appendix** D.

4.4 Visualization Analysis

To further understand how TraceForce-HOI improves HOI reasoning, we visualize the attention responses between answers and visual tokens in Fig. 4. Compared with Qwen2.5-VL-SFT, TraceForce-HOI produces more localized and temporally consistent attention patterns around the manipulated objects, hand regions, and contact areas. In the example, the question requires identifying not only the involved entities but also the sequential interaction process: holding the knife, turning on the faucet, and wiping the blade under running water. Qwen2.5-VL-SFT correctly recognizes several entities but fails to associate them with the evolving interaction states, leading to an incomplete response. Its attention responses are mainly distributed over salient visual regions, without maintaining stable correspondence between the hands and the relevant objects. In contrast,

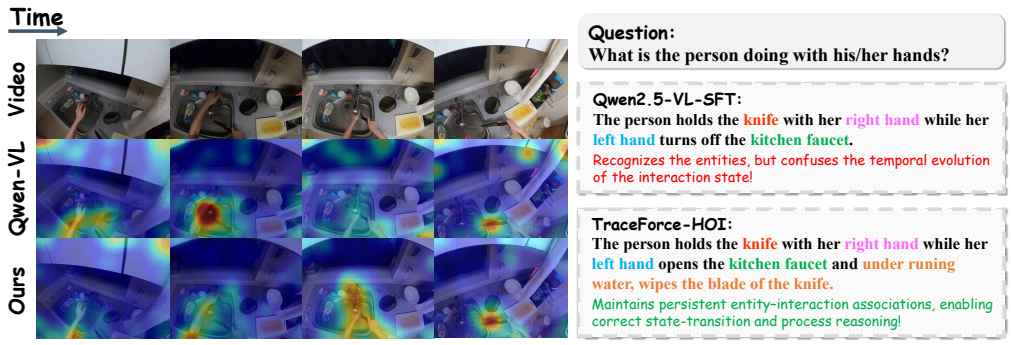


Figure 4: Qualitative comparison between Qwen2.5-VL-SFT and TraceForce-HOI. TraceForce-HOI produces more localized and temporally consistent attention patterns around the interacting hands, objects, and contact regions, enabling better preservation of HOI evolution.

TraceForce-HOI consistently attends to the interacting hand, knife, and faucet regions across different temporal moments. This behavior suggests that the proposed Persistent Interaction Traces provide more structured visual grounding, allowing the model to preserve object identity and contact relationships during temporal reasoning. The visualization supports our hypothesis that explicit interaction trace representations help MLLMs preserve consistent hand-object associations and capture interaction evolution for HOI reasoning.

4.5 Ablation Study

Table 2: Left: component analysis. Row 2 constructs PIT via per-frame extraction and IoU-based association. Right: forcing-layer index ablation.

PIT	FSI	Force	Obj.	Avg5.	Layer	Obj.	Avg5.
			64.2	65.9	8	65.8	79.9
✓			64.6	75.6	16	66.4	80.0
✓	✓		65.4	77.0	**24**	**66.6**	**80.1**
✓	✓	✓	**66.6**	**80.1**	32	66.5	79.9

We evaluate the contribution of each component and the placement of trace forcing in Table 2. Starting from the baseline, Persistent Interaction Traces (PIT) substantially improve performance by explicitly organizing instance identities, contact cues, and interaction states. Fast-Slow Trace Identification (FSI) provides a further gain through more reliable temporal association, while Contact-Aware Trace Forcing achieves the best overall performance by preserving these interaction structures in intermediate representations. We further vary the forcing-layer index with the full model. Layer 24 performs best, while performance remains stable across nearby layers. Shallow layers mainly encode local appearance patterns and have not yet formed mature entity-interaction semantics, whereas very deep layers increasingly aggregate visual evidence into question-conditioned semantics for answer prediction, which may blur the fine-grained instance and contact distinctions targeted by trace forcing. Additional fine-grained ablation studies are provided in **Appendix** C.

5 Conclusions

We presented TraceForce-HOI, a unified framework for fine-grained HOI video understanding that explicitly models interaction structures beyond the implicit reasoning of multimodal large language models. Motivated by the need to associate transient contact cues with persistent hand and object identities, TraceForce-HOI combines Persistent Interaction Traces in pixel and latent spaces, Fast-Slow Trace Identification for contact-sensitive and temporally consistent structure construction, and Contact-Aware Trace Forcing to preserve these structures in high-level visual representations. Extensive experiments demonstrate its effectiveness in understanding interaction processes, state transitions, and object-part changes. This perspective supports more explicit, persistent, and interaction-grounded reasoning for longer-horizon and open-world embodied scenarios. Despite its effectiveness, the current formulation remains largely confined to 2D representations. Future work may extend Persistent Interaction Traces toward 3D-aware representations with articulated hand modeling, enabling more precise reasoning about spatial geometry, contact dynamics, and physical interactions in complex embodied environments.

AI USE STATEMENT

We used generative AI tools to assist with code implementation and improve the clarity and presentation of the manuscript. These tools were not used to conduct experiments, generate reported results, or produce bibliographic references. All AI-assisted code was manually reviewed for correctness, and all cited literature was independently located and verified by the authors. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses existing research datasets and pretrained models in accordance with their applicable licenses and terms of use. We did not recruit participants or collect additional personal data. The proposed method is intended for research on video understanding and does not aim to infer personal identities or sensitive attributes. Nevertheless, it may inherit biases and limitations from the underlying data and models, and its use in real-world applications requires appropriate safeguards and further evaluation.

REPRODUCIBILITY STATEMENT

The main paper describes the model architecture, learning objectives, datasets, evaluation protocol, and training configuration. The appendix and supplementary materials provide the necessary algorithmic and implementation details. We plan to release the complete code, training and inference configurations, generated interaction traces, and model checkpoints.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *CVPR*, pp. 7061–7071, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, RISHI HAZRA, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *ICLR*, 2026. URL https://openreview.net/forum?id=r35clVtGzw.

Qirui Chen, Shangzhe Di, and Weidi Xie. Grounded multi-hop videoqa in long-form egocentric videos. In *AAAI*, volume 39, pp. 2159–2167, 2025.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Yura Choi, Roy Miles, Rolandos Alexandros Potamias, Ismail Elezi, Jiankang Deng, and Stefanos Zafeiriou. Do you see what i am pointing at? gesture-based egocentric video question answering. In *CVPR*, pp. 18534–18544, 2026.

Qiaohui Chu, Haoyu Zhang, Meng Liu, Yisen Feng, Haoxiang Shi, and Liqiang Nie. Intention-guided cognitive reasoning for egocentric long-term action anticipation. In *AAAI*, volume 40, pp. 17436–17444, 2026.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Lingwei Dang, Juntong Li, Zonghan Li, Hongwen Zhang, Liang An, Wei Min, Yebin Liu, and Qingyao Wu. Harmohoi: Harmonizing appearance and 3d motion for multi-view hand-object interaction synthesis, 2026a. URL https://arxiv.org/abs/2607.17097.

Lingwei Dang, Ruizhi Shao, Hongwen Zhang, Wei Min, Yebin Liu, and Qingyao Wu. Svimo: Synchronized diffusion for video and motion generation in hand-object interaction scenarios. *NeurIPS*, 38:76689–76718, 2026b.

Junhao Du, Jialong Xue, Anqi Li, Jincheng Dai, and Guo Lu. Unified spatiotemporal token compression for video-llms at ultra-low retention. In *CVPR*, pp. 17661–17671, 2026.

Yicheng Feng, Yijiang Li, Wanpeng Zhang, Sipeng Zheng, Hao Luo, Zihao Yue, and Zongqing Lu. Videoorion: Tokenizing object dynamics in videos. In *ICCV*, pp. 20401–20412, 2025a.

Yigui Feng, Qinglin Wang, Yang Liu, and Jie Liu. Fre-res: Frequency-residual video token compression for efficient video mllms. *arXiv preprint arXiv:2605.16366*, 2026.

Yisen Feng, Haoyu Zhang, Meng Liu, Weili Guan, and Liqiang Nie. Object-shot enhanced grounding network for egocentric video. In *CVPR*, pp. 24190–24200, 2025b.

Chenhui Gou, Ziyu Ma, Zicheng Duan, Haoyu He, Feng Chen, Akide Liu, Bohan Zhuang, Jianfei Cai, and Hamid Rezatofighi. An empirical study on how video-llms answer video questions. In *CVPR*, pp. 18587–18597, 2026.

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 *CVPR*, pp. 18995–19012, 2022.

Chaohong Guo, Yihan He, Yongwei Nie, Fei Ma, Xuemiao Xu, and Chengjiang Long. T2sgrid: Temporal-to-spatial gridification for video temporal grounding. In *CVPR*, pp. 3443–3454, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Zeyi Huang, Yuyang Ji, Xiaofang Wang, Nikhil Mehta, Tong Xiao, Donghyun Lee, Sigmund Vanvalkenburgh, Shengxin Zha, Bolin Lai, Licheng Yu, et al. Building a mind palace: Structuring environment-grounded semantic graphs for effective long video analysis with llms. In *CVPR*, pp. 24169–24179, 2025a.

Ziyao Huang, Zixiang Zhou, Juan Cao, Yifeng Ma, Yi Chen, Zejing Rao, Zhiyong Xu, Hongmei Wang, Qin Lin, Yuan Zhou, et al. Homa: Towards generic human-object interaction in multimodal driven human animation with weak conditions. In *SIGGRAPH Asia*, pp. 1–12, 2025b.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jae Yong Lee, Daniel Scharstein, Akash Bapat, Hao Hu, Andrew Fu, Haoru Zhao, Paul Sammut, Xiang Li, Stephen Jeapes, Anik Gupta, et al. Ego-1k-a large-scale multiview video dataset for egocentric vision. In *CVPR*, pp. 19854–19863, 2026a.

Seungjae Lee, Yoonkyo Jung, Inkook Chun, Yao-Chih Lee, Zikui Cai, Hongjia Huang, Aayush Talreja, Tan Dao, Yongyuan Liang, Jia-Bin Huang, et al. Tracegen: World modeling in 3d trace space enables learning from cross-embodiment videos. In *CVPR*, pp. 20721–20731, 2026b.

Seungjae Lee, Yoonkyo Jung, Jusuk Lee, Jonghun Shin, Amir Hossein Shahidzadeh, Yao-Chih Lee, H Jin Kim, Jia-Bin Huang, and Furong Huang. μ_0: A scalable 3d interaction-trace world model. *arXiv preprint arXiv:2606.13769*, 2026c.

Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long ZENG, and Haoang Li. Spatial forcing: Implicit spatial representation alignment for vision-language-action model. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=euMVC1DO4k.

Yayuan Li, Aadit Jain, Filippos Bellos, and Jason J Corso. Mistake attribution: Fine-grained mistake understanding in egocentric videos. In *CVPR*, pp. 23966–23976, 2026b.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *EMNLP*, pp. 5971–5984, 2024.

Xinhang Liu, Yuxi Xiao, Donny Y. Chen, Jiashi Feng, Yu-Wing Tai, Chi-Keung Tang, and Bingyi Kang. Trace anything: Representing any video in 4d via trajectory fields. In *ICLR*, 2026a. URL https://openreview.net/forum?id=BqaChqppVh.

Yushan Liu, Peibo Sun, Shoujie Li, Yifan Xie, Lingfeng Zhang, Xintao Chao, Shiyuan Dong, Fang Chen, Xiao-Ping Zhang, and Wenbo Ding. Oa-wam: Object-addressable world action model for robust robot manipulation. *arXiv preprint arXiv:2605.06481*, 2026b.

Arsha Nagrani, Jasper Uijlings, Shyamal Buch, Tobias Weyand, Sudheendra Vijayanarasimhan, Bo Hu, Ramin Mehran, David A Ross, and Cordelia Schmid. Minerva-ego: Spatiotemporal hints for egocentric video understanding. In *CVPR*, pp. 38859–38869, 2026.

Baoqi Pei, Yifei Huang, Jilan Xu, Guo Chen, Yuping He, Lijin Yang, Yali Wang, Weidi Xie, Yu Qiao, Fei Wu, and Limin Wang. Modeling fine-grained hand-object dynamics for egocentric video representation learning. In *ICLR*, 2025. URL https://openreview.net/forum?id=P6G1Z6jkf3.

Chiara Plizzari, Alessio Tonioni, Yongqin Xian, Achin Kulshrestha, and Federico Tombari. Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. In *CVPR*, pp. 24129–24138, 2025.

Yayun Qi and Xinxiao Wu. What to trust? a trust-aware knowledge-guided method for zero-shot object state understanding in videos. In *AAAI*, volume 40, pp. 8484–8492, 2026.

YUEJIAO SU, Yi Wang, Lei Yao, Yawen Cui, and Lap-Pui Chau. Interaction-aware representation modeling with co-occurrence consistency for egocentric hand-object parsing. In *ICLR*, 2026. URL https://openreview.net/forum?id=RYwQ0xQcAh.

Zhihao Sun, Zhiying Du, Xitong Yang, and Zuxuan Wu. Handworld: Hand-centric unified video action generation. In *CVPR*, pp. 15976–15985, 2026.

Yolo Y Tang, Jing Bi, Pinxin Liu, Zhenyu Pan, Zhangyun Tan, Qianxiang Shen, Jiani Liu, Hang Hua, Junjia Guo, Yunzhong Xiao, et al. Video-lmm post-training: A deep dive into video reasoning with large multimodal models. *arXiv preprint arXiv:2510.05034*, 2025.

Masatoshi Tateno, Gido Kato, Hirokatsu Kataoka, Yoichi Sato, and Takuma Yagi. Handyvqa: A video qa benchmark for fine-grained hand-object interaction dynamics. In *CVPR*, pp. 3455–3465, 2026a.

Masatoshi Tateno, Alexandros Stergiou, Risa Shinoda, Yoichi Sato, and Dima Damen. Do egocentric video-language models capture both hand-and object-centric cues? *arXiv preprint arXiv:2607.08514*, 2026b.

Tze Ho Elden Tse, Runyang Feng, Linfang Zheng, Jiho Park, Yixing Gao, Jihie Kim, Ales Leonardis, and Hyung Jin Chang. Collaborative learning for 3d hand-object reconstruction and compositional action recognition from egocentric rgb videos using superquadrics. In *AAAI*, volume 39, pp. 7437–7445, 2025.

Jiankang Wang, Zhihan Zhang, Zhihang Liu, Yang Li, Jiannan Ge, Hongtao Xie, and Yongdong Zhang. Spacevllm: Endowing multimodal large language model with spatio-temporal video grounding capability. In *AAAI*, volume 40, pp. 9912–9920, 2026a.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV*, pp. 396–416. Springer, 2024.

Yuqing Wang, Zhijie Lin, Ceyuan Yang, Yang Zhao, Fei Xiao, Hao He, Qi Zhao, Zihan Ding, Fuyun Wang, Shuai Wang, et al. Representation forcing for bottleneck-free unified multimodal models. *arXiv preprint arXiv:2605.31604*, 2026b.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. In *ICLR*, 2026. URL https://openreview.net/forum?id=ULXYZCms41.

Junbin Xiao, Shenglang Zhang, Pengxiang Zhu, and Angela Yao. Ego-grounding for personalized question-answering in egocentric videos. In *CVPR*, pp. 40537–40547, 2026.

Ziyi Xu, Zejing Rao, Juan Cao, Xiaoqiang Liu, Zhixue Fang, Haoxian Zhang, Songlin Tang, and Fan Tang. Geohoi: Geometry-enhanced human-object interaction video generation via hierarchical multi-modal injection. In *CVPR*, pp. 3739–3748, 2026.

Yufei Ye, Jiaman Li, Ryan Rong, and C Karen Liu. Whole: World-grounded hand-object lifted from egocentric videos. In *CVPR*, pp. 3481–3491, 2026.

Taegyoon Yoon, Yegyu Han, Seojin Ji, Jaewoo Park, Sojeong Kim, Taein Kwon, and Hyung-Sin Kim. Egoxtreme: A dataset for robust object pose estimation in egocentric views under extreme conditions. In *CVPR*, pp. 40696–40706, 2026.

Haoyu Zhang, Qiaohui Chu, Meng Liu, Haoxiang Shi, Yaowei Wang, and Liqiang Nie. Exo2ego: Exocentric knowledge guided mllm for egocentric video understanding. In *AAAI*, volume 40, pp. 12502–12510, 2026a.

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. In *CVPR*, pp. 1805–1815, 2025.

Jusheng Zhang, Xiaoyang Guo, Kaitong Cai, Qinhan Lv, Yijia Fan, Wenhao Chai, Jian Wang, and Keze Wang. Hybrid token compression for vision-language models. In *CVPR*, pp. 31889–31899, 2026b.

Peiwen Zhang, Yufan Deng, Shangkun Sun, Juncheng Ma, Duomin Wang, Jonas Du, Zilin Pan, Ye Huang, Hao Liang, Songyan Huang, et al. Physisforcing: Physics reinforced world simulator for robotic manipulation. *arXiv preprint arXiv:2606.28128*, 2026c.

Zhenhao Zhang, Ye Shi, Lingxiao Yang, Suting Ni, Qi Ye, and Jingya Wang. Openhoi: Open-world hand-object interaction synthesis with multimodal large language model. *NeurIPS*, 38:166582–166612, 2026d.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *CVPR*, pp. 6586–6597, 2023.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Zhifan Zhu, Siddhant Bansal, Shashank Tripathi, and Dima Damen. Reconstructing objects along hand interaction timelines in egocentric video. In *CVPR*, pp. 3955–3965, 2026.

APPENDIX

A DETAILED FAST-SLOW TRACE IDENTIFICATION

Overview. The Fast-Slow Trace Identification module converts an input egocentric video into the Persistent Interaction Trace $\mathcal{M} = \{\mathbf{H}_t^L, \mathbf{H}_t^R, \{\mathbf{O}_t^k\}_{k=1}^K, \mathbf{A}_t, \mathbf{C}_t\}_{t=1}^T$ used by TraceForce-HOI. Its two branches decouple transient interaction discovery from temporally consistent instance propagation. The Fast branch performs interaction-aware perception on a dense timeline and associates frame-local object observations into coarse tracklets without invoking expensive video mask propagation. The Slow branch then propagates these observations only on a sparse timeline, consolidates duplicate tubes, and retains the most interaction-relevant persistent instances. Finally, frame-wise interacted-object observations are attributed to the persistent tracks, and localized contact evidence is reconstructed from the refined hand and object masks. Algorithm 1 summarizes the complete procedure.

A.1 FAST BRANCH: DENSE INTERACTION-AWARE IDENTIFICATION

Dual-timescale sampling and frame alignment. Given an input video $\mathcal{V}$, we construct a dense Fast timeline $\mathcal{T}_{\text{fast}}$ and a sparse Slow timeline $\mathcal{T}_{\text{slow}}$ by uniformly sampling the full temporal extent of the video. We use $f_{\text{fast}} = 10$ FPS for dense interaction-aware perception and $f_{\text{slow}} = 2$ FPS for the final visual input. For the approximately five-second clips in HanDyVQA, the Slow timeline therefore typically contains ten frames.

We decode the ordered union

$$\mathcal{T}_{\text{master}} = \text{SortedUnion}\left(\mathcal{T}_{\text{fast}}, \mathcal{T}_{\text{slow}}\right), \tag{3}$$

and perform frame-wise interaction-aware perception on all frames in $\mathcal{T}_{\text{master}}$. Because every Slow frame is explicitly included in the master timeline, its semantic observation is obtained from the exact same video frame rather than through temporal nearest-neighbor matching. The additional Fast frames are used only to discover short-lived interaction evidence and construct coarse object tracklets. After identity-consistent propagation, the RGB frames, persistent masks, and HOI Mask Video are retained only on $\mathcal{T}_{\text{slow}}$.

Dense semantic parsing. For every master frame, InterFormer SU et al. (2026) predicts semantic masks for the left hand, right hand, and the objects interacting with the left hand, right hand, or both hands. We merge the three interacted-object roles into a binary interacted-object foreground and decompose it into 8-connected components $\mathcal{D}_t = \{\mathbf{D}_t^n\}$. Components smaller than 300 pixels are removed, while left- and right-hand masks smaller than 100 pixels are treated as invalid observations. The hand-specific object role is retained only as diagnostic metadata; persistent object identity is independent of whether an object is currently manipulated by the left hand, right hand, or both hands. Thus, a handover does not create a new object identity.

Motion and appearance descriptors. Each coarse tracklet stores its latest mask, the three most recent centroid observations, and a running appearance descriptor. We estimate constant-velocity motion from the latest two centroids and translate the latest mask to the current frame. The predicted horizontal and vertical displacements are clipped to $\pm W/3$ and $\pm H/3$, respectively. The appearance descriptor is a 54-dimensional vector composed of a normalized 16-bin histogram for each RGB channel (48 dimensions), together with the normalized RGB mean and standard deviation (six dimensions). The resulting vector is ℓ_2-normalized, and appearance similarity s_{app} is measured by cosine similarity. For regions containing more than $20{,}000$ pixels, pixels are uniformly subsampled before computing the descriptor.

Algorithm 1 Fast-Slow Trace Identification

Input: Raw video $\mathcal{V}$; Fast and Slow sampling rates $f_{\text{fast}} = 10$ and $f_{\text{slow}} = 2$; discovery capacity $K_{\text{disc}} = 16$; output capacity $K = 8$
Output: Slow-timeline RGB frames, persistent hand and object tubes, active/contact masks, and HOI Mask Video

$\mathcal{T}_{\text{fast}} \leftarrow \textsc{UniformSample}(\mathcal{V}, f_{\text{fast}})$
$\mathcal{T}_{\text{slow}} \leftarrow \textsc{UniformSample}(\mathcal{V}, f_{\text{slow}})$
$\mathcal{T}_{\text{master}} \leftarrow \textsc{SortedUnion}(\mathcal{T}_{\text{fast}}, \mathcal{T}_{\text{slow}})$
$\{\mathbf{I}_t\} \leftarrow \textsc{Decode}(\mathcal{V}, \mathcal{T}_{\text{master}})$; $\{\mathbf{S}_t\} \leftarrow \textsc{InterFormer}(\{\mathbf{I}_t\})$
$\mathcal{G} \leftarrow \emptyset$ *// coarse object tracklets*
for each $t \in \mathcal{T}_{\text{master}}$ **do**
  $\mathcal{D}_t \leftarrow \textsc{ObjectComponents}(\mathbf{S}_t)$; remove components with area < 300
  $\widehat{\mathcal{G}} \leftarrow \textsc{MotionPredict}(\mathcal{G}, t)$
  $\mathcal{P}_t \leftarrow \textsc{ValidPairs}(\mathcal{D}_t, \widehat{\mathcal{G}})$
  $\mathcal{G} \leftarrow \textsc{GreedyAssociate}(\mathcal{G}, \mathcal{D}_t, \mathcal{P}_t, K_{\text{disc}})$
end for
for each coarse tracklet $g_k \in \mathcal{G}$ **do**
  $t_k^{\text{seed}} \leftarrow$ earliest valid observation of g_k
  $\mathcal{S}_k \leftarrow \{t_k^{\text{seed}}\} \cup \mathcal{T}_{\text{slow}}$
  $\mathbf{O}^k \leftarrow \textsc{BidirectionalSAM3}(\mathcal{S}_k, g_k)$
  Retain $\mathbf{O}^k$ at $\mathcal{T}_{\text{slow}}$ and restore prompt evidence
end for
$(\mathbf{H}^L, \mathbf{H}^R) \leftarrow \textsc{JointHandPropagation}(\mathcal{T}_{\text{slow}}, \{\mathbf{S}_t\})$
repeat
  $\mathcal{U} \leftarrow \textsc{FindDuplicateTubes}(\{\mathbf{O}^k\})$
  Merge each group in $\mathcal{U}$ and re-propagate its aggregated observations
until no tube is merged
Rank tubes by contact, observations, visibility, and accumulated area
$\{\mathbf{O}^k\}_{k=1}^K \leftarrow \textsc{TopK}(\{\mathbf{O}^k\}, K)$
for each $t \in \mathcal{T}_{\text{slow}}$ **do**
  $\mathbf{A}_t \leftarrow \textsc{AssignActiveObjects}(\mathbf{S}_t, \{\mathbf{O}_t^k\}_{k=1}^K)$
  $\mathbf{C}_t \leftarrow \textsc{ConstructContact}(\mathbf{H}_t^L, \mathbf{H}_t^R, \{\mathbf{O}_t^k\}_{k=1}^K)$
  $\mathbf{P}_t \leftarrow \textsc{RenderHOIMask}(\mathbf{H}_t^L, \mathbf{H}_t^R, \{\mathbf{O}_t^k\}_{k=1}^K, \mathbf{C}_t)$
end for
return $\{\mathbf{I}_t, \mathbf{H}_t^L, \mathbf{H}_t^R, \{\mathbf{O}_t^k\}_{k=1}^K, \mathbf{A}_t, \mathbf{C}_t, \mathbf{P}_t\}_{t \in \mathcal{T}_{\text{slow}}}$

For a frame-local candidate $\mathbf{D}$ and a motion-predicted track mask $\widehat{\mathbf{O}}$, we define

$$\begin{cases} \text{IoU}(\mathbf{D}, \widehat{\mathbf{O}}) = \dfrac{|\mathbf{D} \cap \widehat{\mathbf{O}}|}{|\mathbf{D} \cup \widehat{\mathbf{O}}|}, \\ \text{IoM}(\mathbf{D}, \widehat{\mathbf{O}}) = \dfrac{|\mathbf{D} \cap \widehat{\mathbf{O}}|}{\min(|\mathbf{D}|, |\widehat{\mathbf{O}}|)}, \\ s_{\text{area}} = \dfrac{\min(|\mathbf{D}|, |\widehat{\mathbf{O}}|)}{\max(|\mathbf{D}|, |\widehat{\mathbf{O}}|)}, \\ d_{\text{ctr}} = \dfrac{\|c(\mathbf{D}) - c(\widehat{\mathbf{O}})\|_2}{\sqrt{H^2 + W^2}}. \end{cases} \tag{4}$$

The IoM term is particularly useful when a coarse semantic mask covers only part of an object or when the two masks have substantially different areas.

Online coarse-track association. A candidate–track pair is considered valid if it satisfies at least one of the following gates:

$$\begin{cases} \textit{Spatial:} & \text{IoU} \geq 0.08 \text{ or } \text{IoM} \geq 0.28, \\ \textit{Local:} & \Delta t \leq 8,\ d_{\text{ctr}} \leq 0.16, \\ & s_{\text{app}} \geq 0.72,\ s_{\text{area}} \geq 0.20, \\ \textit{Re-ID:} & \Delta t \leq 24,\ d_{\text{ctr}} \leq 0.42, \\ & s_{\text{app}} \geq 0.94,\ s_{\text{area}} \geq 0.32. \end{cases} \tag{5}$$

Here, Δt is measured in discrete master-timeline steps. Each valid pair is ranked by

$$\begin{aligned} s(\mathbf{D}, \widehat{\mathbf{O}}) = {} & 2.4 \max(\mathrm{IoU}, \mathrm{IoM}) \\ & + 1.1 s_{\mathrm{app}} + 0.45 s_{\mathrm{area}} \\ & + 0.35 \max\left(0, 1 - \frac{d_{\mathrm{ctr}}}{0.42}\right) - 0.025 \Delta t. \end{aligned} \tag{6}$$

We sort all valid pairs in descending order and perform one-to-one greedy matching. Each unmatched candidate initializes a new coarse tracklet until the discovery capacity $K_{\mathrm{disc}} = 16$ is reached. After a successful association, the running descriptor is updated as

$$\mathbf{d}_{\mathrm{track}} \leftarrow \mathrm{Normalize}(0.75 \mathbf{d}_{\mathrm{track}} + 0.25 \mathbf{d}_{\mathrm{current}}) \,. \tag{7}$$

This branch is intentionally recall-oriented: it aggregates dense observations and preserves short-lived interaction evidence, but does not produce the final object boundaries and never invokes SAM3 on the dense sequence.

A.2 Slow Branch: Sparse Propagation and Identity Consolidation

Seed-plus-Slow-frame propagation. For each coarse object tracklet k, we select its earliest valid InterFormer observation as the propagation seed t_k^{seed} and construct a temporary SAM3 sequence

$$\mathcal{S}_k = \{t_k^{\mathrm{seed}}\} \cup \mathcal{T}_{\mathrm{slow}}. \tag{8}$$

The earliest observation is always inserted as a mask prompt. Other observations that lie exactly on Slow-frame indices are also used as prompts; multiple components assigned to the same track in one frame are first merged into a single mask. All prompts are written into the SAM3 memory, after which we propagate forward from the earliest prompt and backward from the latest prompt. Only masks at the Slow-frame indices are retained. Consequently, SAM3 operates only on the sparse Slow sequence, optionally augmented with one seed frame, rather than propagating over the complete dense timeline.

The two hands are propagated jointly in one SAM3 state as two predefined identities. All valid left- and right-hand masks at the Slow frames are supplied as prompts, preventing identity reassignment based on their spatial ordering in an individual frame. For each object tube, the SAM3 output is preferred. If the propagated mask is empty at a frame, we recover the largest non-empty mask among the available fallback tubes. If an InterFormer anchor exists at that frame, it is unioned with the recovered result so that exact observation evidence is not removed by propagation.

Iterative identity consolidation. Dense discovery may produce multiple coarse tracklets for the same physical object, particularly before and after an occlusion. We compare two propagated tubes only on Slow frames where both are non-empty. In addition, we mark an *observation conflict* when both tracks contain InterFormer observations at the same master frame and any corresponding observation pair has $\mathrm{IoM} < 0.50$. Two tubes are considered duplicates if either of the following conditions holds:

1. **Repeated high overlap:** at least three shared non-empty frames satisfy $\mathrm{IoU} \geq 0.70$ and $\mathrm{IoM} \geq 0.90$;
2. **Stable tube overlap:** there are at least three shared non-empty frames, the mean cross-frame IoU is at least 0.25, the maximum IoM is at least 0.90, and no observation conflict exists.

All duplicate relations are grouped with union–find. We aggregate the InterFormer observations in each group, run sparse bidirectional propagation again, and retain the original member tubes as fallback masks. The detection–merging–repropagation procedure is repeated until the number of tubes no longer decreases.

Interaction-aware ranking. After consolidation, candidate tubes are sorted in descending lexicographic order according to

$$\begin{aligned} \mathcal{R}_k = \big(& \mathbf{1}[N_{\mathrm{contact}} > 0],\ N_{\mathrm{contact}},\ N_{\mathrm{obs}}, \\ & N_{\mathrm{visible}},\ N_{\mathrm{pixels}} \big). \end{aligned} \tag{9}$$

Configuration	Setting
Backbone	Qwen2.5-VL-3B-Instruct
Model-input sampling rate	2 FPS
Fast-branch sampling rate	10 FPS
Input resolution	480×854
Maximum object slots	$K = 8$
Trace-mark dimension	256
Forcing layer	24th LLM block
Track margins	$m_+ = 0.6,\ m_- = 0.2$
Auxiliary loss weights	$\lambda_{\text{track}} = 0.005,\ \lambda_{\text{contact}} = 0.01$
LoRA rank / scale	$32/64$
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	0.01
Global batch size	32
Temperature	$\tau = 0.1$

Table 3: Implementation settings of TraceForce-HOI.

where N_{contact} is the accumulated number of contact pixels, N_{obs} is the number of InterFormer observation frames on the master timeline, N_{visible} is the number of visible Slow frames, and N_{pixels} is the accumulated tube area on those frames. Internal track identifiers break complete ties deterministically. We retain the top $K = 8$ persistent object tubes; remaining tubes are stored only as preprocessing metadata and are not provided as explicit trace inputs.

A.3 Active-Object and Contact Construction

Active-object attribution. At each Slow frame, the three InterFormer interacted-object roles are merged and decomposed into frame-local active-object components. Each raw component is compared with every retained persistent object mask using IoU and IoM. Candidate pairs satisfying

$$\text{IoU} \geq 0.10 \qquad \text{or} \qquad \text{IoM} \geq 0.50 \tag{10}$$

are sorted by $\max(\text{IoU}, \text{IoM})$ and matched one-to-one greedily. A matched component activates the corresponding complete persistent-instance mask rather than only its potentially partial semantic region:

$$\mathbf{A}_t = \bigcup_{k \in \mathcal{K}_t^{\text{active}}} \mathbf{O}_t^k. \tag{11}$$

If no raw component can be aligned to a persistent object, we apply a conservative contact-based fallback and select only the single object with the largest positive contact area. This avoids marking every object near a hand as active.

Localized contact evidence. Contact is reconstructed from the final hand and persistent-object masks rather than copied from the frame-wise semantic labels. Let $\mathbf{H}_t = \mathbf{H}_t^L \vee \mathbf{H}_t^R$. For object k, we compute

$$\begin{aligned} \mathbf{C}_t^k &= \text{Dilate}(\mathbf{H}_t, r) \wedge \text{Dilate}(\mathbf{O}_t^k, r), \\ \mathbf{C}_t &= \bigvee_{k=1}^{K} \mathbf{C}_t^k. \end{aligned} \tag{12}$$

We use a square kernel with base size five and scale the number of morphology iterations with the input width:

$$n_{\text{iter}} = \max\left(1, \left\lfloor \frac{W}{456} \right\rfloor\right), \qquad r = 2n_{\text{iter}}. \tag{13}$$

For the commonly used width $W = 640$, this gives $r = 2$, corresponding to a 5×5 neighborhood. The per-object masks $\mathbf{C}_t^k$ retain contact attribution and pixel counts, while $\mathbf{C}_t$ is the role-agnostic binary region used by the visual input. Contact is recomputed independently at every Slow frame and is not temporally propagated. It represents localized proximity-based interaction evidence within a small spatial neighborhood rather than an exact physical contact point.

User prompt.

<video>
The first video is the original first-person RGB video.

<video>
The second video is the HOI Mask Video. Color legend: cyan = left hand, magenta = right hand, distinct colors = different persistent object tracks, and white = contact region. The two videos are time-aligned and both receive learnable instance-aware trace marks. Use both videos to answer the question.

Carefully watch the first-person-view video and pay attention to the cause and sequence of events, object details and movements, and human actions and poses.

Question: {question}
{selection_instruction}

Options:
(A) {option_A}
(B) {option_B}
...

Answer format:
{answer_format}

Single-choice questions.

Choose only one option from the following list.

Answer format:
(A) <Description of Option A>

Multi-select Objects questions.

Choose all options that apply from the following list.

Answer format:
(A) <Description of Option A>
(C) <Description of Option C>

Table 4: Prompt template and answer formats used for HanDyVQA.

B Additional Implementation Details

B.1 Trace Construction and HOI Mask Rendering

Trace construction. We use InterFormer SU et al. (2026) for dense frame-level hand–object parsing in the Fast branch and SAM3 Carion et al. (2026) for sparse mask propagation in the Slow branch. Both models are kept frozen and are applied offline to construct the Persistent Interaction Traces. The resulting hand and object masks are retained only on the Slow timeline and are temporally aligned with the visual frames provided to the MLLM.

HOI Mask Video. The HOI Mask Video is rendered from the persistent hand and object masks. The left and right hands are represented by cyan and magenta, respectively, while the contact region is shown in white. Each persistent object track is assigned a distinct color that remains unchanged throughout the video. The object-to-color assignment is permuted on a per-sample basis to prevent the model from associating a particular color with a fixed track rank.

The active-object state is not rendered explicitly because changing the appearance of an object across active and inactive frames may weaken its temporal identity. Instead, active-object information is encoded through the instance-aware trace marks. Objects are rendered first, followed by the two hands and the contact region, such that localized contact evidence remains visible. The RGB video and the HOI Mask Video use identical frame indices, spatial resolution, and geometric preprocessing.

B.2 Trace Encoding and Contact-Aware Trace Forcing

We implement TraceForce-HOI using Qwen2.5-VL-3B-Instruct Bai et al. (2025b). The RGB video and the HOI Mask Video are processed by a shared frozen visual encoder. We use at most $K = 8$ persistent object slots, and each learnable trace mark has a dimension of 256. Object tracks are randomly assigned to the available slot embeddings during training, preventing the model from exploiting a fixed ordering of object instances.

The trace marks are projected to the visual-token dimension and residually added to the spatially aligned RGB and HOI-mask tokens. Average pooling is used to map the hand, object, and active-object masks to the visual-token grid. Max pooling is used for the contact mask because contact evidence is spatially sparse and may otherwise be weakened by average pooling.

Table 5: Ablation of input representations. We separately remove the HOI mask video and trace marks while retaining both alignment objectives and their supervision masks.

	Objects	Action	Process	Location	State	Parts	Avg5.
w/o mask video	66.5	81.9	81.1	73.6	76.2	70.8	76.7
w/o trace marks	65.3	84.4	84.5	76.8	80.1	72.3	79.6
Full method	**66.6**	**84.6**	**86.0**	**76.8**	**80.0**	**73.3**	**80.1**

Forcing implementation. Contact-Aware Trace Forcing is applied to the RGB visual-token hidden states extracted from the 24th language-model block. The selected hidden states are reshaped according to the temporal and spatial visual-token grid. Object, contact, and background representations are then obtained using the trace-derived token weights, followed by projection and ℓ_2 normalization.

For object-instance pooling, we retain only tokens with unambiguous instance ownership. A frame does not contribute to $\mathcal{L}_{\text{track}}$ when no valid historical observation of the corresponding object is available. Similarly, a frame does not contribute to $\mathcal{L}_{\text{contact}}$ when it contains no valid contact evidence.

Loss configuration. For Causal Track Alignment, the positive and negative margins are set to $m_+ = 0.6$ and $m_- = 0.2$, respectively. The auxiliary loss weights are $\lambda_{\text{track}} = 0.005$ and $\lambda_{\text{contact}} = 0.01$. Historical object representations used as alignment targets are detached from gradient computation. The forcing projection heads and auxiliary objectives are used only during training and are removed at inference.

B.3 Training Details

The model-input videos are sampled at 2 FPS, while the Fast trace-identification branch operates at 10 FPS. The input resolution is 480×854. We fine-tune the language model using LoRA with rank 32 and scaling factor 64. The model is optimized using AdamW with an initial learning rate of 1×10^{-4}, a cosine learning-rate schedule, weight decay of 0.01, and a global batch size of 32. The language-modeling loss is applied only to the answer tokens. All experiments are conducted on a single NVIDIA A800 80GB GPU.

B.4 Prompt and Answer Formatting

The prompt describes the two visual streams, their temporal alignment, and the semantic meaning of the colors in the HOI Mask Video. No additional textual tokens are introduced for the trace marks; all learnable marks are injected directly into the visual-token representations. Questions in the *Objects* category are formulated as multi-choice questions, while all remaining categories use a single-choice format. The original order of the answer options is preserved.

The letters in the multi-choice example illustrate the expected output structure and do not indicate fixed answers. The target response contains no additional explanation. For a single-choice question, the target consists of the selected option letter followed by its complete option text. For a multi-choice question, all correct options are listed in their original dataset order, with one option on each line.

C Ablation Study

We complement the ablation study with a finer-grained analysis of input representations, interaction cues, and alignment objectives. These experiments examine the contribution of each component and distinguish the effects of explicit input cues from those of training-time supervision.

Table 6: Ablation of explicit interaction cues. Each variant removes the specified input cue while retaining both alignment objectives and their supervision masks. Identity refers to track-specific colors and object-slot embeddings.

	Objects	Action	Process	Location	State	Parts	Avg5.
w/o identity	66.4	84.7	84.9	76.3	80.1	73.2	79.8
w/o active-object	65.9	84.0	85.1	76.2	79.7	73.0	79.6
w/o contact	64.2	84.6	84.7	75.5	79.7	73.0	79.5
Full method	**66.6**	**84.6**	**86.0**	**76.8**	**80.0**	**73.3**	**80.1**

Table 7: Ablation of alignment objectives. We separately disable Track Alignment and Contact Alignment while retaining all input representations and interaction cues.

	Objects	Action	Process	Location	State	Parts	Avg5.
w/o Track Alignment	66.0	84.4	84.5	76.8	79.6	72.9	79.6
w/o Contact Alignment	64.3	84.5	84.5	75.5	78.4	73.0	79.2
Full method	**66.6**	**84.6**	**86.0**	**76.8**	**80.0**	**73.3**	**80.1**

C.1 Input Representations

As shown in Table 5, removing the mask video causes pronounced degradation in Process and State while leaving object recognition nearly unchanged, suggesting that the structured visual stream helps organize how interactions unfold and alter objects over time. Trace marks further improve Objects, Process, and Parts, supporting their role in strengthening entity-aware grounding and fine-grained interaction reasoning. These results support encoding interaction traces in both pixel and latent spaces, with structured visual evidence enriched by explicit token-level annotations.

C.2 Interaction Cues

Table 6 reveals distinct effects of the three cues: removing identity encoding most affects Process, removing active-object cues produces broadly distributed declines, and removing contact cues particularly affects Objects and Location. These patterns suggest complementary roles in maintaining temporal correspondence, prioritizing relevant objects, and grounding interactions spatially.

C.3 Alignment Objectives

Table 7 shows that both alignment objectives contribute beyond the supplied interaction cues. Removing Track Alignment particularly affects Process, supporting its role in maintaining the temporal correspondence needed to interpret manipulation sequences. Removing Contact Alignment degrades Objects, Process, Location, and State, suggesting that contact-grounded supervision helps connect interacting entities with their actions and resulting changes. With all inputs held fixed, these results support explicitly regularizing intermediate representations to make persistent identities and contact relations more useful for downstream HOI reasoning.

D Computational Cost Analysis

The Fast and Slow branches of Fast-Slow Trace Identification take 1.48 and 11.02 seconds per video, respectively, with the latter including mask propagation, trace association, merging, and export. MLLM inference takes approximately 1.02 seconds per sample. Peak GPU memory is 3.2GB for the Fast branch, 5.4GB for the Slow branch, and 9.0GB for MLLM inference. Using VideoOrion's approximately 11.5s runtime as a contextual reference, we consider this overhead a reasonable trade-off for improved HOI understanding. Constructed traces can be cached across queries, and the training-only forcing heads introduce no inference-time computation. Advances in segmentation and propagation, together with implementation optimizations, could further reduce trace construction latency.

E Additional Qualitative Results

As shown in Figures 5, 6, 7, 8, 9 and 10, we provide additional qualitative examples covering the six question categories in HanDyVQA, including action, location, objects, object part, process, and state understanding. Together, these examples demonstrate the effectiveness of Persistent Interaction Traces across different aspects of fine-grained hand–object interaction understanding.



Question:
What is the person doing with his/her hands?

Qwen2.5-VL-SFT:
The person closes the blade cover of the cutting tool with his hand to secure the blade inside.

TraceForce-HOI:
The person opens the blade cover of the cutting tool with his left hand while holding the tool with his right hand.

Figure 5: Qualitative example for action understanding.



Question:
Where does the person drop the sieve bowl?

Qwen2.5-VL-SFT:
The person drops the sieve bowl onto the counter next to the pot.

TraceForce-HOI:
The person drops the sieve bowl on top of a pot.

Figure 6: Qualitative example for location understanding.



Question:
What object is used by the hands?

Qwen2.5-VL-SFT:
paint brush

TraceForce-HOI:
paint can, paint brush.

Figure 7: Qualitative example for object understanding.

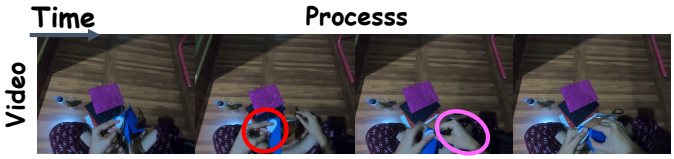


Question:
What part of the plant is plucked?

Qwen2.5-VL-SFT:
The leaves of the plant are plucked.

TraceForce-HOI:
The stem of the plant is plucked.

Figure 8: Qualitative example for object-part understanding.

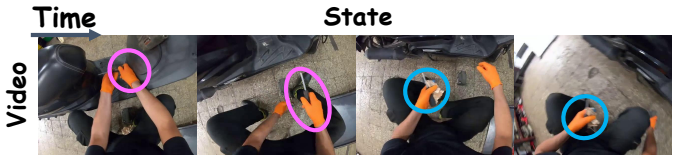


Question:
How does the person adjust the blue paper and the sky blue paper art in her left hand?

Qwen2.5-VL-SFT:
The person pinches the blue paper between her left hand fingers, using her right hand to carefully rotate it.

TraceForce-HOI:
The pereon holds the blue paper and the sky blue paper art between her left hand thumb and index finger and shifts it by adjusting the left hand fingers slightly forward with a support from the right hand.

Figure 9: Qualitative example for process understanding.



Question:
How did the state of the screwdriver change?

Qwen2.5-VL-SFT:
The screwdriver remained in the right hand after being twisted.

TraceForce-HOI:
The screwdriver was twisted for a moment and then transferred to his left hand.

Figure 10: Qualitative example for object-state understanding.