ANY2EGO: EGOCENTRIC VIEWPOINT TRANSLATION TAKING TEXTUAL MODALITY AS ANCHOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Egocentric video is the format in which embodied agents see the world and is difficult to collect, whereas third-person recordings of the same activities are abundant. Translating these recordings requires recovering both the geometry and the appearance of interactions from the first-person viewpoint. Spatial priors constrain geometry, but leave appearance and scene details underspecified; using a text prompt alone does not determine which details supplement these priors or where they should influence generation. We introduce **Any2Ego**, which learns a first-person textual anchor complementary to hand–object interaction (HOI) masks and controls its influence through HOI-conditioned spatial routing. A Qwen3-VL-8B converter learns structured descriptions through supervised caption distillation with sentence-level weighting, followed by generation-grounded reinforcement that rewards agreement between added sentences and rendered previews. A lightweight gate then uses HOI features to modulate caption cross-attention at each video token. On held-out Ego-Exo4D clips, Any2Ego substantially improves perceptual fidelity over our reproduced EgoExo-Gen, better preserves viewpoint-dependent scene structure, and produces interactions whose appearance and temporal evolution more closely match the target egocentric videos.[1]

1 INTRODUCTION

Egocentric video is the format in which embodied agents see the world, with the hands inside the field of view, a viewpoint that moves with the head, and heavy self-occlusion of the manipulated object (Grauman et al., 2024). Ego4D (Grauman et al., 2022) and EPIC-KITCHENS-100 (Damen et al., 2022) have expanded first-person recordings of daily activities, while Ego-Exo4D provides synchronized observations across viewpoints. Yet collecting additional egocentric footage still requires a camera wearer, whereas third-person recordings of the same activities exist in quantity. Egocentric viewpoint translation, which synthesizes the first-person video of a manipulation task from a third-person observation of it, therefore turns a plentiful resource into a scarce one. The value of such data rests entirely on whether the generated video preserves the interaction that the demonstration shows.

This translation involves more than a change of camera pose. The performer's body occludes parts of the workspace, and the egocentric view emphasizes the hands and manipulated objects, making dense cross-view correspondence difficult. Two complementary questions arise: where does the interaction occur in the target view, and how should its appearance and surrounding scene be described? In image synthesis, ControlNet (Zhang et al., 2023) and T2I-Adapter (Mou et al., 2024) show how spatial controls can complement text prompts. For viewpoint translation, the spatial condition must additionally be predicted in the target view, and the description must capture that view's interaction and scene. Spatial conditions address the first question, while language can explicitly represent the second. Their usefulness depends on whether the description adds information to the spatial condition and whether the generator applies it in the right places.

Existing approaches establish the value of spatial conditioning. EgoExo-Gen (Xu et al., 2025) transfers HOI masks across views, while EgoX (Kang et al., 2025) uses target-camera geometry to guide viewpoint translation. General backbones such as Stable Video Diffusion (Blattmann et al., 2023),

[1] Code is available in the supplementary materials.

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.

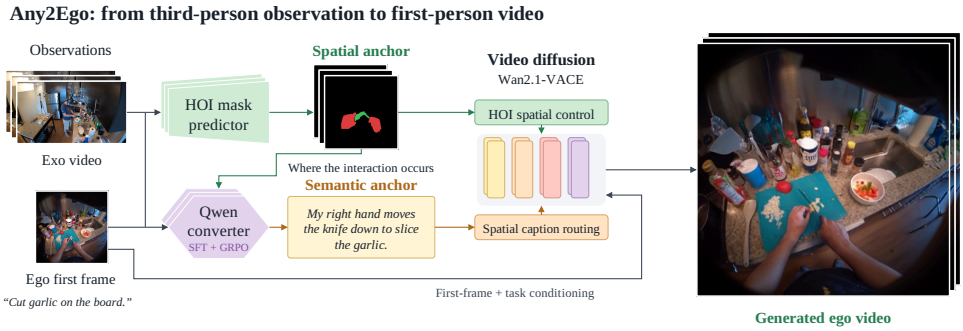


Figure 1: Any2Ego combines a spatial interaction anchor with a learned first-person textual anchor. The converter supplies complementary descriptions, and HOI-conditioned spatial routing modulates their influence on video generation.

ConsistI2V (Ren et al., 2024), Wan (Wan et al., 2025), and VACE (Jiang et al., 2025) also support visual or textual conditioning. Spatial constraints leave appearance and scene details open, while an available text prompt need not supply the missing first-person cues. This motivates our question: how can a first-person description be *learned to complement an interaction prior*, and where should it influence generation? A generic description may repeat visible content; a shared injection strength also couples its influence across regions with different conditioning needs.

We introduce **Any2Ego**, as illustrated in Figure 1. First, we follow EgoExo-Gen's cross-view memory-attention design to predict an egocentric HOI mask sequence from the exocentric video and the ego first frame with its mask. Next, we train a Qwen3-VL-8B converter using supervised caption distillation with sentence-level weighting (SFT), followed by generation-grounded reinforcement learning (GRPO). Given these visual inputs, the predicted masks, and task narration, it produces a structured first-person caption describing hand–object interactions, scene, motion, and lighting. Finally, we inject the predicted masks through Wan2.1-VACE's spatial-control branch and the caption through an added cross-attention branch. HOI-conditioned spatial routing controls the caption's contribution at each video token. Together with the ego first frame, these two conditions guide the generator to synthesize an egocentric video of the observed activity.

Experiments on Ego-Exo4D show that caption injection and spatial routing improve frame-level consistency and overall video quality. Our contributions are:

- **Learning Complementary Textual Anchors:** We develop a first-person converter that turns exocentric observations and initial ego conditions into structured descriptions for video generation. Supervised distillation and generation feedback connect caption production to the downstream generator, providing a learned textual condition alongside HOI masks.
- **HOI-Conditioned Caption Routing:** We introduce a lightweight per-token gate that uses spatial-control features to modulate the added caption response. A fixed-response analysis characterizes shared-scaling error for localized target corrections, motivating spatial modulation while keeping the spatial backbone frozen.
- **Consistent Empirical Gains:** Any2Ego improves frame-level fidelity over our reproduced EgoExo-Gen, increasing SSIM by 8.6% and PSNR by 5.8%, while reducing LPIPS by 25.8% and FVD by 21.8%. Our 8B converter achieves a 29.3% relative gain in composite score over the strongest evaluated off-the-shelf VLM; matched routing and anchor-content ablations support the textual condition and its spatial use.

2 Related Work

Egocentric Video Generation. EgoExo-Gen (Xu et al., 2025) predicts egocentric HOI masks with cross-view memory attention and conditions a diffusion model on them. EgoX (Kang et al., 2025) relies on externally supplied camera intrinsics and target ego-pose trajectories for geometric guidance. It uses annotated ego poses on Ego-Exo4D and manually selected camera extrinsics for in-the-wild examples (their Appendix F.2). The visual inputs used to construct its evaluation-time

exo/ego prompts are not fully specified (their Appendix F.4 and Table 6). These additional pose requirements and the incomplete prompt-input specification preclude a reproducible comparison under our input setting, so we exclude EgoX from the matched-input evaluation. Any2Ego uses the ego first frame without a target-pose trajectory, predicts the HOI anchor, and learns a complementary first-person caption whose contribution is modulated by HOI features.

Controllable Video Diffusion. Stable Video Diffusion (Blattmann et al., 2023) and ConsistI2V (Ren et al., 2024) animate a single image, while CogVideoX (Yang et al., 2024) and LTX-Video (HaCohen et al., 2024) support text-conditioned video generation. Wan (Wan et al., 2025) and VACE (Jiang et al., 2025) further support visual controls, including masks and reference frames. Any2Ego builds on these conditioning capabilities: it retains the native narration path and spatial-control backbone, and learns a separate caption path whose contribution is modulated by HOI features. This design allows the learned textual anchor and its routing to be evaluated on a shared spatially conditioned generator.

3 ANY2EGO

Any2Ego combines a predicted spatial anchor with a learned first-person textual anchor. The converter learns which descriptive content complements the input conditions, and the routing module controls its contribution to the video representation. We first define the inputs and spatial predictor, then describe caption learning, HOI-conditioned injection, and the evaluation calibration.

3.1 FRAMEWORK

Let $X = \{X_t\}_{t=0}^{T-1}$ denote the exocentric video, S its HOI masks, and n the task narration. We use the first-frame-conditioned setting: I_0 is the first RGB frame of the paired ego clip, and M_0 is its cached three-class HOI mask, obtained with the same hands23 detector used for S (Cheng et al., 2023). The first frame supplies initial appearance and M_0 initializes the spatial predictor. Any2Ego comprises a cross-view predictor f_{seg}, a caption converter f_{cap}, and a Wan2.1-VACE-1.3B generator f_{gen} (Wan et al., 2025; Jiang et al., 2025):

$$\widehat{M} = f_{\text{seg}}(X, S, I_0, M_0), \qquad c \sim f_{\text{cap}}(\cdot \mid X, I_0, \widehat{M}, n), \qquad \widehat{Y} = f_{\text{gen}}(I_0, \widehat{M}, n, c), \tag{1}$$

where $\widehat{M}_0 = M_0$ and f_{seg} predicts the masks for $t \geq 1$. The converter produces caption c from the exocentric observation, initial ego view, mask sequence, and narration; the generator uses these anchors to produce $\widehat{Y}$, as illustrated in Figure 2a.

3.2 INTERACTION-GROUNDED SPATIAL ANCHOR

To predict the subsequent interaction geometry from the initial condition, we implement f_{seg} following the cross-view memory-attention formulation of EgoExo-Gen (Xu et al., 2025). A shared ResNet-18 (He et al., 2016) encoder embeds both views, a mask-conditioned value encoder fuses the egocentric first-frame condition, cross-view memory attention lets each exocentric query retrieve the matching egocentric appearance, and a decoder predicts background, hand, and object labels. Our primary segmentation loss combines class-weighted cross-entropy with a soft Dice term on the two foreground classes,

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Dice}}, \tag{2}$$

where $\mathcal{L}_{\text{CE}}$ uses weights 1:5:5 over background, hand, and object. The predictor exports a three-class sequence in the format expected by the generator's spatial-control branch.

3.3 FIRST-PERSON TEXTUAL ANCHOR

To supply the semantics that a mask cannot express, f_{cap} converts the visual conditions into a structured first-person caption with six fields, namely `HANDS`, `OBJECT`, `CONTACT`, `SCENE`, `MOTION`, and `LIGHTING`. We build it from Qwen3-VL-8B-Instruct (Bai et al., 2025), freeze the visual encoder and the aligner, and adapt the language model with rank-16 LoRA (Hu et al., 2022). Its input $u = (X, I_0, \widehat{M}, n, \iota)$ collects the exocentric video, the egocentric first frame, the predicted masks, the narration, and a viewpoint instruction ι. The difficulty is that a capable vision-language model

already describes much of what the masks and the first frame make visible, so imitating a teacher caption directly spends most of its gradient on content the generator already sees, which the two training stages below address.

Supervised Caption Distillation. We use Qwen3.8-27B as the teacher to describe the paired training ego video, yielding offline targets $c^\star$. The student learns these targets from u, with the ego first frame as its only ego RGB input. This sequence-level distillation (Kim and Rush, 2016) transfers target-view supervision into a converter usable at inference. A baseline caption b uses only the initial ego frame, HOI masks, and narration. We apply heuristic sentence weights using similarity to b and lexical cues for actions or newly visible content. Highly similar sentences receive weight 0.3, lower-similarity sentences passing the cue filter receive 2.5, and the remainder receive 1.0 (Appendix A). This operational proxy prioritizes selected descriptions rather than measuring information gain. For a target of length L, the weighted objective is

$$\mathcal{L}_{\text{cap}}(\theta) = -\mathbb{E}_{(u,c^\star)}\left[\frac{1}{L}\sum_{j=1}^{L} w_j \log \pi_\theta\big(c_j^\star \mid u, c_{<j}^\star\big)\right], \tag{3}$$

where π_θ is the converter policy with trainable LoRA parameters θ. Reweighting leaves the target text unchanged and only redistributes gradient mass, so every target field remains supervised while selected sentences receive greater emphasis.

Generation-grounded Caption Reinforcement. To align the converter with what the generator can actually use, we continue from the distilled checkpoint with GRPO (Shao et al., 2024) and define a reward on rendered video instead of on text similarity. For each input the rollout policy samples $G = 8$ candidates, a frozen generator renders a preview $\widehat{Y}_i$ for candidate c_i, and the preview is cropped to the union bounding box of the HOI masks so the reward is measured where the interaction is. Let $\mathcal{S}_i$ be the new sentences of c_i, C_M the crop operator, and e_I, e_T unit-normalized CLIP (Radford et al., 2021) embeddings. With F preview frames, the reward is

$$r_i = \frac{1}{F|\mathcal{S}_i|}\sum_{t=1}^{F}\sum_{s\in\mathcal{S}_i} e_I\Big(C_M(\widehat{Y}_{i,t})\Big)^\top e_T(s), \tag{4}$$

where the candidate is scored on its full text when no new sentence is identified. We optimize Eq. 4 with the clipped GRPO objective under group-normalized advantages and a KL term to a fixed reference policy, updating only the converter LoRA while the generator and CLIP stay frozen. Appendix A specifies how the preview generator is trained and frozen; Figure 2b summarizes converter training.

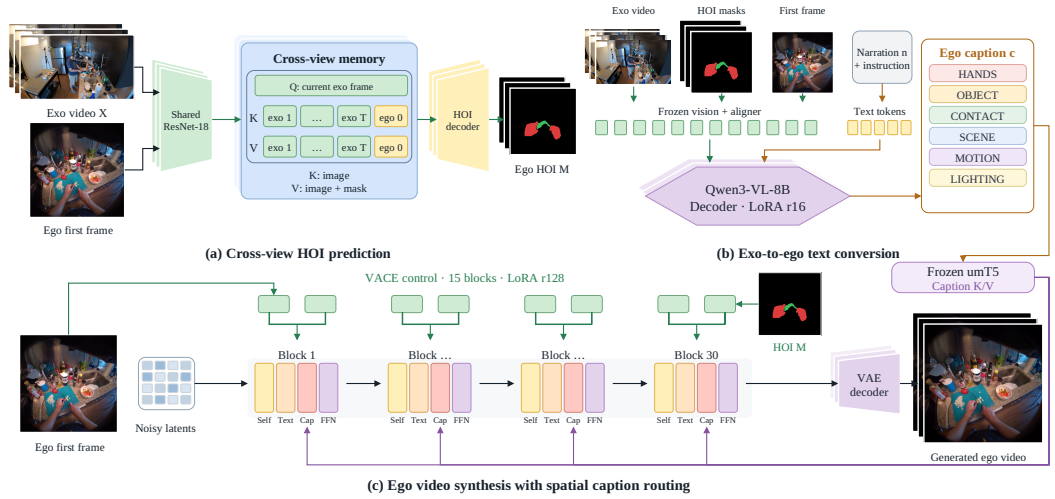
3.4 HOI-Conditioned Spatial Routing

To let the textual anchor act where its content applies, we design the injection path so its strength varies across the frame. The predicted masks $\widehat{M}$ enter Wan2.1-VACE through its 15-block control branch, while I_0 provides the image condition and n uses the native text path. The caption c is encoded by a frozen umT5 into E_c, and we attach a caption cross-attention branch to each of the 30 main blocks, reusing the video queries $Q^{(\ell)}$. Each branch first maps E_c to the block width through a learned affine input projection $\widetilde{E}_c^{(\ell)} = P_c^{(\ell)}(E_c)$, then forms caption keys and values,

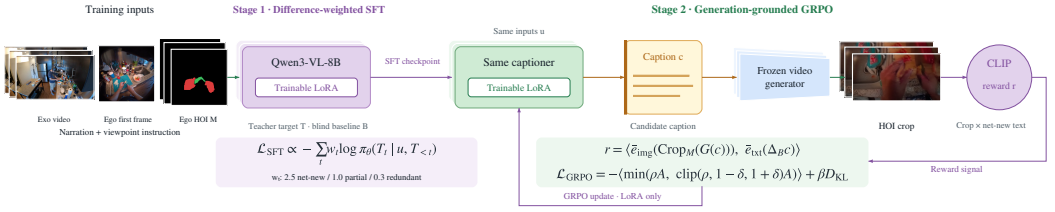
$$A_c^{(\ell)} = \text{softmax}\left(\frac{Q^{(\ell)}\left(K_c^{(\ell)}\right)^\top}{\sqrt{d}}\right)V_c^{(\ell)}, \tag{5}$$

where d is the head dimension. Suppressing block indices and affine biases, $K_c = \text{RMSNorm}(\widetilde{E}_c W_{K,c})$ and $V_c = \widetilde{E}_c W_{V,c}$: the separate learned K/V projections follow input adaptation, and the keys are normalized before attention. The caption response is scaled and added to the native attention before the frozen output projection,

$$O_i^{(\ell)} = W_O^{(\ell)}\left[A_{\text{native},i}^{(\ell)} + g^{(\ell)} s_i^{(\ell)} A_{c,i}^{(\ell)}\right], \qquad s_i^{(\ell)} = \sigma\left(\left(w^{(\ell)}\right)^\top \text{sg}(m_i) + b^{(\ell)}\right), \tag{6}$$



(a) Any2Ego pipeline.



(b) Training of the cross-view caption converter.

Figure 2: (a) Any2Ego produces the spatial anchor with a cross-view mask predictor and the textual anchor with a distilled and reinforcement-refined converter, then combines them in a mask-conditioned backbone whose caption branch is gated per video token by the dense spatial features of the control branch. (b) Supervised distillation uses heuristic sentence weights based on baseline similarity and action/reveal cues, and generation-grounded reinforcement scores each candidate by agreement between its rendered preview and new sentences within the interaction region.

where m_i is the pre-block HOI feature of the control branch aligned with video token i, sg stops gradients, and σ is the sigmoid. Setting $s_i^{(\ell)} = 1$ recovers global injection, which scales caption responses at all video tokens by the same learned per-block scalar $g^{(\ell)}$. Spatial routing additionally modulates this strength with HOI-conditioned $s_i^{(\ell)}$; Section 5.2 compares the two under matched conditions. Setting $g^{(\ell)} = 0$ removes the caption. We initialize $g^{(\ell)} = 0.1$ and $w^{(\ell)} = b^{(\ell)} = 0$, so routing starts from the neutral factor 0.5 while the spatial branch remains frozen.

Spatial Modulation of Caption Responses. Consider one block at one denoising step. Fix its projected caption responses $a_i = W_O A_{c,i}$ and denote the target local corrections by r_i. The response-fitting objectives are

$$\mathcal{E}_{\text{global}}(g) = \sum_i \|r_i - g a_i\|_2^2, \qquad \mathcal{E}_{\text{route}}(g, s) = \sum_i \|r_i - g s_i a_i\|_2^2. \tag{7}$$

For a localized target, let $r_i = \kappa a_i$ on $\mathcal{F}$ and $r_i = 0$ elsewhere, with $\kappa > 0$. Define response energies $E_{\mathcal{F}} = \sum_{i \in \mathcal{F}} \|a_i\|_2^2$ and $E_{\overline{\mathcal{F}}} = \sum_{i \notin \mathcal{F}} \|a_i\|_2^2$. For nonzero total energy, the optimal global gain and residual are

$$g^* = \kappa \frac{E_{\mathcal{F}}}{E_{\mathcal{F}} + E_{\overline{\mathcal{F}}}}, \qquad \mathcal{E}_{\text{global}}(g^*) = \kappa^2 \frac{E_{\mathcal{F}} E_{\overline{\mathcal{F}}}}{E_{\mathcal{F}} + E_{\overline{\mathcal{F}}}}, \tag{8}$$

Equation 8 characterizes the shared-scaling trade-off for these fixed responses. An ideal spatial factor of 1 on $\mathcal{F}$ and 0 elsewhere gives zero residual at $g = \kappa$, motivating control over response amplitude at each location.

Our implementation factorizes this control: $A_{c,i}^{(\ell)}$ depends on caption tokens and video queries, while $s_i^{(\ell)}$ supplies an HOI-conditioned gain. During training, caption projections and key normalization change the responses a_i and the gate learns their spatial scaling, jointly optimized by Eq. 9. Their product couples caption-dependent content with interaction-conditioned modulation. Section 5.2 evaluates the resulting learned branches.

Two-stage Generator Training. To keep any comparison between injection designs attributable to the injection itself, stage one fits rank-128 LoRA on the VACE control branch with the backbone frozen, which reproduces EgoExo-Gen's mask conditioning and is shared by all variants, and stage two freezes that base and trains only the caption projections, key normalization, and gates under caption dropout. Both stages use the flow-matching objective of the backbone (Lipman et al., 2023),

$$\mathcal{L}_{\text{gen}}(\psi) = \mathbb{E}_{z,\epsilon_z,\tau}\left[\omega(\tau)\,\text{MSE}(P_+ v_\psi(z_\tau, \tau; \mathcal{C}),\, P_+(\epsilon_z - z))\right], \tag{9}$$

where $z_\tau = (1-\tau)z + \tau\epsilon_z$, z is the encoded egocentric video, ϵ_z is Gaussian noise, $\omega(\tau)$ the scheduler weight, P_+ excludes the fixed first latent frame, and $\mathcal{C}$ collects $(I_0, \widehat{M}, n)$ in stage one and $(I_0, \widehat{M}, n, c)$ in stage two. Because stage two changes only the caption path, the reproduced baseline, the global gate, and the full model differ in exactly one design decision.

3.5 Freeze-Frame Calibration

To compare the injection designs above we need a metric that responds to generated motion, and on short horizons the common frame-level metrics do not. An egocentric clip of roughly 1.6 seconds retains much of the appearance of its given first frame, so a model earns frame-level credit for holding that frame in place. We make this measurable with a no-model control that repeats the given frame at all 13 positions, which has zero generated motion and therefore sets the score a method must exceed before its frame-level numbers carry information,

$$\widehat{Y}_t^{\text{freeze}} = I_0, \qquad t = 0, \ldots, 12. \tag{10}$$

This control reaches 0.601 SSIM at 21.88 dB PSNR on our evaluation pool, which already exceeds all six open-source generators we evaluate on both metrics, while its FVD is 357.2 and its residual correlation with the true motion is zero by construction. We therefore report FVD as the primary metric, read SSIM and PSNR against the calibration value, and add the residual-frame diagnostics of Section 5.4. It costs one frame copy and applies to any short-horizon egocentric benchmark; we include it in the comparison table below.

4 Experiments

4.1 Dataset and Metrics

Ego-Exo4D. We use synchronized exocentric/egocentric cooking recordings from Ego-Exo4D (Grauman et al., 2024), following EgoExo-Gen's annotation, conditioning, and split protocols (Xu et al., 2025). All comparisons use the same random N=1,000-clip pool, covering 281 takes, 47 settings, 13 tasks, and 10 institutions, with source takes disjoint from training (Appendices B and D).

Video metrics. We compare 13 aligned generated/ground-truth frames at 256×256. SSIM (Wang et al., 2004) measures structural similarity (↑), PSNR measures pixel reconstruction quality (↑, dB), and LPIPS (Zhang et al., 2018) measures learned perceptual distance (↓); these scores are averaged over frames and clips. FVD (Unterthiner et al., 2018) measures the distance between generated and real video distributions in I3D feature space (↓), pooled over all 1,000 clips. Following Section 3.5, FVD is primary; freeze-frame calibration contextualizes SSIM and PSNR.

Caption metrics. Qwen3.5-27B judges each generated caption against 13 ground-truth EGO frames. It scores HANDS, OBJECT, CONTACT, SCENE, MOTION, and LIGHTING, plus overall visual faithfulness, from 1 (incorrect or absent) to 5 (fully accurate). *Label-mean* averages the six field scores; *Overall* is the separate holistic score. *Comp.* starts from Overall and subtracts 1 for recording-equipment mentions and 0.3 for grammar failures. *Rule-0* is the equipment-mention rate (↓), and *Gram.* is the grammaticality pass rate (↑). We report caption-averaged scores (↑) and percentage rates.

Table 1: Caption quality on 1,000 clips (metrics: Section 4.1). Best values in **bold**.

Caption source	Params	Comp. ↑	Overall ↑	Label-mean ↑	Rule-0 ↓	Gram. ↑
Any2Ego converter	8B	**4.043**	**4.051**	**4.281**	0.0%	97.5%
Weighted SFT	8B	3.689	3.790	4.160	0.6%	68.2%
Qwen3-VL-32B	32B	3.128	3.131	3.529	0.0%	99.1%
Qwen3-VL-235B	235B	3.079	3.079	3.430	0.0%	**99.9%**
Qwen3.5-9B	9B	3.003	3.004	3.354	0.0%	99.8%
Qwen3-VL-8B	8B	2.829	2.831	3.164	0.0%	99.3%

Table 2: Any2Ego leads all compared open-source generators and EgoExo-Gen (reproduced on the same backbone) on all four metrics. Frame-level metrics saturate on short horizons, since the freeze-frame calibration outscores all six open-source generators on SSIM and PSNR while ranking near the bottom on FVD. Sees exo denotes third-person video input; best values are in **bold**.

Method	Sees exo	SSIM ↑	PSNR ↑	LPIPS ↓	FVD ↓
Freeze first frame (no model)	no	0.601	21.88	0.187	357.2
Wan2.1-VACE-1.3B	yes	0.178	10.14	0.711	1118.7
Wan2.1-Fun-1.3B-Control	yes	0.368	12.30	0.844	2361.1
Stable Video Diffusion	no	0.438	16.10	0.386	177.0
Wan2.2-TI2V-5B	no	0.553	20.74	0.225	127.6
LTX-Video	no	0.574	20.85	0.210	121.6
CogVideoX-5B-I2V	no	0.578	21.09	0.211	148.3
EgoExo-Gen (reproduced)	yes	0.608	21.06	0.194	42.6
Any2Ego (ours)	yes	**0.660**	**22.28**	**0.144**	**33.3**

4.2 Experimental Setup

Any2Ego uses Wan2.1-VACE-1.3B (Wan et al., 2025; Jiang et al., 2025), a Qwen3-VL-8B-Instruct converter, frozen umT5, and a ResNet-18 mask predictor. Weighted SFT and GRPO use ms-swift, with colocated vLLM for GRPO rollout. Our models generate 13 frames at 480×480 using 50 denoising steps and a fixed seed; baselines retain their released samplers. All training and inference use NVIDIA RTX PRO 6000 GPUs. Appendix B details shared frame loading.

4.3 Main Results

Capability on Cross-view Semantic Conversion. We evaluate caption faithfulness on the N=1,000 pool using the protocol of Section 4.1. Table 1 compares the weighted-SFT checkpoint and the full converter with off-the-shelf models from 8B to 235B under identical inputs. Our 8B converter achieves the highest composite score, 4.043. Appendix G shows four complete captions and their differences in hand visibility and scene grounding. The full converter attains 97.5% judged grammaticality, compared with 68.2% for weighted SFT. Field completeness differs from sentence completeness: SFT can retain all six labels yet produce incomplete clauses. Appendix C discusses the student's cross-view learning demands and generation feedback.

Capability on Egocentric Viewpoint Translation. We then evaluate the full pipeline to assess its *cross-view translation capability*, where each model receives its native conditioning inputs and must produce the 13 frames of the real egocentric continuation.

We compare Any2Ego with video-to-video generators that receive the raw exocentric video, with image-to-video generators that receive the first frame and the task prompt, each run from its released checkpoint, and with the freeze-frame calibration. Since EgoExo-Gen (Xu et al., 2025) is reproduced on the same backbone, the comparison against it isolates what the textual anchor adds. As shown in Table 2, Any2Ego leads on all four metrics and reaches 33.3 FVD, while EgoExo-Gen, the strongest prior method here, already separates itself from every open-source generator. More importantly, the freeze-frame row shows why the frame-level columns must be read with care, since a video with no generated motion outscores all six open-source generators on SSIM and PSNR while placing near the bottom on FVD. Frame-level agreement alone cannot explain this viewpoint translation.

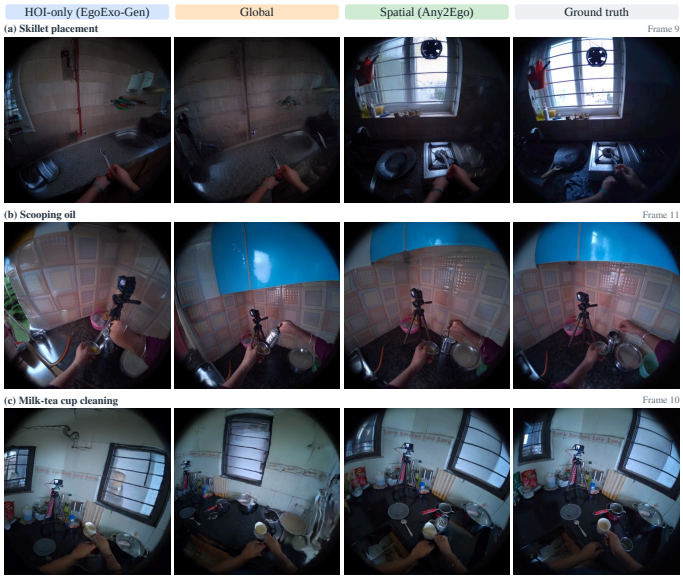


Figure 3: Any2Ego preserves scene structure unspecified by masks: (a) cooking-area layout during skillet placement (frame 9), (b) the overhead cabinet during oil scooping (frame 11), and (c) the window and background during milk-tea cup cleaning (frame 10).

Table 3: Comparison of three conditioning strategies on the same HOI-conditioned base. Global and Spatial (Any2Ego) inject the same captions through a global gate and HOI-conditioned spatial gates, respectively. Best values are in **bold**.

Method	**SSIM** ↑	**PSNR** ↑	**LPIPS** ↓	**FVD** ↓
HOI-only (EgoExo-Gen)	0.608	21.06	0.194	42.6
Global	0.636	21.74	0.163	37.8
Spatial (Any2Ego)	**0.660**	**22.28**	**0.144**	**33.3**

Comparison of Conditioning Strategies. We compare HOI-only (EgoExo-Gen), Global, and Spatial (Any2Ego) on the same clips and HOI-conditioned base. Global and Spatial use identical captions with global and spatial gates, respectively. Table 3 gives quantitative results; Figure 3 compares matched frames, with more cases in Appendix E.

Spatial (Any2Ego) better matches ground-truth framing of the cooking area, overhead cabinet, window, and background than HOI-only and Global. Section 5.2 analyzes both injection strategies.

5 Analysis

5.1 Role of a Dedicated Caption Branch

Table 4 compares a trained native text pathway (Native n), inference-time caption substitution (Replace n), and separate caption branches on 1,000 clips. Native n trains DiT LoRA with captions through the original narration path. Replace n keeps the trained HOI-only (EgoExo-Gen) base fixed. Both use the same GRPO captions as Global and Spatial (Any2Ego) at inference.

Table 4: Frame-level comparison. Native n: trained native text path. Replace n: caption substitution. Global/Spatial: dedicated branches. Best values in bold.

Pathway	**SSIM** ↑	**PSNR** ↑	**LPIPS** ↓
Native n	0.6470	22.195	0.1511
Replace n	0.624	21.67	0.178
Global	0.636	21.74	0.163
Any2Ego	**0.660**	**22.28**	**0.144**

Substitution improves SSIM from 0.608 to 0.624, showing that caption content itself is useful. Global improves all three frame-level metrics over substitution. Even after training the native text pathway, Spatial (Any2Ego) still achieves higher

SSIM and PSNR and lower LPIPS than this control, indicating better frame-level reconstruction fidelity.

Task narration guides the activity, while the detailed caption adds hand, object, and scene information. Separate projections learn to use these details while retaining the narration path; the spatial gain in Eq. 6 controls the caption response. Their combination gives the best frame-level results among the compared paths.

Relative to direct substitution, Spatial (Any2Ego) achieves paired-clip SSIM, PSNR, and LPIPS win rates of 82.7%, 70.2%, and 85.7%, with all three improving together on 67.9%. Wins mean strictly higher SSIM/PSNR or lower LPIPS. Section 5.2 analyzes spatial routing against Global.

5.2 Effect of Spatial Routing

Global and Spatial (Any2Ego) share the frozen HOI-only (EgoExo-Gen) base, captions, projection architecture, and training schedule; Spatial (Any2Ego) additionally learns the HOI-conditioned gate. Spatial (Any2Ego) improves all four metrics (Table 3) and the cooking-area layout, cabinet, and window (Figure 3). Caption-dependent attention determines the semantic response, while interaction-conditioned scaling adjusts its spatial strength. This control adds only 30 projection vectors and biases on existing HOI features. The gate reads dense control features at every video token, allowing caption responses to be modulated in background regions as well as around hands and objects. This explains the relevance of scene-level comparisons: the cabinet and window in Figure 3 test structure beyond the foreground mask, while the shared captions keep semantic content fixed. The masks provide structural context without specifying the appearance of those regions.

5.3 Ablation Study on Anchor Content

To test whether the generator uses both anchors, we keep the trained model fixed and randomize one condition at inference. Captions are replaced by text of equal length; masks are spatially randomized while preserving per-class pixel counts. In Table 5, caption randomization worsens every metric, with original conditions winning on 88.7% of clips; mask randomization causes larger degradations and original conditions win on 100%. Both anchors contribute: masks supply geometry, while captions provide smaller, consistent semantic refinements.

Table 5: Anchor ablations. Win: fraction of clips where original conditions give higher SSIM than randomized inputs.

Condition	SSIM ↑	LPIPS ↓	Win ↑
Original	**0.660**	**0.144**	
caption rand.	0.632	0.166	88.7%
mask rand.	0.423	0.504	100.0%

5.4 Analysis of Residual Motion Fidelity

We compare temporal changes after subtracting each video's own first frame. Residual correlation increases from 0.561 for HOI-only (EgoExo-Gen) to 0.614 for Global and 0.647 (Any2Ego), supporting temporal fidelity beyond retained appearance. Residual PSNR likewise improves from 19.80 to 20.47 and 21.17 dB, while the residual-magnitude ratio moves from 1.18 to 1.14 and 1.04 (target 1). Thus, Spatial improves agreement with observed temporal changes while bringing their amplitude closer to ground truth. Correlation and amplitude together distinguish matching the continuation from simply producing more change; these residuals include object motion, camera motion, and lighting variation. This complements FVD by testing paired temporal changes rather than aggregate video-feature distributions. Appendix F gives the protocol and full results (Table 7).

6 Conclusion

Any2Ego learns complementary first-person captions through distillation and rendered-preview feedback, and controls their contribution through HOI-guided gating with a fixed spatial backbone. Caption and routing comparisons show that learning the textual condition and controlling its spatial influence both improve egocentric viewpoint translation, with gains in scene structure and interaction fidelity.

AI USE STATEMENT

In this work, we used generative AI tools only for final-stage language polishing and grammar refinement of the manuscript. We have not used generative AI tools for generating research ideas, designing methods, conducting experiments, implementing code, analyzing data, interpreting results, or producing scientific claims, and other generative AI-assisted tasks requiring disclosure are not applicable to this work. We have reviewed all AI-assisted revisions to ensure their correctness and consistency with the authors' intended content. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of this work by providing detailed descriptions of the proposed method, experimental settings, implementation details, and evaluation protocols in the main paper and supplementary materials. The supplementary materials include the source code required to reproduce our experiments, and we will publicly release the code after the review process to further facilitate verification and future research. The content of this paper can be completely reproduced.

ETHICS STATEMENT

This work uses only the publicly documented Ego-Exo4D dataset of cooking activities, which is released for research use, and we introduce no new human data collection. The generative model studied here is scoped to short kitchen-activity clips, and we do not evaluate or recommend it for generating egocentric video of individuals without consent or in contexts that are sensitive to safety, privacy, or identity.

REFERENCES

S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025.

A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani, and R. Rombach. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. *arXiv preprint arXiv:2311.15127*, 2023.

T. Cheng, D. Shan, A. Hassen, R. Higgins, and D. Fouhey. Towards A Richer 2D Understanding of Hands at Scale. *NeurIPS*, 2023.

D. Damen, H. Doughty, G. M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision*, 130:33–55, 2022.

K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, et al. Ego4D: Around the World in 3,000 Hours of Egocentric Video. *CVPR*, 2022.

K. Grauman, et al. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. *CVPR*, 2024.

Y. HaCohen, N. Chiprut, B. Brazowski, D. Shalem, D. Moshe, E. Richardson, E. Levin, G. Shiran, N. Zabari, O. Gordon, et al. LTX-Video: Realtime Video Latent Diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *CVPR*, 2016.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-Rank Adaptation of Large Language Models. *ICLR*, 2022.

Z. Jiang, Z. Han, C. Mao, J. Zhang, Y. Pan, and Y. Liu. VACE: All-in-One Video Creation and Editing. *arXiv preprint arXiv:2503.07598*, 2025.

T. Kang, K. Kim, D. Kim, M. Park, J. Hyung, and J. Choo. EgoX: Egocentric Video Generation from a Single Exocentric Video. *arXiv preprint arXiv:2512.08269*, 2025.

Y. Kim and A. M. Rush. Sequence-Level Knowledge Distillation. *EMNLP*, pp. 1317–1327, 2016.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow Matching for Generative Modeling. *ICLR*, 2023.

C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan. T2I-Adapter: Learning Adapters to Dig Out More Controllable Ability for Text-to-Image Diffusion Models. *AAAI*, 2024.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning Transferable Visual Models From Natural Language Supervision. *ICML*, 2021.

W. Ren, H. Yang, G. Zhang, C. Wei, X. Du, W. Huang, and W. Chen. ConsistI2V: Enhancing Visual Consistency for Image-to-Video Generation. *arXiv preprint arXiv:2402.04324*, 2024.

Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, M. Zhang, Y. K. Li, Y. Wu, and D. Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024.

T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly. Towards Accurate Generative Models of Video: A New Metric & Challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Team Wan, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, et al. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025.

Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

J. Xu, Y. Huang, B. Pei, J. Hou, Q. Li, G. Chen, Y. Zhang, R. Feng, and W. Xie. EgoExo-Gen: Ego-centric Video Prediction by Watching Exo-centric Videos. *ICLR*, 2025.

Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng, et al. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv preprint arXiv:2408.06072*, 2024.

R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *CVPR*, 2018.

L. Zhang, A. Rao, and M. Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. *ICCV*, 2023.

A Training Configuration

Mask Predictor. We train f_{seg} on all 108,338 annotated clips at 256^2 working resolution with an ImageNet-initialized ResNet-18 encoder, using AdamW at learning rate 5×10^{-5} with distributed data parallelism for 20 epochs. The loss follows Eq. 2, and we retain the ego-to-exo query curriculum of the reference design. While the curriculum uses ground-truth ego queries, we additionally apply the same CE-plus-Dice loss, with unit weight, to the exo-query path conditioned on the initial ego frame and mask. This auxiliary training loss is intended to reduce reliance on ground-truth ego queries; it is disabled once the ego-query mixing weight reaches zero and is not used at inference.

Teacher Supervision. Qwen3.8-27B receives exocentric video, HOI masks, narration, and sampled frames across the paired training ego video. Its six-field text supervises a student whose ego RGB input is only the first frame. The baseline caption uses that frame, masks, and narration without exo video.

Sentence Weights. Within each field, we take the maximum of content-word Jaccard and character-sequence similarity to baseline sentences. A score of at least 0.50 gives weight 0.3. Otherwise, an action/reveal cue (e.g., opening, lifting, pouring, or contents) without a low-information cue (e.g., static or slight motion) gives 2.5; all remaining segments receive 1.0. Segment weights are broadcast to tokens and multiply cross-entropy; weight markers are removed before tokenization.

Caption Converter. SFT and GRPO use the same 32,558-clip common subset indexed in the supplementary code. SFT uses rank-16 LoRA with $\alpha = 32$. We first train a caption branch on the HOI-conditioned Wan2.1-VACE backbone using teacher captions, then freeze it for GRPO. GRPO starts from an SFT checkpoint, samples eight candidates per input, and updates rank-16 LoRA at learning rate 10^{-6}. CLIP scores generated interaction crops against added caption content (Eq. 4). Previews use 8 steps at 256^2, versus 50 at 480^2 for full generation; only the converter is updated.

Generator. All generator training, including the reward-preview model, uses the same 32,558 clip IDs as the caption converter. Final Global and Spatial branches train on GRPO captions. To separate the spatial path from the caption path, stage one trains rank-128 LoRA on the VACE control branch under bfloat16 with gradient checkpointing. Stage two keeps the backbone, the stage-one control LoRA, and umT5 frozen, and trains the caption projections, key normalization, and gates of Eq. 6, where spatial routing adds one token-wise projection on the detached HOI features. A single anisotropic resize to the target square is applied identically in training, inference, and evaluation.

B Evaluation Protocol

All methods are evaluated on the same fixed split of N=1,000 clips at 13 frames, where CogVideoX uses its native 30 steps and we retain the first 13 of its 49 output frames without temporal resampling. Per-clip PSNR, SSIM, and LPIPS are computed after resizing to 256×256, and FVD is pooled over the whole split with I3D features. The exo-conditioned baselines receive raw exocentric RGB as control input, the text-capable image-to-video baselines receive the egocentric first frame and the task prompt, and Stable Video Diffusion receives the first frame alone.

Coverage is counted by mapping each evaluated clip to its source take and counting distinct take, physical-setting, task, and institution entries in the Ego-Exo4D metadata, which yields the 281 takes, 47 settings, 13 tasks, and 10 institutions reported in Section 4.1. Each take contributes between 1 and 16 clips and task counts range from 3 to 240 clips, so coverage within cooking is not uniform.

C Per-field Caption Results

Table 6 shows that our converter leads on four fields. The full converter scores 4.17 and 4.18 on hands and objects, and 4.70 on scene, versus 4.09 for the weighted-SFT model. Appendix G illustrates hand-visibility and scene-grounding differences with complete captions.

Grammaticality. Cross-view distillation requires the 8B student to produce the teacher's first-person descriptions from exocentric observations and initial ego conditions. Beyond learning six field labels, it must jointly learn viewpoint translation, interaction reasoning, and coherent language generation. The gap between the student's task competence and the teacher targets can yield in-

Table 6: Per-field caption faithfulness on the fixed 1,000-clip pool. The full converter leads on four fields, including scene and lighting. Values are mean judge scores from 1 to 5 under the protocol of Table 1. Best values are in **bold**.

Caption source	Hands	Object	Contact	Scene	Motion	Lighting
Any2Ego converter	4.17	4.18	**4.06**	**4.70**	**3.95**	**4.63**
Weighted SFT	**4.19**	**4.20**	4.01	4.09	3.89	4.58
Qwen3-VL-32B	3.09	3.34	3.20	4.16	3.14	4.25
Qwen3-VL-235B	2.94	3.38	3.11	4.15	3.12	3.89
Qwen3.5-9B	3.22	3.43	3.31	3.52	3.22	3.42
Qwen3-VL-8B	2.87	3.15	2.89	3.68	2.93	3.47

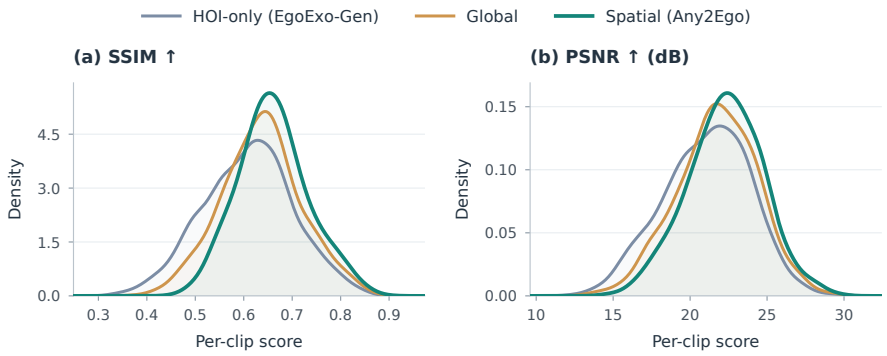


Figure 4: The textual anchor shifts the whole evaluation pool instead of a favorable subset, since per-clip (a) SSIM and (b) PSNR distributions move together across EgoExo-Gen, Any2Ego without spatial routing, and Any2Ego under a shared kernel bandwidth.

complete clauses even after the field structure is learned. GRPO then optimizes agreement between candidate descriptions and rendered interactions, providing indirect feedback for coherent, visually grounded expression. Following this refinement, judged grammaticality increases from 68.2% to 97.5% (Table 1).

D Per-clip Score Distributions

Mean scores can hide a small set of clips that carries an improvement, so we examine whether the gains hold across the pool. We compute per-clip scores over the 13 frames of every clip in the fixed split and compare the three Any2Ego variants under a shared kernel bandwidth. As shown in Figure 4, median SSIM moves from 0.613 to 0.638 and to 0.657 across EgoExo-Gen, Any2Ego without spatial routing, and Any2Ego, and median PSNR moves from 21.27 to 21.80 and to 22.35 dB, which tracks the means in Table 3 and indicates that the improvement is not driven by a tail. The distribution of spatial routing is close to symmetric, with a 10th to 90th percentile interval of $[0.566, 0.760]$ for SSIM and $[19.02, 25.22]$ dB for PSNR. Overall, the per-clip distributions show that the textual anchor shifts the whole pool instead of a favorable subset.

E Additional Qualitative Examples

As shown in Figure 5, the pot-placement and omelet cases differ mainly in stove layout and viewpoint, where spatial routing follows the ground-truth framing of the cooking area more closely, and the tea-bag case gives a closer view of the hands, the handled object, and the countertop. Each row compares full frames at the same time step across methods, and frame indices start at 0.

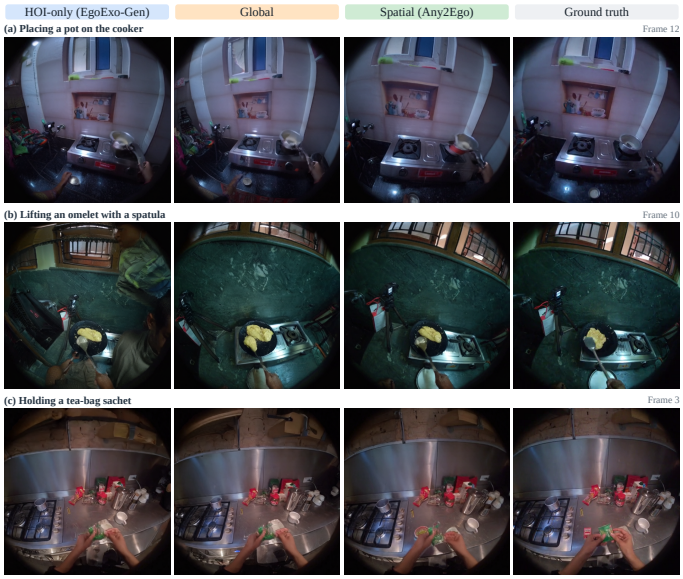


Figure 5: Spatial routing follows the ground-truth framing of the cooking area more closely on three further cases, including (a) placing a pot on the cooker at frame 12, (b) lifting an omelet with a spatula at frame 10, and (c) holding a tea-bag sachet at frame 3. Each row compares HOI-only (EgoExo-Gen), Global, Spatial (Any2Ego), and ground truth at the same time index, using the same four-column layout as Figure 3.

Table 7: Residual-frame analysis after subtracting each video's own first frame shows that Any2Ego generates real motion while image-to-video baselines mostly retain the given frame. ΔCorr is the residual correlation and ΔPSNR the residual fidelity, both higher is better, and the magnitude ratio has target 1. Best values are in **bold**.

Method	**ΔCorr** $\uparrow$	**ΔPSNR** $\uparrow$	**Mag. ratio** $(\rightarrow 1)$
Ground truth	*1.000*		*1.000*
Stable Video Diffusion	0.108	14.85	1.77
Wan2.2-TI2V-5B	0.221	18.45	0.73
LTX-Video	0.233	18.86	0.61
CogVideoX-5B-I2V	0.198	18.76	0.54
Wan2.1-VACE-1.3B	0.003	18.09	0.71
Wan2.1-Fun-1.3B-Control	0.319	10.35	5.28
Any2Ego	**0.647**	**21.17**	**1.04**
− spatial routing	0.614	20.47	1.14
− textual anchor	0.561	19.80	1.18

F Residual Temporal Fidelity

For each generated and ground-truth video, we subtract its own first frame from frames $t = 1, \ldots, 12$. We compare the resulting residuals using Pearson correlation (ΔCorr), residual PSNR (ΔPSNR), and the ratio of mean absolute residual magnitudes. Higher correlation and PSNR indicate closer agreement, while the magnitude ratio has target 1. The freeze-frame control has zero residual magnitude and is assigned correlation 0. These measures capture temporal appearance changes, including those caused by object motion, camera motion, and lighting.

As shown in Table 7, Any2Ego reaches 0.647 residual correlation and 21.17 dB residual PSNR, with a magnitude ratio of 1.04. Global and HOI-only (EgoExo-Gen) reach correlations of 0.614 and 0.561, respectively, while the strongest image-to-video baseline reaches 0.233. The results complement the frame-level metrics by assessing whether generated changes match the ground-truth continuation.



Figure 6: Shared visual evidence: the left hand ($t = 0$), spatula–omelet interaction ($t = 4$), and barred window ($t = 12$).

G Complete Caption Comparison

Four complete six-field ego captions for the same clip are shown below. Additional selected cases and full 13-frame visualizations are provided in the supplementary material.

GRPO Qwen3-VL-8B

HANDS: My right hand holds a spatula, positioned over the skillet; my left hand rests near the stove's control panel.
OBJECT: A spatula is held in my right hand, interacting with an omelet cooking in a black skillet.
CONTACT: The spatula touches the omelet, shifting it slightly within the skillet.
SCENE: The view shows a dark green countertop, a gas stove, and a window with bars in the background.
MOTION: My right hand moves the spatula to shift the omelet within the skillet, while my left hand remains steady near the stove's control panel.
LIGHTING: The scene is dimly lit, with shadows cast by the hands and objects, and reflections visible on the metallic surfaces of the stove.

Weighted SFT Qwen3-VL-8B

HANDS: My right hand holds a black spatula, positioned over the skillet. My left hand is visible near the edge of the stove, resting or stabilizing.
OBJECT: A yellowish-brown omelet sits inside a black skillet on a gas burner.
CONTACT: The spatula blade is inserted under the edge of the omelet, making contact to lift and shift it.
SCENE: A dark green marble countertop surrounds a silver gas stove. In the background, there are metal shelves filled with pots and pans, and a window with bars letting in daylight.
MOTION: The right hand moves the spatula slightly, nudging the omelet within the pan. The omelet shifts position slightly as the spatula pushes against it.
LIGHTING: Natural light comes from the window in the background, illuminating the scene. The foreground is slightly shadowed compared to the bright background.

Complete caption comparison (continued)

Qwen3.5-9B Off-the-shelf

HANDS: My right hand enters the view holding a spatula. My left hand is not visible.
OBJECT: I see a round, yellowish omelet cooking in a black skillet.
CONTACT: The spatula in my right hand touches the omelet, moving it around the pan.
SCENE: The skillet sits on a silver gas stove. The countertop is dark green marble.
MOTION: I use the spatula to slide the omelet across the surface of the skillet.
LIGHTING: The kitchen is brightly lit, casting clear shadows from the spatula and the omelet.

Qwen3-VL-235B A22B-Instruct

HANDS: My right hand holds the spatula handle, fingers curled around it, sleeve visible at the wrist; my left hand is not in view.
OBJECT: The omelet is round, pale yellow with browned edges, resting in a black skillet; as I move it, the skillet's interior and the omelet's underside briefly appear.
CONTACT: My right hand grips the spatula firmly; the spatula's edge slides under the omelet and nudges it forward without lifting.
SCENE: The dark green countertop surrounds the silver gas stove; the skillet sits centered on the burner, with a window and wall visible in the upper background.
MOTION: I slide the spatula under the omelet from the right, pushing it slightly forward in the skillet; the omelet shifts position while the skillet remains stationary.
LIGHTING: Overhead light casts a soft shadow from my hand onto the skillet; the omelet's surface reflects a slight sheen, and the countertop remains evenly lit.

For this clip, GRPO gives the most faithful overall visual account, followed by SFT. GRPO correctly distinguishes the two hands: the right hand manipulates the spatula, while the left stays near the stove controls. Its contact and motion fields describe the same small displacement of the omelet, linking the acting hand, tool, and object consistently. The surrounding gas stove, dark green surfaces, barred window, and dim work area also match the video. Its advantage is therefore agreement across the interaction and scene fields, rather than a longer description.

SFT captures both hands, the spatula–omelet contact, and the small pushing motion, so it retains the main interaction more faithfully than the off-the-shelf models. Its weakness is scene completion: it adds "metal shelves filled with pots and pans", which are absent from the visible background across the sequence. This plausible kitchen detail makes the caption less faithful even though the central action is correct. SFT thus provides a useful interaction description but introduces an unsupported scene element that GRPO avoids.

Both off-the-shelf models recognize the omelet, skillet, and spatula action, but incorrectly state that the left hand is not visible. This is a concrete viewpoint error: the target video shows the left hand beside the stove throughout the sequence. Qwen3.5-9B additionally describes a brightly lit kitchen, missing the shadowed work area. Qwen3-VL-235B provides more detailed object and motion descriptions, yet still omits a visible participant in the interaction. Greater descriptive detail does not resolve that error.

The comparison shows why filling all six fields is insufficient: the fields must form a mutually consistent account of the observed first-person view. GRPO best preserves that consistency here, SFT retains the core interaction but adds a scene hallucination, and the baselines misrepresent hand visibility. This is the behavior targeted by generation-grounded refinement: descriptions that agree with the visual interaction and its surroundings.