CrossScope: A Role-Asymmetric World Model for Joint Dual-Scope Surgical Video Prediction

Anonymous submission

Abstract

Visual world models typically learn future dynamics from a single observation stream, limiting their ability to model cooperative systems with multiple independently moving observers. We investigate this challenge in Mother–Child endoscopic retrograde cholangiopancreatography (ERCP), where two flexible scopes provide complementary yet role-dependent views without a calibrated stereo relationship. Unlike conventional multi-view fusion that assumes symmetric information exchange, we formulate **role-asymmetric dual-scope future prediction**, where cross-view evidence is selectively transferred according to the prediction target and its underlying spatial requirements. We propose **CrossScope**, a dual-stream surgical world model that preserves view-specific experts while enabling target-specific evidence routing through geometry-guided residual interactions. CrossScope learns two complementary communication directions: geometric motion cues from the Mother view guide Child-view future dynamics, while pose-aligned Child appearance supports Mother-view prediction only when valid spatial correspondence is established. This design allows each scope to contribute task-relevant evidence without compromising its view-specific representation. To evaluate this problem, we establish a paired dual-scope benchmark comprising synchronized phantom and real-world ERCP episodes, with evaluations assessing visual fidelity, structural preservation, target localization, and motion consistency. Experiments demonstrate that CrossScope consistently outperforms strong surgical video generation baselines, validating the importance of role-aware evidence routing for multi-observer visual world modeling.

Introduction

Visual world models aim to learn observation dynamics and predict future states of an environment from partial observations (Ha and Schmidhuber 2018; Bruce et al. 2024). Recent advances in surgical world models have demonstrated promising progress in forecasting surgical video dynamics, including long-horizon predictions that capture instrument–tissue interactions (Pan et al. 2026). However, existing approaches primarily focus on a single visual stream, implicitly assuming that future dynamics can be inferred from one observer. This assumption becomes insufficient for cooperative procedures involving multiple independently controlled views, where different observers may provide complementary but non-equivalent evidence.

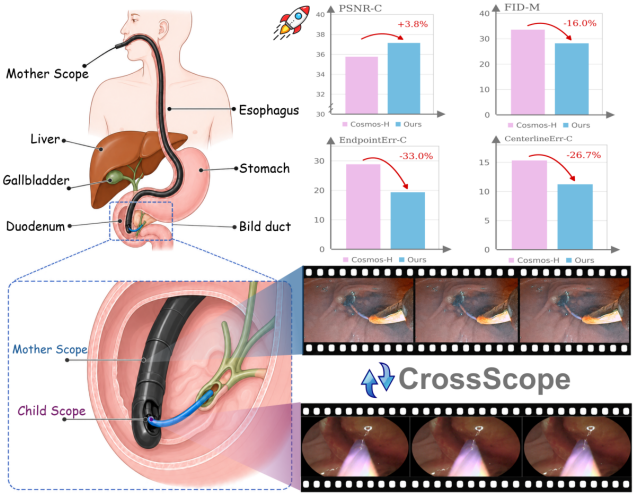

Figure 1: CrossScope for role-asymmetric dual-scope future prediction. Mother and Child scopes provide complementary wide- and near-field observations; directional M2C and C2M routes jointly forecast both streams. The plots show representative phantom-benchmark gains over Cosmos-H-Surgical.

Mother–Child endoscopic retrograde cholangiopancreatography (ERCP) presents a distinct dual-scope prediction problem. The wide-view Mother duodenoscope provides global spatial context and visualizes the Child-scope tip, whereas the near-field Child cholangioscope advances through the Mother into narrow bile ducts to reveal local anatomy (Parsi 2011). The two flexible scopes are independently controlled rather than fixed as a calibrated stereo pair. Their relative pose, image scale, spatial overlap, and visible content therefore vary throughout a procedure. The challenge is therefore not to treat the two views as interchangeable inputs, but to determine which cross-scope evidence can condition each prediction target. We formulate this setting as *role-asymmetric dual-scope future prediction*: jointly forecasting both streams from current anchors and strictly causal

Child history.

Yet existing surgical video methods do not model role-dependent exchange across independently moving scopes. Conditional generators enrich one target with visual, object-level, structured-scene, or action cues (Li et al. 2024b; Iliash et al. 2024; Chen et al. 2025; Biagini, Navab, and Farshad 2025; Yeganeh et al. 2024; Sivakumar et al. 2025; Koju et al. 2025), while SAW and Routed Visual Control add trajectories or image-aligned kinematics (Rapuri et al. 2026; Li et al. 2026). These controls improve target-specific prediction but do not specify what evidence should transfer from another independently moving view. Multi-view reconstruction assumes shared geometry or jointly optimized poses (Özsoy et al. 2022; Li et al. 2024a; Huang et al. 2024; Stilz et al. 2026). Temporal memory retains evidence within one stream or shared representation, while surgical foundation models focus on representation learning (Cheng and Schwing 2022; Xiao et al. 2025; Lu et al. 2026; Wu et al. 2026; Yang et al. 2026). These approaches do not match Mother–Child ERCP: Mother-plane geometry can guide Child motion, whereas Child appearance can inform Mother prediction only after causal observation and spatial alignment.

To address this challenge, we introduce **CrossScope**, a role-asymmetric world model for joint dual-scope surgical video prediction. Instead of treating multiple views as symmetric observations, CrossScope models cross-scope interaction through target-dependent evidence allocation, allowing each observer to contribute information according to its predictive role. We further establish a paired dual-scope benchmark with synchronized phantom and real-world ERCP episodes to evaluate future prediction beyond visual fidelity, including structural consistency, target localization, and motion accuracy. Comprehensive experiments demonstrate that CrossScope consistently outperforms strong video generation baselines, highlighting the importance of role-aware evidence routing for multi-observer visual world modeling.

Our contributions are fourfold:

1. **Role-asymmetric dual-scope future prediction.** We formulate joint future prediction for Mother–Child ERCP, where current observations and causal history provide directional, non-interchangeable evidence across views.
2. **A role-preserving read/write contract.** Separate experts expose pre-injection states through a read-only interface, while cross-scope communication is restricted to zero-initialized, target-specific residual writes.
3. **Target-specific geometric conditioning.** M2C derives Child motion conditions from a predicted Mother-plane trajectory; C2M separately translates pose-aligned Current and History, then blends their residuals only where geometry licenses both sources.
4. **A paired dual-scope benchmark and protocol.** We provide phantom and real-world episodes with role-specific evaluation of frame fidelity, Mother structure, Child localization, and Child trajectories.

Dual-Scope Benchmark

We introduce a **paired dual-scope benchmark** for evaluating multi-observer visual world models in Mother–Child ERCP (Figure 2). Unlike conventional single-view surgical video datasets, each episode contains synchronized wide-view Mother and near-field Child streams with independently controlled scope motions.

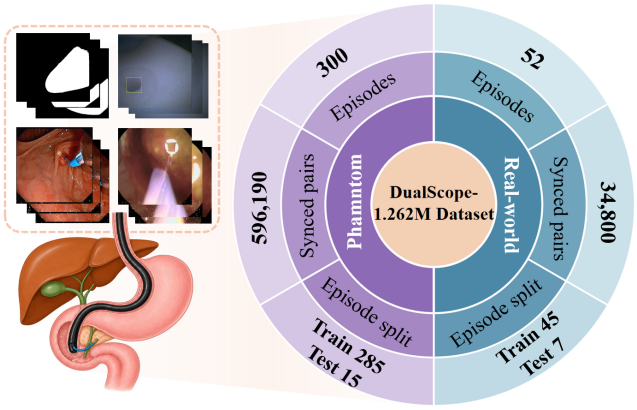

Figure 2: Dual-scope benchmark composition. The benchmark comprises paired phantom and real-world Mother–Child ERCP episodes with synchronized wide-view Mother and near-field Child streams; 1.262M denotes both views counted as individual frames.

This setting enables systematic evaluation of future prediction under complementary yet asymmetric observations, where non-shared motion, view-dependent visibility, and role-specific prediction errors can be measured explicitly. The benchmark consists of both phantom and real-world ERCP data. The phantom subset contains 300 paired episodes with 596,190 synchronized time points, including 285 training episodes and 15 test episodes. The real-world subset contains 52 paired episodes with 34,800 synchronized time points, split into 45 training and seven held-out test episodes. All splits are episode-disjoint, and the real-world evaluation further ensures patient- and case-level separation between training and testing.

Both domains provide complementary view-specific annotations, including Mother background/instrument segmentation masks and Child papilla bounding boxes. These annotations enable evaluation across multiple dimensions, including structural preservation in the Mother view, target localization in the Child view, and future motion consistency across both streams. The paired synchronization and annotation protocol were verified by clinicians and are described in Appendix C. The retrospective real-world study was conducted under ethics approval with a waiver of informed consent, and all videos were de-identified. The phantom videos, annotations, data splits, and evaluation code will be released, while real-world data will remain available only through controlled access following applicable ethics and data-use requirements.

Method

CrossScope treats dual-scope interaction as target-dependent evidence allocation rather than symmetric fusion (Figure 3). Separate Mother and Child DiT experts preserve view-

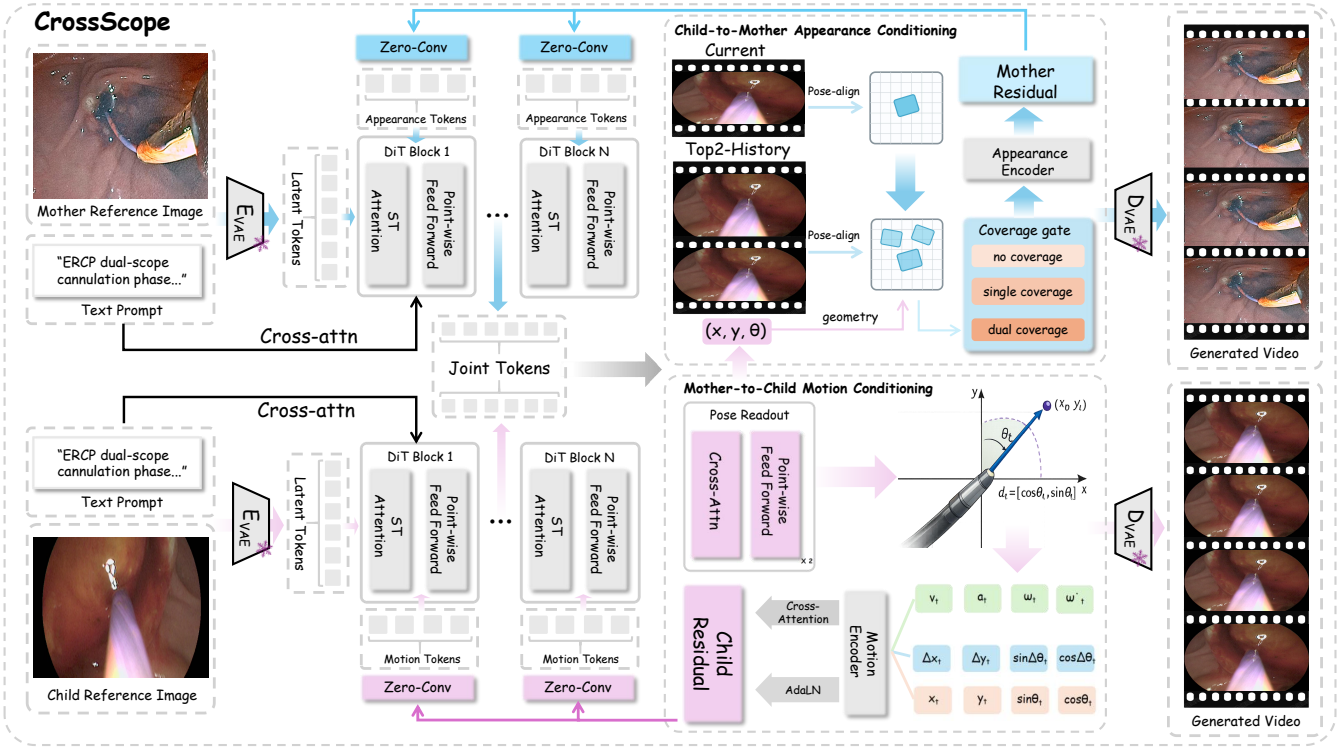

Figure 3: CrossScope architecture. Separate DiT experts expose read-only pre-injection states. Frame 0 is the observed anchor, and future latent groups cover frames 1–4 and 5–8. Pose Readout maps Mother states to a Mother-plane trajectory, which M2C converts into Child motion residuals. C2M pose-aligns Current and up to two causal History observations, then writes or blends Mother residuals only under geometric support.

specific statistics, and their pre-injection states form a read-only interface. Pose Readout estimates a Mother-plane Child-tip trajectory from the Mother partition and anchor; M2C and C2M convert relevant evidence into target-specific residuals without altering this snapshot. The Mother coordinate frame locates Child motion but does not make Child appearance globally transferable. Thus, M2C conditions Child motion from the trajectory, whereas C2M writes Mother appearance only where geometry licenses observed Child evidence. We first define the causal read/write contract, then specify M2C and C2M, before describing joint learning and the causal inference schedule.

Causal Formulation and Role-Asymmetric Dual-Stream Backbone

Although synchronized, the independently actuated streams do not form a calibrated stereo pair and cannot be represented by a shared latent with interchangeable evidence. We therefore preserve view-specific experts and permit only explicit, target-directed residual communication. Given synchronized wide-view Mother and near-field Child videos, we predict eight frames after their observed anchors. Let $I_d^{0:8}$ be a nine-frame clip for $d \in \{M, C\}$, c a fixed prompt, and $\mathcal{H}_C = \{(I_{C,j}, \hat{p}_j, a_j)\}$ a same-episode causal cache of Child observations whose times precede the anchor by at least 33 synchronized time points, with cached predicted Mother-plane poses and ages. We model

$$p_\theta\left(I_M^{1:8}, I_C^{1:8} \mid I_M^0, I_C^0, \mathcal{H}_C, c\right). \tag{1}$$

The VAE encodes each stream into $Z_d = (z_d^0, z_d^1, z_d^2)$: z_d^0 represents frame 0, z_d^1 frames 1–4, and z_d^2 frames 5–8. Separate DiT experts process the two streams while conditioning on c. At routing depth r, their pre-injection states form the read-only Joint Tokens $J^{(r)} = [X_M^{(r)}; X_C^{(r)}]$. For a within-stream update $\bar{X}_d^\ell = F_d^\ell(X_d^\ell; c)$, cross-scope communication is restricted to

$$\begin{aligned} X_M^{\ell+1} &= \bar{X}_M^\ell + R_{C\to M}^\ell, \\ X_C^{\ell+1} &= \bar{X}_C^\ell + R_{M\to C}^\ell. \end{aligned} \tag{2}$$

Zero-initialized output projections recover the uncoupled experts at initialization. All routing reads come from the pre-injection snapshot, so neither residual depends on the other path's same-layer write.

Mother-to-Child Motion Conditioning

The near-field Child stream lacks a stable reference for global displacement. The lower M2C path therefore resolves Child-motion ambiguity through a Mother-plane Child-tip trajectory. Pose Readout forms queries from the Mother anchor z_M^0 and flow-matching timestep embedding e_σ, then attends

to the Mother partition of the pre-injection states:

$$\begin{aligned}\hat{P} &= \mathcal{R}\Big(X_M^{(r)}, z_M^0, e_\sigma\Big) = (\hat{p}_0, \dots, \hat{p}_8),\\ \hat{p}_t &= \Big(\hat{x}_t, \hat{y}_t, \widehat{\cos\theta}_t, \widehat{\sin\theta}_t, \hat{\nu}_t\Big).\end{aligned} \tag{3}$$

The resulting $\hat{P}$ is expressed in the Mother plane, while M2C writes only to the Child expert. Positions lie in $[-0.5, 0.5]^2$, $\hat{\nu}_t$ is the predicted validity score used to mask invalid states, and stop-gradient prevents generation losses from redefining this interface. Appendix E provides quantitative diagnostics and visual overlays.

The Mother partition supplies wide-view localization and the pinned anchor. This shared coordinate convention gives both routes a compact geometric bottleneck without transferring appearance.

Because a single endpoint displacement cannot represent both local turning and accumulated advance, M2C encodes $\mathrm{sg}(\hat{P})$ at fine and coarse temporal scales, where sg denotes stop-gradient and hats are omitted below:

$$\begin{aligned}\phi_t = [&x_t, y_t, \cos\theta_t, \sin\theta_t,\\ &\Delta x_t^{\mathrm{ego}}, \Delta y_t^{\mathrm{ego}}, \sin\Delta\theta_t, \cos\Delta\theta_t,\\ &v_t, \omega_t, a_t, \alpha_t] \in \mathbb{R}^{12},\\ m_g^{\mathrm{coarse}} = [&\Delta x_g^{\mathrm{ego}}, \Delta y_g^{\mathrm{ego}},\\ &\sin\Delta\theta_g, \cos\Delta\theta_g], \quad g \in \{1, 2\}.\end{aligned} \tag{4}$$

Here $v_t, \omega_t, a_t, \alpha_t$ denote linear speed, angular velocity, linear acceleration, and angular acceleration. The eight descriptors, indexed by $t \in \{0, \dots, 7\}$, encode transitions $(0 \to 1), \dots, (7 \to 8)$: z_C^1 cross-attends to $\phi_{0:3}$ and z_C^2 to $\phi_{4:7}$. AdaLN uses the corresponding coarse endpoint conditions $(0, 4)$ and $(4, 8)$. The fine cross-attention and coarse AdaLN branches produce R_{attn}^ℓ and R_{AdaLN}^ℓ, whose sum forms the Child residual

$$R_{M\to C}^\ell = \alpha_{\mathrm{w}} \mathcal{Z}_C^\ell\big(R_{\mathrm{attn}}^\ell + R_{\mathrm{AdaLN}}^\ell\big). \tag{5}$$

The two conditions resolve complementary temporal scales: cross-attended ϕ_t preserves adjacent displacement and heading changes, while coarse AdaLN summarizes net motion over each four-transition group. Their combination retains transition-level variation while constraining accumulated displacement over the predicted horizon.

Zero-initialized $\mathcal{Z}_C^\ell$ is enabled through the warmup α_{w}. To supervise the motion written to the Child expert rather than only the intermediate pose, an auxiliary head decodes a Child box trajectory from layer-averaged residual fields and applies trajectory, endpoint, ROI, and validity losses.

Child-to-Mother Appearance Conditioning

The Child view may reveal local anatomy that is not visible in the wider Mother view, but the observed Child appearance is neither globally registered to the Mother view nor valid as future evidence. C2M therefore establishes Mother-plane support from observed sources before writing any appearance residual. It projects the inference-available anchor z_C^0 once with $\hat{p}_0$ and broadcasts the static canvas to future Mother groups. Historical slices use their cached poses, preventing pose–appearance mismatch. The future trajectory guides ROI selection but contributes no appearance; only observed Current and History latents do. For local Child coordinate u and Mother-plane pose $p = (x, y, \theta)$,

$$\pi(p, u) = \begin{bmatrix} x \\ y \end{bmatrix} + \mathrm{Rot}(\theta)\big(Au + \delta(u)\big), \tag{6}$$

where the global footprint A and bounded correction $\delta(u)$ are learned through Mother flow matching. Only valid in-frame projections write the static canvas, defining spatial support without calibrated stereo geometry or pixel copying.

The current Child anchor alone may not cover all Mother-view regions relevant to the predicted trajectory. To complement rather than duplicate Current evidence, Geometry-first Top2-History selects up to two same-episode cached observations at least 33 synchronized points before the anchor, without appearance similarity or a learned router. Starting from Current coverage, its score combines valid in-frame footprint, age decay, and incremental trajectory-ROI coverage. It accepts the highest positive-gain candidate, updates the coverage union, and repeats once, so the second source must add support; unsupported candidates have zero gain. Selected latent slices form one History canvas with slotwise maximum coverage. Thus, History contains zero, one, or two causal observations. Appendix A.3 gives the exact score and placement contract.

Geometric support specifies where evidence is admissible, but not how Child-domain features should be expressed in the Mother expert. Pose alignment gives source $q \in \{\mathrm{cur}, \mathrm{hist}\}$ a value canvas V_s^q and coverage C_s^q at Mother slot s, with mask $m_s^q = \mathbf{1}[C_s^q \geq \tau]$. For future Mother group $k \in \{1, 2\}$, the translator produces $T_{k,s}^{q,\ell} = \mathcal{E}_{\mathrm{app}}^\ell(V_s^q)$ and the zero-initialized projection produces $Q_{k,s}^{q,\ell} = \mathcal{Z}_M^\ell(T_{k,s}^{q,\ell})$. Let $b_s = (m_s^{\mathrm{cur}}, m_s^{\mathrm{hist}})$. At a fixed (k, s, ℓ), write $R = R_{C\to M,k,s}^\ell$, $Q_{\mathrm{cur}} = Q_{k,s}^{\mathrm{cur},\ell}$, and $Q_{\mathrm{hist}} = Q_{k,s}^{\mathrm{hist},\ell}$, and denote the sigmoid gate predicted from $\bar{X}_{M,k,s}^\ell$ by $\gamma \in [0, 1]$. The coverage gate writes

$$R = \begin{cases} 0, & b_s = (0, 0),\\ Q_{\mathrm{cur}}, & b_s = (1, 0),\\ Q_{\mathrm{hist}}, & b_s = (0, 1),\\ \gamma Q_{\mathrm{cur}} + (1 - \gamma) Q_{\mathrm{hist}}, & b_s = (1, 1). \end{cases} \tag{7}$$

Geometry therefore licenses each write: unsupported slots remain strict Null, single-source slots use the available translated residual, and dual-source slots blend residuals only after both sources are licensed. The trust gate cannot create support, leaving the Mother expert to complete remaining unsupported regions.

Learning and Causal Inference

Training jointly supervises future generation and the geometric conditions that govern residual writes. Both streams apply flow matching (Lipman et al. 2023) to non-anchor latent groups, while training combines generation, pose, and M2C residual-trajectory supervision:

$$\mathcal{L} = \lambda_M \mathcal{L}_{\mathrm{FM}}^M + \lambda_C \mathcal{L}_{\mathrm{FM}}^C + \lambda_P \mathcal{L}_{\mathrm{pose}} + \alpha_{\mathrm{w}} \lambda_{\mathrm{aux}} \mathcal{L}_{\mathrm{aux}}. \tag{8}$$

Model	Child View				Mother View			
	PSNR↑	SSIM↑	FID↓	LPIPS↓	PSNR↑	SSIM↑	FID↓	LPIPS↓
(a) Frame Fidelity								
HunyuanVideo-I2V	34.332	0.942	18.213	0.0848	35.112	0.953	34.324	0.0352
Cosmos-H-Surgical	<u>35.123</u>	<u>0.943</u>	<u>17.431</u>	<u>0.0821</u>	35.774	<u>0.969</u>	33.617	<u>0.0347</u>
Wan2.2	35.003	0.940	17.986	0.0854	35.001	0.956	34.618	0.0355
Early-Fusion Wan2.2	30.790	0.915	20.190	0.0877	35.432	0.966	33.378	0.0354
Symmetric Cross-Attention MoT	33.326	0.931	18.369	0.0852	<u>36.298</u>	0.966	<u>29.019</u>	0.0351
CrossScope	**35.906**	**0.951**	**17.402**	**0.0819**	**37.147**	**0.973**	**28.232**	**0.0334**

Model	Mother Structure	Child Localization		Child Trajectory	
	Mask IoU↑	Box IoU↑	Det. Recall↑	Endpoint Err.↓	Centerline Err.↓
(b) Task Fidelity and Motion					
HunyuanVideo-I2V	<u>0.946</u>	0.789	0.909	36.068	19.126
Cosmos-H-Surgical	0.937	<u>0.802</u>	0.913	<u>28.880</u>	15.339
Wan2.2	0.944	0.792	0.911	35.093	18.544
Early-Fusion Wan2.2	0.933	0.724	0.907	42.590	21.733
Symmetric Cross-Attention MoT	0.941	0.798	<u>0.917</u>	31.767	<u>12.414</u>
CrossScope	**0.948**	**0.839**	**0.927**	**19.343**	**11.238**

Table 1: End-to-end comparison on the phantom benchmark. All systems use the common training split and evaluation windows. Panel (a) reports view-specific frame fidelity and panel (b) task fidelity and motion; trajectory errors are in pixels. Bold and underlined values mark the best and second-best results.

The pose loss covers position, orientation, validity, and coarse endpoint motion. Stop-gradient protects the geometric interface while flow matching trains how each route uses it; Mother flow matching directly supervises C2M without an appearance proxy. Loss decompositions and weights are given in Appendix B.

At inference, the schedule enforces the same causal evidence boundary. At each denoising step, routing modules read the pre-injection snapshot, and Pose Readout uses the Mother partition. Current uses its predicted anchor pose and each earlier History observation its cached pose, so inference needs no ground-truth pose. C2M admits only z_C^0 and earlier observations as appearance evidence; future tokens remain internal denoising variables. Routes are recomputed as latents evolve, and both anchors are restored after each scheduler update; Appendices B and E provide the corresponding routing configuration and pose diagnostics.

Experiments

Experimental Setup

Training details. All systems use rank-128 LoRA (Hu et al. 2022) for 10,000 steps on the same split. CrossScope assigns full- and half-width 30-layer Wan2.2-TI2V-5B DiTs to Child and Mother, respectively, initializing the Mother by interpolation; the experts use separate adapters and a shared frozen VAE. Each window subsamples 33 synchronized points at stride four into one anchor and eight targets. Phantom tests use 860 causal windows from 15 episodes; real-world evaluation uses a patient- and case-disjoint 45/7 split within one cohort. We report PSNR, SSIM, FID, and LPIPS (Heusel et al. 2017; Zhang et al. 2018) alongside frozen task evaluators applied identically to generated and target clips. Appendices D and B detail metric aggregation and implementation.

Baselines. We compare HunyuanVideo-I2V, Wan2.2, and Cosmos-H-Surgical (Kong et al. 2024; Wan Team 2025; He et al. 2025) as independently fine-tuned, view-specific controls, alongside Early-Fusion Wan2.2 and Symmetric Cross-Attention MoT (Symmetric MoT). The first three predict each view independently, whereas Early-Fusion tests direct input coupling. Symmetric MoT retains separate experts but uses mirror-symmetric, zero-initialized cross-attention residuals without Pose Readout, M2C descriptors, or C2M placement. The comparison thus separates independent prediction, generic coupling, and directional routing. All complete systems use the same paired windows, observed anchors, prompt, optimization budget, and evaluation protocol. Table 1 reports end-to-end results, and Table 2 gives the same-backbone route analysis. Appendices B and D provide the corresponding configuration and evaluator details.

Main Results

View-specific generation fidelity. The main comparison asks whether independent prediction, input fusion, symmetric exchange, or directional routing best preserves two visually distinct targets. Table 1(a) combines distortion measures (PSNR and SSIM) with perceptual measures (FID and LPIPS), requiring agreement with target frames and their appearance statistics. The independent baselines maintain competitive Child appearance, whereas Early-Fusion and Symmetric MoT concentrate their gains more strongly in the Mother view and do not consistently preserve Child fidelity. Thus, increasing cross-view exchange alone does not determine which evidence is useful to each target. CrossS-

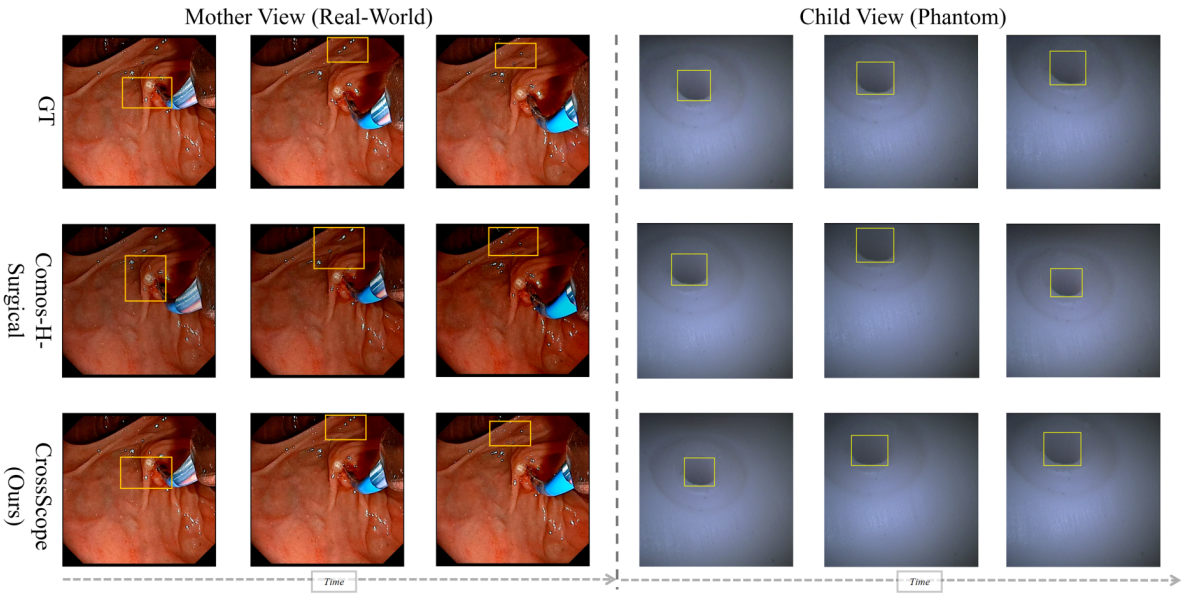

Figure 4: Qualitative future prediction across views. Independent examples are shown for a real-world Mother sequence (left) and a phantom Child sequence (right). Rows show ground truth, Cosmos-H-Surgical, and CrossScope; yellow boxes mark comparison regions.

cope yields the strongest reported overall profile among the evaluated systems: relative to Cosmos-H-Surgical, it reduces Mother FID by 16.0% and improves Child PSNR and SSIM, while retaining low Child perceptual errors. This role-balanced pattern is consistent with separate view experts and target-specific residual exchange.

Task fidelity and motion. Frame fidelity alone does not establish whether dual-scope geometry is preserved. Table 1(b) therefore measures Mother foreground structure (mask IoU), Child localization and visibility (box IoU and recall), and path and terminal motion (centerline and endpoint errors). Symmetric MoT achieves the lowest centerline error among the controls, but its endpoint error remains above Cosmos-H-Surgical, showing that mean path agreement need not recover the final motion state. CrossScope leads all reported task metrics, reducing endpoint error by 33.0% relative to Cosmos-H-Surgical while improving centerline alignment and localization coverage. The gains thus extend from appearance to spatial structure and motion without an observed frame-fidelity trade-off.

Ablation Study

Directional specialization. Table 2 asks whether the gains follow the intended direction of each route within the same MoT backbone and optimization schedule. Adding M2C concentrates its largest improvements on Child localization and trajectory quality, reducing endpoint error by 18.5% over MoT-only while leaving Mother fidelity nearly unchanged. C2M shows the complementary profile: it improves every Mother frame-fidelity measure and reduces Mother FID by 16.8%, while its Child-motion changes remain smaller. This selective response matches the architecture: pose-derived transition descriptors write motion residuals only to the Child expert, whereas pose-aligned appearance evidence writes only to the Mother expert. The alignment supports route-

Model	Child View			
	PSNR↑	SSIM↑	FID↓	LPIPS↓
MoT-only	34.868	0.939	17.959	0.0855
CrossScope-M2C	<u>35.113</u>	<u>0.941</u>	<u>17.534</u>	0.0838
CrossScope-C2M	34.927	<u>0.941</u>	17.633	<u>0.0831</u>
CrossScope	**35.906**	**0.951**	**17.402**	**0.0819**

Model	Mother View			
	PSNR↑	SSIM↑	FID↓	LPIPS↓
MoT-only	35.871	0.965	37.575	0.0356
CrossScope-M2C	35.934	0.967	37.200	0.0356
CrossScope-C2M	<u>36.542</u>	<u>0.969</u>	<u>31.277</u>	<u>0.0343</u>
CrossScope	**37.147**	**0.973**	**28.232**	**0.0334**

Model	Mother Structure	Child Localization		Child Trajectory	
	Mask IoU↑	Box IoU↑	Det. Recall↑	End. Err.↓	Cent. Err.↓
MoT-only	0.943	0.793	0.919	26.593	13.727
CrossScope-M2C	0.943	<u>0.816</u>	<u>0.923</u>	<u>21.665</u>	<u>12.437</u>
CrossScope-C2M	<u>0.947</u>	0.795	0.919	25.537	13.426
CrossScope	**0.948**	**0.839**	**0.927**	**19.343**	**11.238**

Table 2: Route-level directional ablation on the phantom benchmark. The three blocks report Child frame fidelity, Mother frame fidelity, and task fidelity/motion; trajectory errors are in pixels. Bold and underlined values mark the best and second-best results.

specific functional specialization rather than a uniform effect of generic cross-view communication across the two experts.

C2M spatial-support behavior. Figure 5 isolates how C2M forms admissible Mother-plane writes. Current coverage remains localized around the anchor-aligned region, whereas selected causal History contributes complementary support. Their composition yields Current-only, History-only, and dual-source states while retaining Null elsewhere, matching the geometry-licensed residual contract in Equation 7.

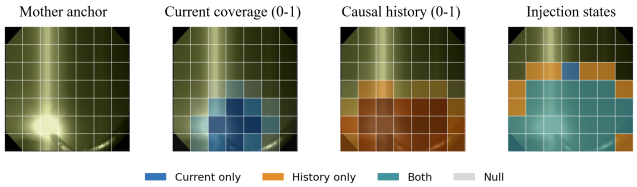

Figure 5: C2M spatial-support diagnostic. A held-out phantom window from the 10,000-step C2M-only diagnostic. Columns show the Mother anchor, Current and causal-History coverage, and the resulting 7×7 injection states; additional cases appear in Appendix Figure 7.

Route complementarity. If the routes were redundant, the full model would match the stronger single-route variant for each target. Instead, adding C2M to CrossScope-M2C lowers Mother FID from 37.200 to 28.232, while adding M2C to CrossScope-C2M lowers Child endpoint error from 25.537 to 19.343 pixels. C2M supplies Mother appearance evidence absent from motion-only routing, whereas M2C corrects Child displacement left by appearance-only routing. Their combination improves both role-specific metric families rather than reallocating fidelity between streams.

Real-World Evaluation

Held-out quantitative performance. We test whether phantom trends persist under real-world appearance, illumination, and scope-motion variation. We fine-tune Cosmos-H-Surgical and CrossScope on 45 real episodes and evaluate seven patient- and case-disjoint episodes from the same cohort. Across the role-specific measures in Figure 6, Mother PSNR increases by $1.57\,\mathrm{dB}$, while endpoint and centerline errors decrease by 16.5% and 21.5%. Consistent gains across both roles match the phantom pattern. Appendix D reports exact values and metric definitions.

Qualitative comparison. In Figure 4, CrossScope more closely preserves broad-context local structure in the Mother example and follows near-field lumen position and appearance in the Child example than Cosmos-H-Surgical. These independent sequences visually contextualize the structure, localization, and motion trends reported by the aggregate quantitative evaluation.

Cross-setting evidence. Together, complete-system comparisons distinguish directional routing from independent prediction, input fusion, and symmetric exchange; same-backbone ablations link each route to its target; and real-world evaluation tests the joint pattern under natural acquisition variation. Concurrent gains in Mother fidelity and structure and in Child fidelity, localization, and motion indicate a role-consistent forecast rather than an isolated metric change. View-specific measures test whether routing preserves both predictive roles rather than reallocating fidelity between streams.

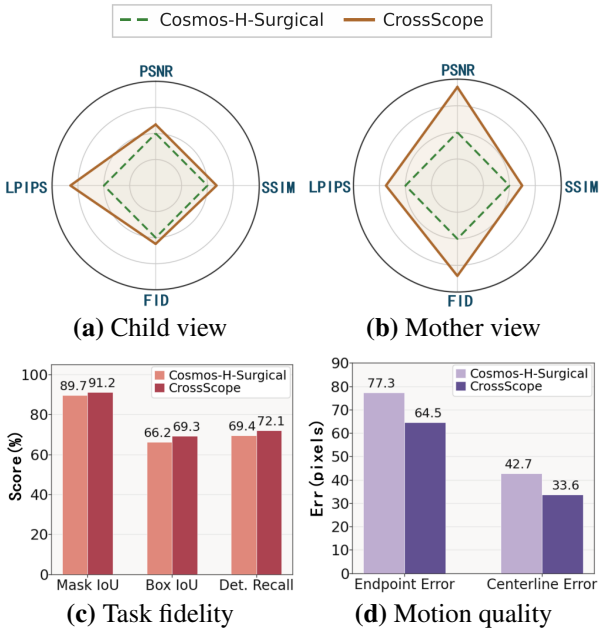

Figure 6: Real-world quantitative comparison on seven held-out episodes. Radar scores are normalized to Cosmos-H-Surgical ($= 100$), using reciprocal ratios for lower-is-better metrics. Bars report task-fidelity scores (%) and trajectory errors (pixels). Exact values are provided in Appendix D.

Conclusion

CrossScope structures dual-scope forecasting around what to share, which target receives it, and where to write it. In Mother–Child ERCP, role-preserving experts use a Mother-plane trajectory for Child motion and admit pose-aligned Child appearance to Mother prediction only under geometric support. Across phantom and held-out real-world episodes in the paired benchmark, CrossScope improves the reported measures of frame fidelity, structure, localization, and motion; same-backbone ablations associate each route with its intended role. These results support target-specific evidence routing as a design principle for paired flexible scopes.

The present study is limited to short-horizon prediction in one dual-scope procedure, with real-world evaluation drawn from one cohort. Future work will extend CrossScope to longer horizons, additional dual-scope procedures, and prospective validation across centers and devices before considering clinical decision-support applications.

References

Biagini, D.; Navab, N.; and Farshad, A. 2025. HieraSurg: Hierarchy-Aware Diffusion Model for Surgical Video Generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15968 of *Lecture Notes in Computer Science*, 310–319. Springer Nature Switzerland.

Bruce, J.; Dennis, M. D.; Edwards, A.; Parker-Holder, J.; Shi, Y.; Hughes, E.; Lai, M.; Mavalankar, A.; Steigerwald, R.; Apps, C.; Aytar, Y.; Bechtle, S. M. E.; Behbahani, F.; Chan, S. C. Y.; Heess, N.; Gonzalez, L.; Osindero, S.; Ozair, S.; Reed, S.; Zhang, J.; Zolna, K.; Clune, J.; de Freitas, N.; Singh, S.; and Rocktäschel, T. 2024. Genie: Generative Interactive Environments. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 4603–4623. PMLR.

Chen, T.; Yang, S.; Wang, J.; Bai, L.; Ren, H.; and Zhou, L. 2025. SurgSora: Object-Aware Diffusion Model for Controllable Surgical Video Generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15969 of *Lecture Notes in Computer Science*, 521–531. Springer Nature Switzerland.

Cheng, H. K.; and Schwing, A. G. 2022. XMem: Long-Term Video Object Segmentation with an Atkinson-Shiffrin Memory Model. In *Computer Vision – ECCV 2022*, volume 13688 of *Lecture Notes in Computer Science*, 640–658. Springer Nature Switzerland.

Ha, D.; and Schmidhuber, J. 2018. World Models. arXiv:1803.10122.

He, Y.; Guo, P.; Xu, M.; Li, Z.; Myronenko, A.; Imans, D.; Liu, B.; Yang, D.; Gu, M.; Ji, Y.; Jin, Y.; Zhao, R.; Shen, B.; and Xu, D. 2025. Cosmos-H-Surgical: Learning Surgical Robot Policies from Videos via World Modeling. arXiv:2512.23162.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, Y.; Cui, B.; Bai, L.; Guo, Z.; Xu, M.; Islam, M.; and Ren, H. 2024. Endo-4DGS: Endoscopic Monocular Scene Reconstruction with 4D Gaussian Splatting. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15006 of *Lecture Notes in Computer Science*, 197–207. Springer Nature Switzerland.

Iliash, I.; Allmendinger, S.; Meissen, F.; Kühl, N.; and Rückert, D. 2024. Interactive Generation of Laparoscopic Videos with Diffusion Models. In *Deep Generative Models*, Lecture Notes in Computer Science, 109–118. Springer Nature Switzerland.

Koju, S.; Bastola, S.; Shrestha, P.; Amgain, S.; Shrestha, Y. R.; Poudel, R. P. K.; and Bhattarai, B. 2025. Surgical Vision World Model. In *Data Engineering in Medical Imaging*, volume 16191 of *Lecture Notes in Computer Science*, 1–10. Springer Nature Switzerland.

Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; et al. 2024. HunyuanVideo: A Systematic Framework for Large Video Generative Models. arXiv:2412.03603.

Li, B.; Yang, S.; Peng, B.; Guo, X.; Zhang, E.; Tao, Y.; Duan, J.; Xu, D.; Dou, Q.; Jin, X.; Zeng, W.; Zhao, H.; and Jin, Y. 2026. From Articulated Kinematics to Routed Visual Control for Action-Conditioned Surgical Video Generation. arXiv:2605.08712.

Li, C.; Feng, B. Y.; Liu, Y.; Liu, H.; Wang, C.; Yu, W.; and Yuan, Y. 2024a. EndoSparse: Real-Time Sparse View Synthesis of Endoscopic Scenes Using Gaussian Splatting. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15006 of *Lecture Notes in Computer Science*, 252–262. Springer Nature Switzerland.

Li, C.; Liu, H.; Liu, Y.; Feng, B. Y.; Li, W.; Liu, X.; Chen, Z.; Shao, J.; and Yuan, Y. 2024b. Endora: Video Generation Models as Endoscopy Simulators. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15006 of *Lecture Notes in Computer Science*, 230–240. Springer Nature Switzerland.

Lipman, Y.; Chen, R. T. Q.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2023. Flow Matching for Generative Modeling. arXiv:2210.02747.

Lu, S.; Xiao, Z.; Wei, J.; et al. 2026. Scaling Video Pretraining for Surgical Foundation Models. arXiv:2603.29966.

Özsoy, E.; Örnek, E. P.; Eck, U.; Czempiel, T.; Tombari, F.; and Navab, N. 2022. 4D-OR: Semantic Scene Graphs for OR Domain Modeling. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, volume 13437 of *Lecture Notes in Computer Science*, 475–485. Springer Nature Switzerland.

Pan, W.; Li, W.; Liu, S.; et al. 2026. SurgVista: Long-Horizon Surgical World Modeling with Plausible Instrument-Tissue Dynamics. arXiv:2606.19889.

Parsi, M. A. 2011. Peroral Cholangioscopy in the New Millennium. *World Journal of Gastroenterology*, 17(1): 1–6.

Rapuri, S.; Seenivasan, L.; Schneider, D.; Soberanis-Mukul, R.; He, Y.; Ding, H.; Xu, J.; Yu, C.; Jing, C.; Guo, P.; Xu, D.; and Unberath, M. 2026. SAW: Toward a Surgical Action World Model via Controllable and Scalable Video Generation. Manuscript under review, arXiv:2603.13024.

Sivakumar, S. K.; Frisch, Y.; Ghazaei, G.; and Mukhopadhyay, A. 2025. SG2VID: Scene Graphs Enable Fine-Grained Control for Video Synthesis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15968 of *Lecture Notes in Computer Science*, 511–521. Springer Nature Switzerland.

Stilz, F.; Karaoglu, M.; Tristram, F.; Navab, N.; Busam, B.; and Ladikos, A. 2026. FLex: Joint Pose and Dynamic Radiance Fields Optimization for Stereo Endoscopic Videos. *International Journal of Computer Assisted Radiology and Surgery*, 21(1): 137–146.

Wan Team. 2025. Wan2.2. https://github.com/Wan-Video/Wan2.2. Model release.

Wu, J.; Holm, F.; Chen, C.; et al. 2026. SurgMotion: A Video-Native Foundation Model for Universal Understanding of Surgical Videos. arXiv:2602.05638.

Xiao, Z.; Lan, Y.; Zhou, Y.; Ouyang, W.; Yang, S.; Zeng, Y.; and Pan, X. 2025. WorldMem: Long-Term Consistent World Simulation with Memory. In *Advances in Neural Information Processing Systems*, volume 38, 49632–49652. Curran Associates, Inc.

Yang, S.; Zhou, F.; Mayer, L.; et al. 2026. Large-scale Self-supervised Video Foundation Model for Intelligent Surgery. *npj Digital Medicine*.

Yeganeh, Y.; Lazuardi, R.; Shamseddin, A.; et al. 2024. VISAGE: Video Synthesis Using Action Graphs for Surgery. arXiv:2410.17751.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

A Additional Method Details

This supplement documents the methodological and empirical contracts underlying CrossScope. Section A specifies the information each route may read and the residuals it may write. Section B records the training and inference configuration. Section C details the paired benchmark and its view-specific annotations. Section D defines the evaluation protocol and reports exact real-world values. Section E examines the intermediate pose interface and output-level Child motion. Together, these details connect the directional design in Figure 3 to the measurements reported in the main paper.

A.1 Read–Write Routing Invariant

At every routed transformer layer, CrossScope separates evidence reads from residual writes. Let H_M and H_C denote the pre-injection Mother and Child states. The manuscript term *Joint Tokens* denotes the conceptual paired snapshot $[H_M; H_C]$ available for reading; it is not a third mutable stream. Pose Readout, M2C, and C2M therefore inspect the same pre-write state, but add their outputs only to their designated expert. Neither route can condition on the other route's same-layer residual. This invariant gives the paths independent write semantics: M2C converts a pose-derived transition description into a Child residual, whereas C2M converts pose-aligned observed Child evidence into a Mother residual. Zero-initialized output projections make both paths exact no-ops at initialization, preserving the uncoupled view experts before the routes are learned.

A.2 Mother-to-Child Motion Conditioning

Pose Readout interface. Pose Readout is attached at layer 7. It uses nine learned temporal queries with sinusoidal temporal embeddings and a pooled flow-matching timestep embedding. The queries cross-attend to memory formed from the pre-injection Mother tokens and pinned Mother anchor-token group. Its inference inputs are therefore the Mother state, observed Mother anchor, and denoising timestep; it neither reads Child tokens nor consumes ground-truth pose. The output is a nine-state Mother-plane sequence $(x, y, \cos\theta, \sin\theta, \ell_v)$. Position is bounded to $[-0.5, 0.5]^2$, heading is normalized to the unit circle, and ℓ_v is a validity logit. Before M2C or C2M uses the prediction, the route-facing pose is detached and invalid states are masked by $\ell_v \leq 0$. This prevents generation losses from redefining the geometric interface while leaving the pose objective to supervise it directly.

Sampling-aware descriptor and residual write. M2C transforms the detached pose sequence into a descriptor for each valid transition between temporally sampled states. The per-frame pose cache stores $(x, y, \cos\theta, \sin\theta, q)$, where q is validity. Derivatives are formed after temporal sampling, rather than on the raw synchronized stream, so velocity and acceleration use the interval represented by the generated video. Table 3 lists the twelve components. The four ego-motion components form the compact coarse condition, whereas the complete descriptor supplies fine transition-wise evidence.

Group	Components and interpretation
Global pose	$(x_k, y_k, \cos\theta_k, \sin\theta_k)$: normalized tip position and unit heading at the transition start.
Ego motion	$(\Delta x_k^{\text{ego}}, \Delta y_k^{\text{ego}}, \sin\Delta\theta_k, \cos\Delta\theta_k)$: translation expressed in the current tip frame and wrap-safe heading change.
First order	(v_k, ω_k): translational speed and angular velocity over the sampled interval.
Second order	(a_k, α_k): finite differences of v_k and ω_k; both are zero for the first transition.

Table 3: Twelve-dimensional Mother-plane transition descriptor used by M2C.

For consecutive sampled poses, let $\Delta x_t = x_{t+1} - x_t$ and $\Delta y_t = y_{t+1} - y_t$. Translation is rotated into the current tip frame before it is encoded:

$$\begin{bmatrix} \Delta x_t^{\text{ego}} \\ \Delta y_t^{\text{ego}} \end{bmatrix} = \begin{bmatrix} \cos\theta_t & \sin\theta_t \\ -\sin\theta_t & \cos\theta_t \end{bmatrix} \begin{bmatrix} \Delta x_t \\ \Delta y_t \end{bmatrix}.$$

The heading change is represented by $(\sin\Delta\theta_t, \cos\Delta\theta_t)$ rather than direct angle subtraction, avoiding wrap discontinuities at $\pm\pi$. Translational speed and angular velocity are computed over the sampled interval, and their finite differences produce a_t and α_t. The first transition uses zero second-order terms. Hence, the nine predicted states yield eight fine transition descriptors. The two coarse conditions are recomputed directly from endpoint pairs $(0, 4)$ and $(4, 8)$, rather than by averaging fine transitions; they condition the first and second future latent groups, respectively.

At its enabled Child layers, M2C first writes a validity-masked cross-attention residual from the fine descriptors, then an AdaLN residual conditioned on the corresponding coarse descriptor. The anchor latent group is unchanged. A missing pose state is not replaced by a nominal motion vector; its transition contributes no M2C residual. Fine conditions retain transition-level variation, whereas the two coarse conditions provide the group-level context used by AdaLN. The route remains disabled during the first 3,000 optimization steps, allowing the view experts and Pose Readout to establish stable single-view representations before motion residuals are introduced. Its auxiliary decoder reads the residual fields written at these layers rather than the raw Pose Readout output, so the auxiliary signal supervises the generation pathway it conditions.

A.3 C2M Causal Appearance Placement

C2M has a different information contract from M2C: it writes only pose-aligned appearance from observed Child frames to the Mother expert. The Current source is the observed Child anchor; History consists solely of strictly earlier cached Child observations. Future Child-token states remain internal denoising variables and are never admitted as C2M appearance evidence. The nine predicted poses are grouped into anchor state 0, future states 1–4, and future states 5–8, matching the three VAE latent groups. Current uses the observed anchor latent and its group-0 pose. For each historical observation, its latent slices are placed with their cached poses before evidence is aggregated across slices. This ordering prevents a pose–appearance mismatch from aggregating temporal evidence before spatial alignment.

Figure 5 highlights one representative window in the main paper, while Figure 7 extends this diagnostic to four held-out phantom windows. Together, they separate continuous geometric coverage, which licenses a source at a slot, from the discrete write state, which determines whether the translated residual is written from Current, History, both, or neither.

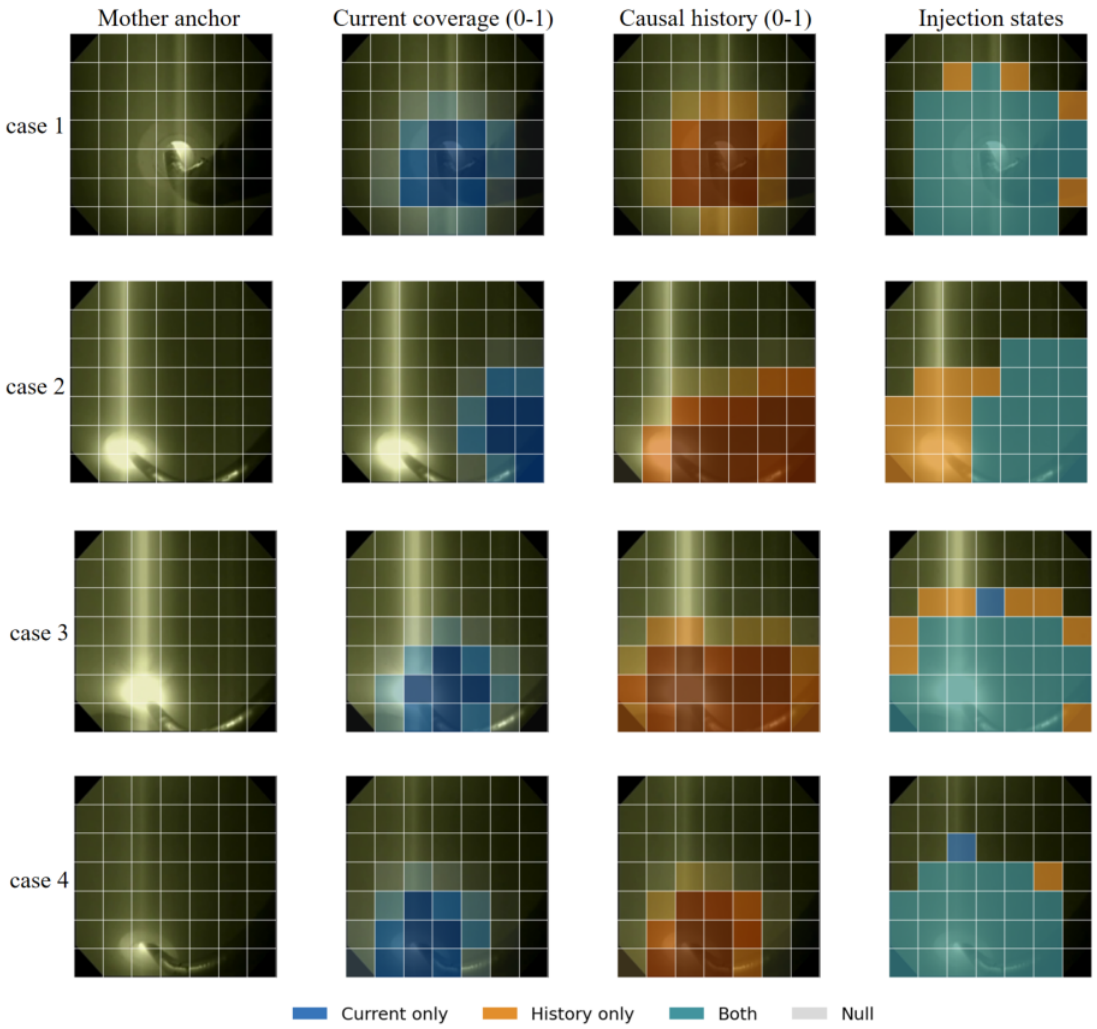

Figure 7: Geometry-licensed C2M evidence placement. Four held-out phantom windows from the 10,000-step C2M-only geometry-canvas ablation are visualized at the final denoising step. Columns show the Mother anchor, continuous Current coverage, continuous causal-History coverage, and the final discrete 7×7 injection state. Blue, orange, teal, and gray denote Current-only, History-only, dual-source, and Null support, respectively. This is a route diagnostic, not a complete-system performance result.

Each admissible source is projected to a 7×7 Mother-plane canvas. The projection produces an in-frame footprint and a coverage value for every Mother slot. C2M then selects complementary History evidence by geometry, not by an appearance-similarity score. For a historical candidate j, let C_s^j be its coverage at slot s, f_j its in-frame footprint ratio, a_j its age in synchronized frames, and C the coverage union already obtained from Current and previously selected History. Its marginal gain is

$$G_j(C) = f_j\, 2^{-a_j/h} \sum_{k=1}^{2} \sum_{s} \rho_{k,s} \left[C_s^j - C_s\right]_+, \qquad [x]_+ = \max(x, 0), \tag{9}$$

where h is the recency half-life and $\rho_{k,s}$ weights slot s by the predicted-trajectory region of interest for future group k. Selection starts from Current coverage. After the highest positive-gain candidate is selected, its support is merged into the union before the second round. The procedure thus returns zero, one, or two causal History observations, and the second observation must add support not already supplied by Current or the first History observation.

The score acts only on geometric coverage and recency; it does not inspect an appearance-similarity feature. Selection therefore determines *where* History can contribute before the learned translator determines *how* its values should be expressed in the Mother expert. If no candidate has positive marginal gain, the History canvas remains empty and C2M reduces to Current-only or Null behavior according to available support.

Geometry establishes support before any C2M residual is written. At every enabled Mother layer, Current and History values are translated *independently* into Mother-domain residuals. A Mother-state-conditioned sigmoid gate blends these translations only at slots supported by both sources. At single-source slots, the available translation is written directly; at unsupported slots, the update is exactly zero. The learned gate therefore chooses how to combine already licensed sources, but cannot create support where geometry provides none. The canvas size, layer schedule, and warm-up are recorded in Table 4.

B Implementation and Reproducibility

B.1 Training and Routing Configuration

All reported systems use the fixed prompt "ERCP dual-scope cannulation phase," the same paired training split, prediction horizon, and evaluation windows. Table 4 records the configuration for the reported 10,000-step CrossScope runs. It distinguishes the common optimization protocol from settings specific to the two directional routes, allowing the route architecture to be reproduced without inferring implementation values from the main text.

Training was performed on one Ubuntu 4XN6lS-80GB node with four NVIDIA H100 GPUs (80 GB each), 88 vCPU cores, and 960 GiB of system memory, using Python 3.10.16, PyTorch 2.7.1, and CUDA 12.8. Each reported result is obtained from one training run per model using training seed 42 and fixed generation seed 12,345.

Setting	Value
Backbone and VAE	Wan2.2-TI2V-5B; 30-layer Child/Mother DiTs with hidden/FFN dimensions $3072/14336$ and $1536/7168$; interpolated Mother initialization, independent rank-128 LoRA adapters, and a shared frozen WanVideoVAE38.
Input and adaptation	224×224 per view; bfloat16; per-device batch size 16; no gradient accumulation.
Optimizer and schedule	AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, learning rate 10^{-4}, zero weight decay, and cosine decay.
Optimization and sampling	10,000 optimization steps; training/split seed 42; fixed generation seed 12,345 and 20 denoising steps at evaluation.
Pose and M2C	Pose Readout at layer 7; M2C layers $[8, 10, 12, 14, 16, 18, 20, 22, 25, 28]$; 3,000-step M2C warm-up.
C2M	Layers $[8, 11, 14, 17, 20, 23, 26, 29]$; 7×7 canvas; 500-step Null warm-up; footprint scale 0.1; kernel width 0.15.
Causal history	Coverage threshold $\tau = 0.05$; recency half-life $h = 200$ frames; history candidates precede the Current anchor by at least 33 frames.

Table 4: Training, routing, and causal-history settings used for the reported CrossScope runs.

B.2 Baselines, Capacity, and Objectives

Every reported model is LoRA-fine-tuned for 10,000 optimization steps. HunyuanVideo-I2V, Wan2.2, and Cosmos-H-Surgical use one view-specific instance for each target stream. Early-Fusion Wan2.2 encodes paired anchors on a horizontally concatenated 224×448 canvas and predicts the paired future in that layout. Symmetric Cross-Attention MoT retains separate Mother and Child latents and experts, but replaces directional routing with mirror-symmetric, zero-initialized cross-attention residuals. It has neither Pose Readout nor the M2C descriptor and C2M geometry contracts. MoT-only disables all three CrossScope additions, whereas CrossScope-M2C and CrossScope-C2M activate only the named path. Before the common evaluator is applied, every output is separated into its two views.

CrossScope contains 688.3 million trainable parameters, including the LoRA adapters, Pose Readout, M2C, and C2M, and excluding frozen VAE and base-backbone weights. This reports the trainable capacity of the complete CrossScope implementation. The directional attribution in the main paper comes from the MoT-only and single-route variants, which retain the same dual-stream backbone and optimization schedule.

The main objective combines Mother and Child flow matching, pose supervision, and an M2C residual-trajectory auxiliary term. Its outer weights are

$$\lambda_M = 1, \qquad \lambda_C = 1, \qquad \lambda_P = 1, \qquad \lambda_{\text{aux}} = 0.1.$$

The Pose Readout objective is

$$\mathcal{L}_{\text{pose}} = \mathcal{L}_{xy} + 0.5\mathcal{L}_{\text{dir}} + 0.05\mathcal{L}_{\text{unit}} + 0.2\mathcal{L}_{\text{valid}} + \mathcal{L}_{\text{coarse}},$$

which supervises normalized position, heading, direction-vector norm, validity, and coarse endpoint motion. The M2C auxiliary objective is

$$\mathcal{L}_{\text{aux}} = \mathcal{L}_{\text{traj}} + \mathcal{L}_{\text{endpoint}} + 0.5\mathcal{L}_{\text{ROI}} + 0.2\mathcal{L}_{\text{valid}}.$$

The auxiliary decoder reads the M2C residual fields after they have been constructed, rather than the raw pose output. Mother flow matching directly supervises C2M's appearance translator and its support-conditioned residual writes.

B.3 Causal Inference Schedule

Table 5 records the information available at each inference-stage operation. Predicted pose is a geometric condition. Future Child latent states may participate in the Child denoising stream, but they are never admitted as C2M appearance evidence; C2M receives only the observed anchor and strictly earlier cached observations.

Stage	Available information	Operation
Anchor preparation	Observed Mother and Child anchors; cached strictly earlier Child observations	Pin the two observed anchors and retain only causal Child history for potential C2M evidence.
Pose Readout	Pre-injection Mother state and Mother anchor tokens	Predict the nine Mother-plane pose states and detach the route-facing pose representation.
M2C update	Detached pose transitions and Child pre-injection state	Write fine motion and coarse AdaLN residuals only to future Child token groups.
C2M update	Observed Child anchor, causal Child history, predicted poses, and Mother pre-injection state	Establish support, translate Current and History evidence, and write only geometry-licensed Mother residuals.
Scheduler step	Updated future latents and the pinned anchors	Recompute routes as latents evolve, then restore both observed anchors.

Table 5: Causal information flow during CrossScope inference.

At rollout start, the observed Mother and Child anchors are VAE-encoded independently and copied into latent group 0, while the two future groups are initialized from noise. Both experts use the same scheduler timestep. After each scheduler update, the saved anchor latents are restored before the next denoising step. Frame 0 is therefore never denoised as a prediction; it remains the observed condition while future latents and route conditions are recomputed throughout inference.

C Dual-Scope Benchmark and Annotation Protocol

C.1 Benchmark Composition and Split

Each synchronized time point is a paired Mother–Child observation: the wide-view Mother and near-field Child frames arise at the same procedural instant while the two scopes remain independently controlled. The phantom collection contains 300 paired episodes (285 training and 15 test), and the real-world collection contains 52 (45 training and seven held-out test). Accordingly, the time-point counts in Table 6 refer to synchronized pairs, not individual view frames. Splits are episode-disjoint; real-world test episodes are patient- and case-disjoint. Experienced ERCP clinicians reviewed pairing, synchronization, and task-specific annotations.

At the model boundary, the two views remain separate tensors rather than a horizontally concatenated canvas. Each raw 480×360 stream is independently resized to 224×224. A 33-point synchronized clip is then sampled at an interval of four to form the nine-frame input–target sequence. This preserves the view-specific image statistics seen by the two experts while retaining a shared temporal reference for paired prediction and evaluation.

C.2 View-Specific Annotation Pipeline

Figure 8 links paired raw observations to the task-specific quantities used throughout the paper. In the Mother view, the segmentation map separates background from the Child-endoscope foreground. A mask-derived extractor records a normalized tip center, unit heading vector, and validity label for each frame. The pose cache is created before temporal sampling, allowing the same per-frame geometric primitive to be differentiated at the interval of a training or evaluation window. These mask-derived poses supervise Pose Readout and provide the Mother-plane geometric reference for both directional routes.

Source	Episodes	Synchronized time points	Paired annotations
Phantom	285 train / 15 test	596,190	Mother segmentation map and mask-derived Child-tip pose; Child papilla box.
Real-world	45 train / 7 test	34,800	Mother segmentation map and mask-derived Child-tip pose; Child papilla box.

Table 6: Composition and view-specific annotations of the Dual-Scope Benchmark. Each synchronized time point contains one paired Mother–Child observation.

The pose extractor follows a deterministic image-plane procedure. It retains the largest valid foreground component, estimates its centerline by binning points along the principal axis, and assigns the endpoint nearest an image boundary as the instrument base. The opposite endpoint is the tip, and the local tangent near that tip is oriented from base to tip. A frame that fails the geometric validity check is marked invalid rather than assigned a synthetic pose. This makes the visual overlays in Figure 8 directly traceable to the Mother segmentation map.

In the Child view, a papilla bounding box specifies target visibility and localization. The fixed detector is applied independently to generated and target clips under the same protocol. Its box centers provide the anchored displacements used for centerline and endpoint errors, whereas box overlap and recall measure localization and coverage. Thus, the paired annotation chain supports Mother structure, Child target visibility, and Child trajectory fidelity in both phantom and real-world episodes.

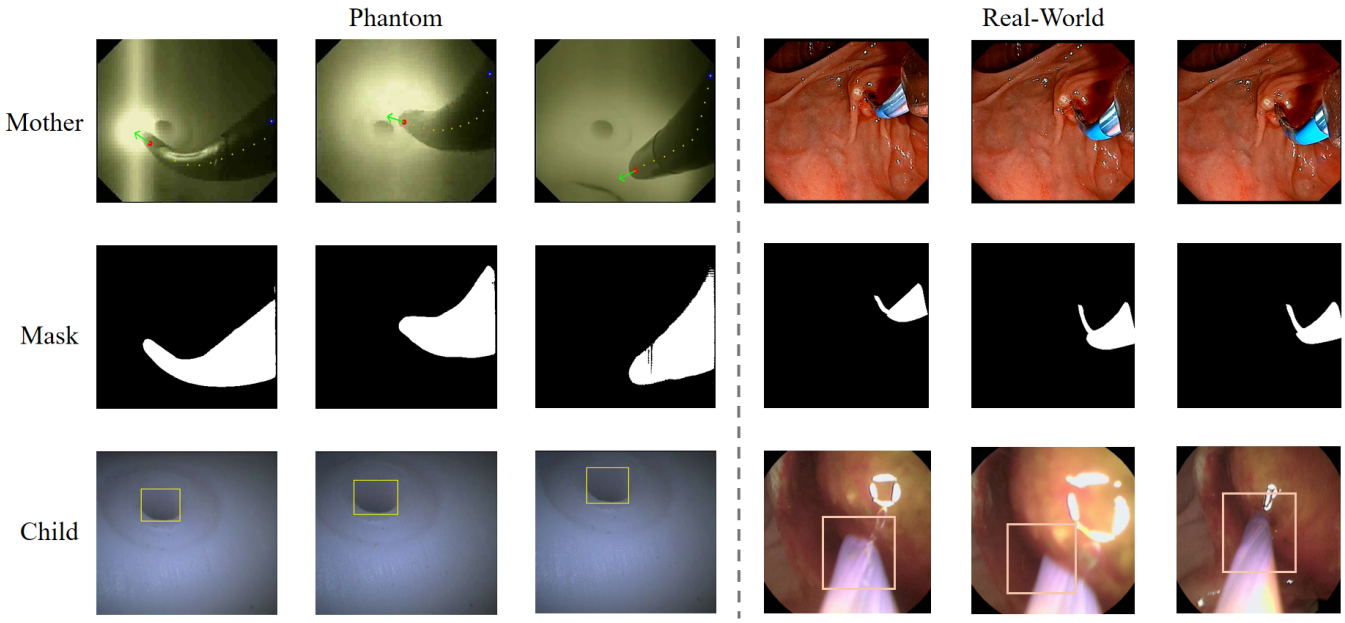

Figure 8: View-specific annotations in the Dual-Scope Benchmark. Each column is a synchronized Mother–mask–Child triplet; phantom and real-world examples are shown. In the Mother row, yellow points trace the mask-derived centerline, the blue circle marks the base, the red point marks the tip, and the green arrow marks its heading. The middle row shows the Mother segmentation map, and the Child row shows the papilla bounding box.

D Evaluation Protocol and Exact Results

D.1 Prediction Windows and Frame Metrics

Each prediction instance begins with 33 consecutive synchronized time points, temporally sampled into nine frames per view. Frame 0 is the observed anchor, and frames 1–8 are future prediction targets. The phantom test split yields 860 non-overlapping windows from 15 held-out episodes; real-world metrics are computed on seven held-out episodes. Window construction, prediction horizon, denoising budget, and generation seed are fixed across methods.

Frame-fidelity metrics assess only generated future frames. PSNR, SSIM, and AlexNet LPIPS are averaged over the eight predicted frames of each view. FID uses standard InceptionV3 pool3 features, pooling generated and target future frames separately for Mother and Child. By contrast, task-specific evaluation retains the anchor when it provides the reference required to measure a trajectory. Table 7 maps each reported metric family to its view, temporal support, and fixed evaluator.

Metric family	View	Temporal support	Evaluation source
PSNR, SSIM, FID, LPIPS	Mother, Child	Eight future frames	Generated and target frames.
Mask IoU	Mother	Anchor plus eight future frames	Fixed Child-endoscope foreground segmenter.
Box IoU and recall	Child	Anchor plus eight future frames	Fixed papilla detector applied to generated and target clips.
Centerline and endpoint error	Child	Common detected frames in the nine-frame sequence	Anchored generated and target box-center displacements.

Table 7: Relationship between reported metric families, temporal support, and fixed evaluation sources.

D.2 Structure, Localization, and Motion Aggregation

Mother structure is evaluated with a binary DeepLabV3-MobileNetV3-Large student trained on phantom Mother frames with SAM2-propagated Child-endoscope masks. Child localization is evaluated with a YOLO11x detector trained on a labeled ERCP Child corpus for papillary-orifice, lumen, and calculus detection. Its highest-confidence papillary-orifice candidate, at a confidence threshold of 0.25, supplies the reported localization and trajectory measurements. Both checkpoints are frozen. Generated and target clips receive identical preprocessing: each square model frame is restored to the native 480×360 aspect ratio before detection. Task-specific quantities use the full conditioned nine-frame sequence. The anchor is included in mask and box measurements for consistent conditioned-sequence evaluation, and establishes the displacement reference for trajectory metrics. Target-side detections define the reference objects that generated clips are expected to reproduce.

The aggregates distinguish coverage from geometric accuracy. Detection recall is defined over all target-side papilla boxes, so a missing generated detection is a false negative. Box IoU is computed only for matched generated and target boxes, and trajectory errors require at least two common detections. Thus, an absent box is not converted into an arbitrary pixel penalty, while recall remains visible alongside conditional localization and trajectory scores. Table 8 states these denominators explicitly.

Metric	Aggregation and missing-detection semantics
Mask IoU	All nine Mother frames; generated and target Child-endoscope foreground masks are scored framewise.
Detection recall	All target-side papilla boxes; a miss is a false negative.
Box IoU	Matched target-side/generated papilla boxes; a miss has no IoU value.
Trajectory errors	Papilla sequences with at least two common detections; no fixed error is inserted for a miss.

Table 8: Evaluator denominators and missing-detection semantics.

For the papilla target detected in both clips, let c_t and $c_t^\star$ be the generated and target box centers at common frames, and let t_0 be their first common frame. We compare anchored displacements $d_t = c_t - c_{t_0}$ and $d_t^\star = c_t^\star - c_{t_0}^\star$. Centerline error is the mean $\|d_t - d_t^\star\|_2$ over common frames, whereas endpoint error is the same quantity at the final common frame. The first measures path alignment across the available horizon; the second exposes accumulated terminal drift. Reporting recall alongside these conditional errors keeps coverage and matched-trajectory accuracy interpretable as complementary quantities.

D.3 Exact Real-World Results

The main paper summarizes the two-model real-world comparison with baseline-relative radar axes and bar plots. Table 9 provides the corresponding unnormalized values, so that the frame-fidelity, Mother-structure, Child-localization, and Child-motion values underlying Figure 6 can be inspected directly.

The table preserves the two view roles rather than averaging them into one score. Its upper panel reports frame fidelity on each stream's native scale, and its lower panel retains the structural, localization, coverage, and motion quantities used by the task evaluators. The main-paper radar is a compact relative visualization; Table 9 supplies absolute values and metric directions without cross-metric normalization. All entries use the same seven held-out episodes, fixed generation protocol, and evaluator definitions described above.

E Pose Readout and Generated Motion Diagnostics

E.1 Pose Readout Diagnostics

Pose Readout is evaluated at optimization step 10,000 on 860 non-overlapping held-out phantom windows. The nine predicted pose states in each window are compared with Mother-plane targets derived from segmentation maps. Position errors are converted from normalized coordinates to the resized image plane, and heading error is the angular distance between predicted and target direction vectors. Table 10 reports position, heading, coarse-motion, and full-descriptor errors.

Model	View	PSNR↑	SSIM↑	FID↓	LPIPS↓
Cosmos-H-Surgical	Child	18.314	0.737	19.153	0.445
Cosmos-H-Surgical	Mother	18.453	0.726	37.240	0.360
CrossScope	Child	**18.627**	**0.750**	**18.925**	**0.418**
CrossScope	Mother	**20.023**	**0.744**	**34.813**	**0.347**

Model	Mask↑	Box↑	Recall↑	End.↓	Cent.↓
Cosmos-H-Surgical	0.897	0.662	0.694	77.283	42.732
CrossScope	**0.912**	**0.693**	**0.721**	**64.535**	**33.557**

Table 9: Exact real-world results on seven held-out episodes. In the lower panel, Mask is Mother Child-endoscope mask IoU; Box and Recall are Child papilla-box IoU and recall; End. and Cent. are endpoint and centerline errors in pixels.

The diagnostics address two route-facing interfaces. Absolute position and heading characterize the Mother-plane geometry used for C2M placement. Coarse four-dimensional and raw twelve-dimensional errors characterize the derived transition quantities consumed by M2C after temporal sampling. Reporting both avoids treating a low absolute-pose error as evidence of accurate motion differences.

Diagnostic	Value
Normalized position L1	0.02795
Per-coordinate position MAE	6.250 px
Euclidean position error	9.750 px
Heading error	5.406°
Coarse-motion 4D L1	0.02194
Raw-motion 12D L1	0.02450

Table 10: Pose Readout diagnostics on 860 held-out phantom windows.

Figure 9 complements the aggregate diagnostics with state-level overlays of predicted and mask-derived pose primitives on held-out Mother frames.

The shuffled-target and constant-prior controls test whether the readout exploits a generic location prior rather than the Mother-stream state. Shuffling preserves the marginal target distribution but breaks the correspondence between each window and its pose target; the constant-prior predictor retains no window-specific geometry. Normalized position L1 rises from 0.02795 to 0.15137 when targets are shuffled across windows, and is 0.11328 for a constant-prior predictor. This gap is consistent with recovery of window-specific Mother-plane information. All 7,740 target pose states in this diagnostic are valid, so validity is not separately tabulated.

E.2 Generated Child Bounding-Box Trajectories

The preceding diagnostic assesses the intermediate Mother-plane pose interface. Figure 10 instead visualizes output-level papilla motion from bounding-box centers independently detected in target and generated Child clips; it does not reuse the Pose Readout target. The X and Y axes are detector-coordinate centers and T is the original frame index, exposing lateral drift or jitter that a two-dimensional overlay can conceal.

For each displayed episode, 20 nine-frame windows are sampled evenly across the full timeline. Each window is generated independently from its observed anchor. Detector centers within a window are averaged into one point placed at the raw center-frame index. A centered five-point moving average is then drawn through the sequence to suppress window-level detector jitter. The left and right panels therefore use matched temporal locations in the target and CrossScope-generated clips, but do not depict an autoregressive chained rollout.

Target and generated clips are passed through the detector independently. A window contributes a point only when the papilla is detected in that clip; no location is interpolated or assigned a default coordinate. Both plots retain the same detector-coordinate system, with no post-hoc spatial alignment. The displayed smoothing shows the episode-level trend only. The main-paper Box, Recall, Centerline, and Endpoint metrics use unsmoothed framewise detections under the rule in Section D.2. These traces therefore complement the aggregate errors temporally; they are not a test of long autoregressive rollout stability.

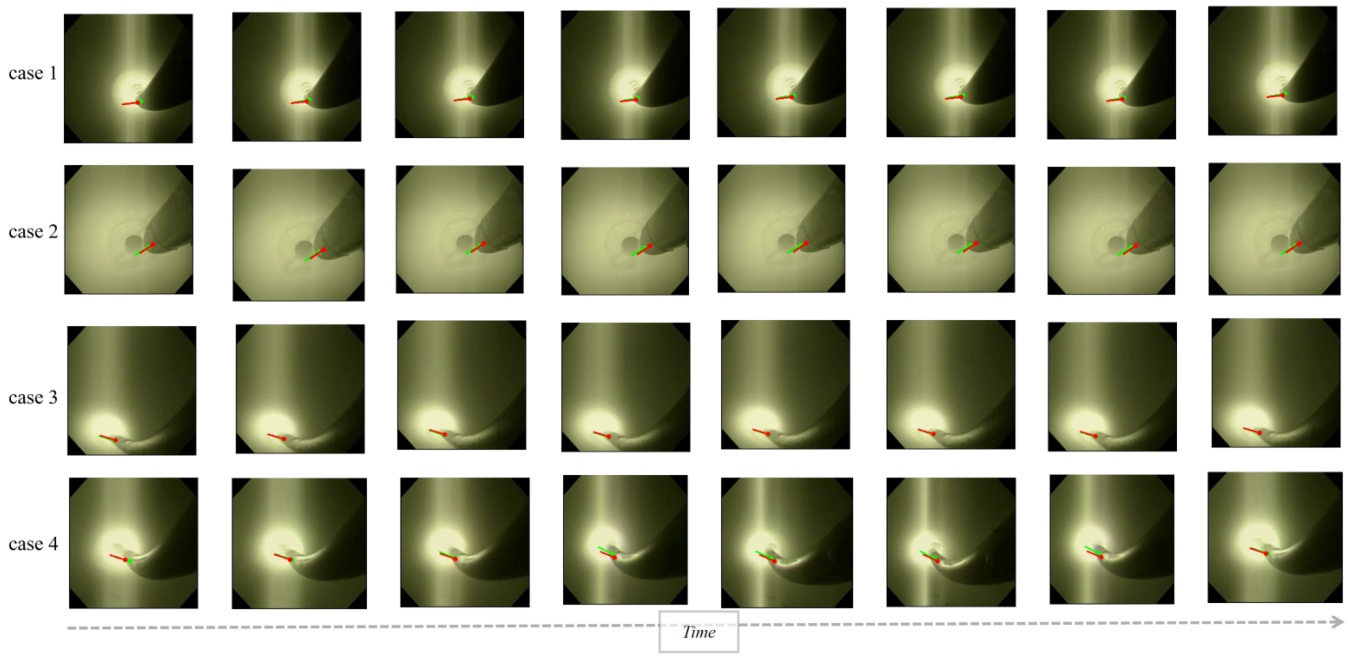

Figure 9: Visual validation of Pose Readout. Four held-out phantom windows at step 10,000 are shown over their nine temporally sampled Mother states. Red markers and rays denote Pose Readout predictions; green markers and rays denote mask-derived targets.

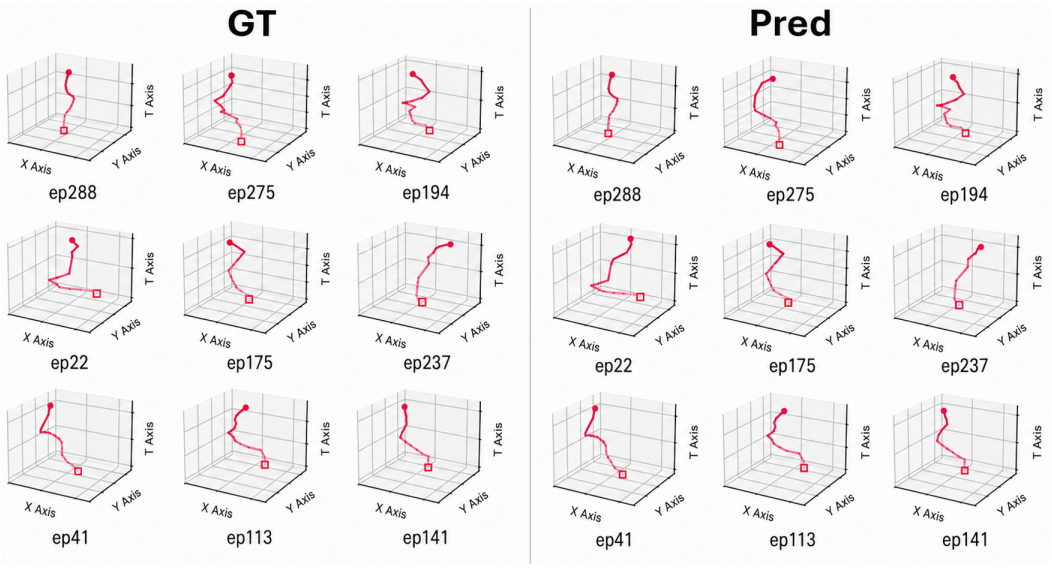

Figure 10: Output-level Child papilla trajectories for nine selected held-out phantom episodes at step 10,000. The left and right blocks compare target and CrossScope-generated mean bounding-box centers from 20 independent nine-frame windows sampled across each episode, in (X, Y, T) space. Open squares and filled circles mark the first and final detections.