ENDOWAM: A GROUNDED WORLD-ACTION MODEL FOR GENERALIZABLE ENDOSCOPIC NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Autonomous endoscopic navigation remains challenging due to tissue deformation, transient occlusions, and rapid viewpoint changes. Existing policies typically predict actions from current observations without explicitly modeling future anatomical dynamics, limiting robustness in safety-critical settings. We introduce **EndoWAM**, to our knowledge the first World Action Model (WAM) for generalizable robotic endoscopic navigation. EndoWAM introduces *future grounding*, which predicts task-relevant target regions in future observations from intermediate denoising features of a video world model. A lightweight diffusion transformer for future target prediction is coupled with a discrete action expert through a shared predictive representation, enabling target-aware dynamics modeling and real-time control with a single denoising pass. We further introduce **EndoMotion**, a robotic endoscopic motion dataset covering ureteroscopy, esophagoscopy, and endoscopic retrograde cholangiopancreatography (ERCP). EndoWAM consistently outperforms strong baselines and alternative grounding strategies, while exhibiting zero-shot generalization to unseen viewpoints, environments, and targets. These results demonstrate the effectiveness of predictive, target-grounded world-action modeling for robust and generalizable navigation in deformable endoscopic environments.

1 INTRODUCTION

Autonomous endoscopic navigation can substantially alleviate clinicians' operational burden during interventional procedures. Achieving robust and consistent navigation is essential across diverse endoscopic applications, including ureteroscopy, esophagoscopy, and endoscopic retrograde cholangiopancreatography (ERCP) Alian et al. (2023); Buxbaum et al. (2023). However, tissue deformation during intervention induces complex and rapidly evolving visual dynamics, while transient occlusions further degrade visual observations, posing significant challenges to vision-based navigation methods. Recent learning-based approaches, including imitation learning (IL) and vision-language-action (VLA) models, have been investigated for specific surgical and endoscopic tasks Kim et al. (2025); Lin et al. (2026). Nevertheless, these approaches typically predict actions from current observations without explicitly modeling future environment dynamics. This limited predictive capability hinders their deployment in safety-critical surgical scenarios, where reliable and predictable behavior is essential.

World Action Models (WAMs) have recently emerged as a promising paradigm for robotic manipulation and visuomotor control Li et al. (2026; 2025a); Bi et al. (2025). Unlike VLA models that directly map observations to actions, WAMs incorporate inverse-dynamics objectives or jointly model visual observations and action distributions, enabling them to capture physical dynamics and task-relevant temporal structure Kim et al. (2026); Ma et al. (2026); Ye et al. (2026). Extending WAMs to robotic endoscopy, however, presents several challenges. First, ethical constraints and the high costs of clinical procedures limit the collection of large-scale navigation data. Second, confined anatomical workspaces restrict viewpoint diversity, providing sparse supervision for learning visual dynamics. Third, deformable anatomy and transient occlusions complicate target-directed navigation and increase the risk of unsafe tissue interactions. Finally, the high inference latency of existing WAMs limits their responsiveness to rapidly changing endoscopic observations, hindering deployment in time-sensitive surgical scenarios.

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.

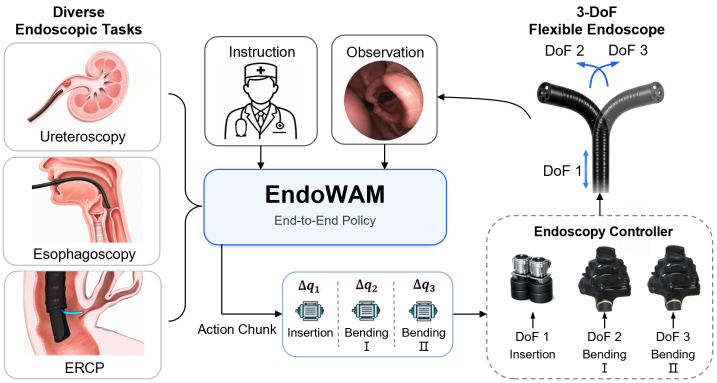


Figure 1: **Overview of EndoWAM for autonomous navigation across diverse endoscopic tasks.** The policy maps the current observation and instruction to action chunks that control the three degrees of freedom of a flexible endoscope.

We address these challenges with **EndoWAM**, which is, to our knowledge, the first World Action Model for generalizable robotic endoscopic navigation, as illustrated in Fig. 1. EndoWAM introduces a future grounding paradigm that uses latent-space supervision from task-relevant regions in future observations to shape predictive visual representations. Specifically, a diffusion transformer (DiT), conditioned on intermediate denoising features from a video world model, predicts future target-region latents. Unlike previous grounding methods that localize targets only in the current observation Huang et al. (2025); Lai et al. (2024); Li et al. (2025c); Khanam & Hussain (2024); Zawalski et al. (2024); Deng et al. (2025); Song et al. (2025), future grounding captures the spatial evolution of task-relevant targets. This design promotes fine-grained, target-aware temporal representations, enabling the action policy to exploit predictive visual dynamics for robust and precise endoscopic navigation. To alleviate the scarcity of temporally aligned endoscopic navigation data, we further introduce **EndoMotion**, a multi-procedure robotic endoscopy dataset comprising long-horizon trajectories across ureteroscopy, esophagoscopy, and ERCP. We collect human teleoperation trajectories in corresponding anatomical phantoms and use a fine-tuned Grounding DINO Liu et al. (2024) to densely annotate task-relevant target regions. These annotations provide target-relative action labels and future target regions for grounding supervision. We additionally introduce viewpoint augmentation by jointly rotating each trajectory and its target annotations across multiple roll angles, substantially increasing visual diversity and viewpoint coverage.

Experiments across diverse long-horizon robotic endoscopic navigation tasks demonstrate that our future grounding paradigm consistently outperforms existing visual grounding strategies. Attention visualizations further show that EndoWAM focuses on task-relevant target regions, enabling precise and temporally coherent navigation. Ablation studies confirm that viewpoint augmentation substantially improves generalization to unseen viewpoints. Moreover, EndoWAM exhibits strong zero-shot generalization to unseen environments and targets. Collectively, these results establish EndoWAM as a robust, target-aware, and generalizable framework for long-horizon navigation in complex and visually constrained endoscopic environments.

In summary, our main contributions are as follows:

- We propose **EndoWAM**, which is, to our knowledge, the **first** World Action Model for generalizable robotic endoscopic navigation. By conditioning the policy on intermediate denoising features of a video world model, EndoWAM incorporates predictive visual dynamics while enabling real-time navigation.
- We introduce a future grounding paradigm that predicts task-relevant regions in future latent states, enabling fine-grained, target-aware temporal representation learning for precise endoscopic navigation.
- We construct **EndoMotion**, a robotic endoscopy motion dataset spanning ureteroscopy, esophagoscopy, and ERCP, with dense target annotations, action labels, and viewpoint augmentation.

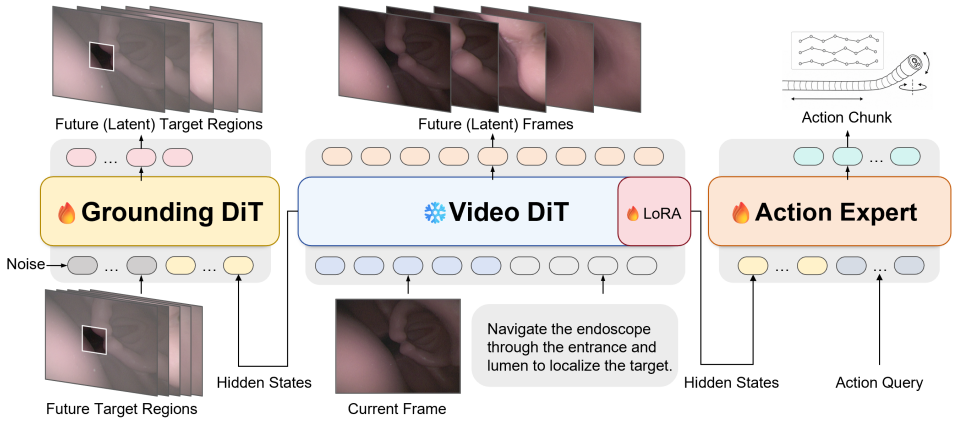


Figure 2: **EndoWAM architecture.** The Video DiT models future latent dynamics from the current frame and instruction. Its hidden states condition the Grounding DiT to reconstruct future target regions and the action expert to generate an action chunk.

- Extensive real-world experiments demonstrate state-of-the-art performance, validate the proposed grounding strategies, and show that the policy generalizes effectively to unseen viewpoints, environments, and targets.

2 Related Work

Vision-Language-Action and World Action Models. Learning-based visuomotor control has evolved from imitation policies Chi et al. (2023); Zhao et al. (2023) to VLAs Kim et al. (2024b); Black et al. (2024); Physical Intelligence et al. (2025). Both learn action policies from demonstrations, while VLAs additionally inherit semantic reasoning capabilities from pretrained vision-language models (VLMs). However, these methods typically predict actions from current observations without explicitly modeling how the environment evolves under robot actions. More recently, World Action Models (WAMs) Li et al. (2026); Cen et al. (2025); Li et al. (2025a); Bi et al. (2025); Kim et al. (2026); Ma et al. (2026); Ye et al. (2026) have incorporated inverse-dynamics objectives or jointly modeled visual observations and action distributions, often building on large-scale video world models Agarwal et al. (2025); Assran et al. (2025). Their predictive representations capture physical dynamics and task-relevant temporal structure beyond the current observation. While existing WAMs primarily target robotic manipulation, EndoWAM extends predictive visuomotor modeling to deformable and partially observable anatomical environments and grounds future visual dynamics in task-relevant target regions.

Robot Learning for Autonomous Surgical Robotics. Surgical robotics is increasingly transitioning from rule-based control to data-driven policies capable of interacting with complex and deformable anatomy Schmidgall et al. (2024); Long et al. (2025); Kim et al. (2025). Most existing approaches rely on imitation learning Kim et al. (2024a); Moghani et al. (2025), while reinforcement learning has also been explored under simulated supervision Singh et al. (2023). In flexible endoscopy, autonomous navigation is typically addressed through visual servoing or platform-specific reinforcement learning Duan et al. (2023); Tian et al. (2024); Zhang et al. (2024). More recently, VLA-based methods have been investigated for scrub-nurse support Li et al. (2025b), instrument tracking Ng et al. (2025), and biliary navigation Lin et al. (2026). However, these methods primarily infer actions from current observations without explicitly modeling future anatomical dynamics. In contrast, EndoWAM conditions action generation on predictive intraluminal representations and, to our knowledge, is the first World Action Model for generalizable robotic endoscopic navigation.

Visual Grounding Methods for Robotic Control. Fine-grained robotic control requires policies to identify and attend to task-relevant visual regions. Existing grounding methods either rely on external perception models or learn grounding jointly with the control policy.

RoboGround Huang et al. (2025) uses LISA Lai et al. (2024) to segment target and background regions as auxiliary inputs, while VIP Li et al. (2025c) crops and enlarges target regions detected by YOLOv11 Khanam & Hussain (2024); both therefore depend on the accuracy of external grounding models. Alternatively, ECoT Zawalski et al. (2024) and GraspVLA Deng et al. (2025) predict bounding boxes before action generation, whereas ReconVLA Song et al. (2025) predicts gaze regions from policy representations to provide implicit grounding supervision. Despite their different formulations, these approaches ground task-relevant regions only in the current observation. In contrast, EndoWAM introduces future grounding, which predicts task-relevant target regions in future observations from intermediate world model representations, thereby injecting target-centric supervision into predictive visual dynamics.

3 METHOD

3.1 PROBLEM FORMULATION

Unlike conventional VLA policies that directly map current visual observations o_t and language instructions l to actions a_t, EndoWAM adopts a *predict-latent-dynamics-then-ground-and-act* paradigm. Given o_t and l, the video world model predicts a latent representation of future visual dynamics $\mathbf{z}_{t+1}$, whose intermediate denoising features jointly support multi-step action generation and future target grounding:

$$\begin{aligned} \mathbf{z}_{t+1} &\sim p_v(\cdot \mid o_t, l)\,, \\ \mathbf{a}_t &\sim p_a\big(\cdot \mid \mathcal{H}\big(\mathbf{z}_{t+1}^{\tau_v}\big)\big)\,, \\ \mathbf{r}_{t+1} &\sim p_r\big(\cdot \mid \mathcal{H}\big(\mathbf{z}_{t+1}^{\tau_v}\big)\big)\,. \end{aligned} \tag{1}$$

Here, $\mathbf{z}_{t+1}^{\tau_v}$ denotes the intermediate latent state at flow step τ_v, which progressively approaches $\mathbf{z}_{t+1}$ as $\tau_v \to 0$. The operator $\mathcal{H}$ extracts hidden representations from the denoising process without explicitly decoding future observations into pixel-space frames.

The variable $\mathbf{r}_{t+1}$ denotes task-relevant target regions in future observations. Training jointly models future visual dynamics, multi-step actions, and target regions through the conditional latent-action-region distribution:

$$\mathbf{z}_{t+1}, \mathbf{a}_t, \mathbf{r}_{t+1} \sim p_{var}(\cdot \mid o_t, l)\,. \tag{2}$$

3.2 MODEL ARCHITECTURE

Overview. EndoWAM is designed for generalizable endoscopic navigation, generating instruction-conditioned three-DoF controls that steer a flexible endoscope toward anatomical targets despite occlusions and specular artifacts. Rather than relying solely on the current observation, EndoWAM grounds its policy in predicted future dynamics. During training, a video world model captures future visual dynamics in a single denoising pass, and the resulting shared representation conditions an action decoder for multi-step control and a Grounding DiT for reconstructing future target-region latents (Fig. 2). At inference the world model still performs the same single denoising pass to produce the shared representation, but future-frame decoding and the Grounding DiT branch are dropped, so future-grounded representation learning adds no deployment cost.

Architecture. The world model is initialized from Cosmos-Predict2.5-2B Agarwal et al. (2025), a pretrained video DiT with a text encoder and video VAE. The instruction is encoded and injected into the DiT tokens through cross-attention, while endoscopic observations are mapped to latent video tokens by the VAE. The current observation is retained as a clean conditioning latent, whereas future latent tokens are perturbed under a rectified-flow schedule and jointly denoised. We apply LoRA Hu et al. (2022) to the attention and feed-forward projections to efficiently adapt the pretrained video prior to the endoscopic domain.

To instantiate $\mathcal{H}$ in Eq. 1, we extract an intermediate activation after a single denoising pass as the shared representation $H = \mathcal{H}(\mathbf{z}_{t+1}^{\tau_v})$ and flatten it into a sequence of tokens. Two prediction heads are built upon H. The action decoder processes learnable action queries through transformer blocks that combine bidirectional self-attention with cross-attention to H, followed by an MLP that predicts per-axis categorical logits. The bidirectional attention enables each action query to incorporate the full temporal context of the action chunk. The Grounding DiT reconstructs the VAE latents of target

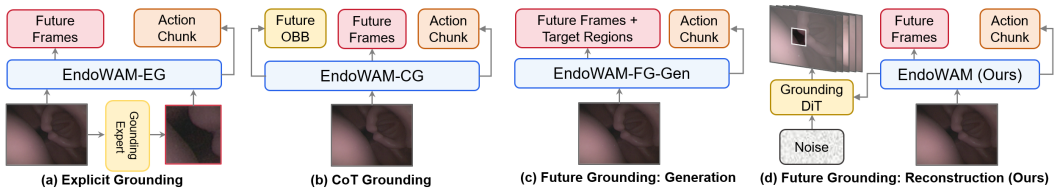


Figure 3: **Conceptual comparison of different paradigms in EndoWAM.** (a) **EndoWAM-EG** performs explicit input grounding by injecting a grounding-expert crop of the current target. (b) **EndoWAM-CG** performs CoT grounding by regressing future OBBs from the shared predictive representation. (c) **EndoWAM-FG-Gen** jointly generates future frames and target regions through the world model's generative pathway. (d) **EndoWAM (Ours)** performs future grounding by reconstructing the latent target regions of future frames with a Grounding DiT.

crops in future frames, conditioned on a projected and resampled representation of H. This future grounding objective encourages H to capture the spatial evolution of task-relevant targets, yielding predictive features for action decoding.

3.3 Training and Inference

Dataset. To provide aligned control and grounding supervision for endoscopic navigation, we construct **EndoMotion**, comprising 515 trajectories and 0.98M frames collected from three procedures: ureteroscopy, esophagoscopy, and ERCP. We use Grounding DINO Liu et al. (2024) to localize the instructed anatomical targets and apply Kalman filtering to obtain temporally consistent oriented bounding-box (OBB) tracks. Each track provides two aligned supervision signals: the target-center offset relative to the endoscopic view is converted into discrete action labels, while target regions in future frames are cropped to provide future grounding supervision. To improve robustness to viewpoint variations, we further introduce viewpoint augmentation by rotating each trajectory and its corresponding OBB annotations at 45° intervals, expanding the dataset to 4,120 sequences and 7.84M frames. By aligning observations, target-relative actions, and future target regions across anatomically distinct procedures and viewpoints, EndoMotion supports the joint learning of predictive dynamics, action generation, and future grounding. Per-procedure statistics are reported in Appendix C, and the full annotation pipeline is detailed in Appendix B.

Training objective. We jointly optimize EndoWAM for predictive visual dynamics, multi-step action generation, and future target grounding, using three corresponding objectives. For video co-training, we sample $\varepsilon \sim \mathcal{N}(0, I)$ and a flow time τ, interpolate the future latents as $x_\tau = (1-\tau)x_0 + \tau\varepsilon$, and regress the velocity field over future positions:

$$\mathcal{L}_{\text{video}} = \mathbb{E}_{\tau,\varepsilon}\Big[\big\|v_\theta(x_\tau, \tau \mid o_t, l) - (\varepsilon - x_0)\big\|_2^2\Big]. \tag{3}$$

Here, τ is resampled for each training example, whereas τ_v in Eq. 1 denotes the flow time at which the shared representation H is extracted.

The action decoder is trained with per-axis weighted cross-entropy under a validity mask m:

$$\mathcal{L}_{\text{action}} = \frac{\sum_{k,a} m_{k,a}\,\mathrm{CE}\big(p_\theta(\cdot \mid q_k, H)_a, y_{k,a}; w_a\big)}{\sum_{k,a} m_{k,a}}, \tag{4}$$

where k indexes the action chunk, a indexes the three motion axes, $y_{k,a}$ is the target action bin, and w_a denotes the corresponding class weights.

For future grounding, the Grounding DiT denoises the target-region latent $\mathbf{z}_n^{\text{tar}}$ of future frame n at diffusion step s under a cosine schedule:

$$\mathcal{L}_{\text{ground}} = \mathbb{E}_{n,s,\varepsilon}\Big[u_n\big\|\varepsilon - \varepsilon_\theta\big(\sqrt{\bar\alpha_s}\mathbf{z}_n^{\text{tar}} + \sqrt{1-\bar\alpha_s}\varepsilon, s, H\big)\big\|_2^2\Big], \tag{5}$$

where the target mask $u_n \in \{0,1\}$ restricts supervision to future frames with detected targets, so missed detections contribute no gradient. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{action}} + \lambda_v\mathcal{L}_{\text{video}} + \lambda_r\mathcal{L}_{\text{ground}}, \tag{6}$$

where λ_v and λ_r balance action learning, video co-training, and future grounding.

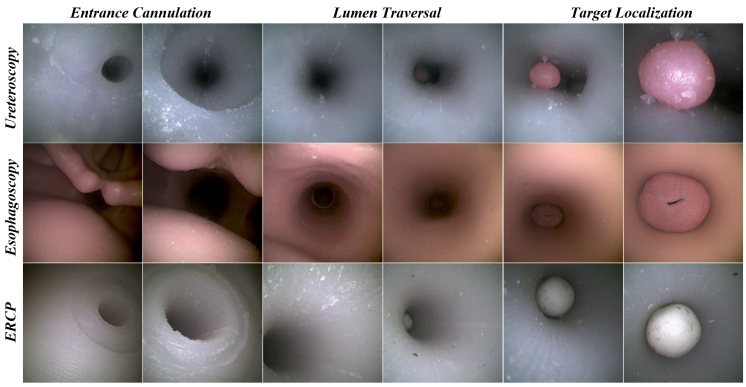


Figure 4: **Closed-loop deployment trajectories.** Representative EndoWAM rollouts across ureteroscopy, esophagoscopy, and ERCP, covering the three navigation phases: entrance cannulation, lumen traversal, and target localization.

3.4 Controlled Variants for Grounding WAM

To investigate how visual attention allocation affects grounding in WAMs, we construct a set of controlled variants inspired by representative grounding strategies in VLA models, as grounding in WAMs remains underexplored. For a fair comparison, all variants share the same backbone, tokenization scheme, and training protocol whenever applicable.

As illustrated in Fig. 3, EndoWAM-EG performs explicit input grounding by augmenting current observation with a cropped and resized target region produced by a fine-tuned grounding expert. EndoWAM-CG directly regresses future OBB parameters $(c_x, c_y, w, h, \sin 2\theta, \cos 2\theta)$ from the shared representation H. EndoWAM-FG-Gen instead incorporates future target regions into the world model's generation target, allowing grounding to be learned through the generative pathway rather than a separate reconstruction branch. These variants represent input-level, regression-based, and generative grounding strategies, respectively, and are evaluated under matched settings in Sec. 4.3.

4 Experiments

We design our experiments around the following questions:

- Can EndoWAM outperform state-of-the-art methods on long-horizon endoscopic navigation under unseen viewpoints? (Sec. 4.2)
- Does our **future grounding** approach outperform alternative visual grounding paradigms? (Sec. 4.3)
- How do the grounding paradigm and other key design choices affect the overall performance of EndoWAM? (Sec. 4.4)
- Can EndoWAM generalize to unseen environments and unseen targets? (Sec. 4.5)
- Is EndoWAM efficient enough for closed-loop endoscopic control compared with existing methods? (Sec. 4.6)

4.1 Implementation Details

Deployment and evaluation protocol. All policies are deployed on the same three-DoF robotic endoscope and evaluated on physical phantoms for ureteroscopy, esophagoscopy, and ERCP. A trial is considered successful if the endoscope reaches the instructed target and maintains it within the field of view until the rollout terminates; failures are categorized as lumen loss, wall impaction, or timeout. Each policy is evaluated over 32 trials per procedure. For each procedure, we randomly sample an initial viewpoint and apply uniformly spaced roll offsets spanning the full

Table 1: **Closed-loop navigation success rates (%) across three endoscopic procedures.** For each procedure, 32 trials are initialized from a randomly sampled viewpoint and evaluated at uniformly spaced roll offsets spanning the full 360° range, with test orientations offset from the eight roll angles used for training augmentation.

Method	Discrete Action	Visual Grounding	Dynamics Prior	Ureteroscopy	Esophagoscopy	ERCP	**Average**
Diffusion Policy Chi et al. (2023)	×	×	×	3.1	0.0	3.1	2.1
Qwen3DiT	×	×	×	9.4	3.1	12.5	8.3
$\pi_{0.5}$ Physical Intelligence et al. (2025)	×	×	×	18.8	6.3	21.9	15.6
Qwen3AE	✓	×	×	15.6	12.5	25.0	17.7
GR00T-N1.7 Bjorck et al. (2025)	×	×	×	25.0	21.9	34.4	27.1
EndoWAM (Ours)	✓	✓	✓	**90.6**	**56.3**	**93.8**	**80.2**
EndoWAM variants							
EndoWAM-CG	✓	✓	✓	46.9	37.5	53.1	45.8
EndoWAM-FG-Gen	✓	✓	✓	43.8	46.9	62.5	51.0
EndoWAM-EG	✓	✓	✓	71.9	43.8	68.8	61.5
EndoWAM (Ours)	✓	✓	✓	**90.6**	**56.3**	**93.8**	**80.2**

360° range. These test viewpoints are deliberately unaligned with the eight angles used for viewpoint augmentation, ensuring that every trial evaluates generalization to an unseen viewpoint. Representative EndoWAM rollouts are shown in Fig. 4, with additional demonstrations provided in Appendix F. Further details on the robotic platform, anatomical phantoms, and success criteria are provided in Appendix A.

Baselines. We compare EndoWAM against five baselines. Diffusion Policy Chi et al. (2023) serves as a non-VLA visuomotor policy, while $\pi_{0.5}$ Physical Intelligence et al. (2025) and GR00T-N1.7 Bjorck et al. (2025) represent large-scale pretrained VLA policies. Qwen3DiT and Qwen3AE share the same Qwen3-VL-2B Bai et al. (2025) backbone but differ in their action heads: the former adopts a diffusion transformer (DiT) action head, whereas the latter uses our discretized action expert (AE). This controlled comparison isolates the effect of action parameterization under a shared visual prior. All baselines receive identical observations and instructions and are trained on the same rotation-augmented EndoMotion dataset for a matched number of optimization steps. Detailed training configurations for each baseline are provided in Appendix E.

Model and optimization. EndoWAM adapts Cosmos-Predict2.5-2B Agarwal et al. (2025) using LoRA Hu et al. (2022) on the attention and feed-forward projections, while keeping the text encoder and video VAE frozen. The shared representation H is extracted from the block-17 activation after a single denoising pass and jointly conditions both the action decoder and the Grounding DiT. Architectural and optimization details are provided in Appendix D.

4.2 Comparison with State-of-the-Art

We compare EndoWAM with representative imitation-learning and VLA policies under the protocol described in Sec. 4.1. To evaluate generalization to unseen viewpoints, all methods are tested over uniformly spaced roll offsets spanning the full 360° range, with test viewpoints unaligned with those used for training augmentation. The comparison spans action-generation mechanisms and examines the contributions of discrete action prediction, visual grounding, and predictive dynamics modeling.

Results. The upper block of Table 1 summarizes the comparison. EndoWAM succeeds in 77 of 96 trials, achieving an average success rate of 80.2% and outperforming the strongest baseline, GR00T-N1.7 (27.1%), by 53.1 percentage points. It also achieves the highest success rate across all three procedures, with 90.6% on ureteroscopy, 56.3% on esophagoscopy, and 93.8% on ERCP. A breakdown of the observed failure modes is reported in Appendix G. These results demonstrate that predictive dynamics priors and our future grounding approach consistently benefit diverse endoscopic navigation tasks, while enabling strong generalization to unseen viewpoints. Moreover, Qwen3AE outperforms Qwen3DiT under the same vision-language backbone, validating the effectiveness of our discrete action expert over the diffusion-based action head in this setting.

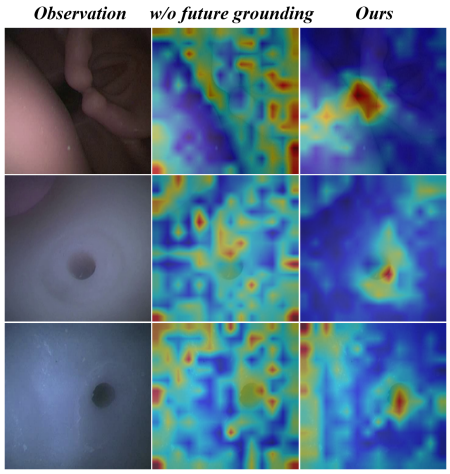


Figure 5: Future grounding concentrates attention on task-relevant targets despite darkness and occlusion, while its removal leads to diffuse attention.

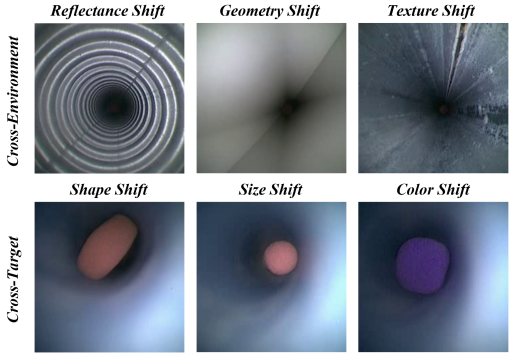


Figure 6: **Zero-shot deployment examples.** Representative observations from EndoWAM rollouts under unseen cross-environment shifts in lumen reflectance, geometry, and texture, and cross-target shifts in target shape, size, and color.

4.3 Comparison with EndoWAM Variants

We evaluate the controlled grounding variants introduced in Fig. 3 and Sec. 3.4 under the same closed-loop protocol and unseen-viewpoint setting as in Sec. 4.2. All variants share matched training and evaluation configurations, with results reported in the lower block of Table 1.

Results. EndoWAM achieves the highest average success rate of 80.2%, outperforming the strongest variant, EndoWAM-EG, by 18.7 points, with gains of 18.7, 12.5, and 25.0 points on ureteroscopy, esophagoscopy, and ERCP, respectively. Directly regressing future OBBs in EndoWAM-CG yields an average success rate of 45.8%, while incorporating future target regions into the generative objective improves performance to 51.0%. Explicit input grounding further increases the average success rate to 61.5%. Nevertheless, all variants remain substantially below our reconstruction-based future grounding, indicating that reconstructing future target regions from predictive denoising features more effectively couples visual dynamics, target-aware attention, and multi-step action generation.

4.4 Ablation Study

Table 2 evaluates the future grounding and viewpoint augmentation through independent single-factor ablations. Removing the future groundingreduces the average success rate from 80.2% to 37.5%, demonstrating the importance of future target supervision for reliable navigation. Removing viewpoint augmentation lowers the average success rate to 28.1%, indicating that exposure to diverse viewpoints is critical for generalization to unseen orientations. Figure 5 provides qualitative evidence: without the Grounding DiT, policy attention is diffuse and frequently captured by specular highlights and mucosal textures, whereas the full model concentrates its attention around the task-relevant target even under darkness and partial occlusion. These results show that the future groundingpromotes target-aware predictive representations, while viewpoint augmentation improves viewpoint robustness.

Table 2: **Ablation of future grounding and viewpoint augmentation.** Closed-loop navigation success rates (%) across three procedures.

Variant	Ureteroscopy	Esophagoscopy	ERCP	**Avg.**
w/o future grounding	37.5	28.1	46.9	37.5
w/o viewpoint aug.	31.3	15.6	37.5	28.1
Full model	**90.6**	**56.3**	**93.8**	**80.2**

Table 3: **Zero-shot generalization to unseen environments and targets.** Success rates (%) over 32 trials per condition for lumen traversal and target localization; results are not directly comparable to the three-stage evaluation in Table 1.

	Cross-environment			Cross-target			
Method	Reflectance Shift	Geometry Shift	Texture Shift	Shape Shift	Size Shift	Color Shift	**Average**
Qwen3AE	31.3	28.1	40.6	28.1	53.1	37.5	36.5
GR00T-N1.7 Bjorck et al. (2025)	53.1	37.5	53.1	43.8	59.4	34.4	46.9
EndoWAM (Ours)	**93.8**	**87.5**	**100.0**	**84.4**	**93.8**	**81.3**	**90.1**

4.5 Generalization Capability

Evaluation setup. We evaluate zero-shot generalization across six conditions unseen during training. As shown in Fig. 6, cross-environment shifts alter lumen reflectance, geometry, and texture, whereas cross-target shifts alter target shape, size, and color. Each condition comprises 32 closed-loop trials following a two-stage protocol of lumen traversal and target localization.

Results. As reported in Table 3, EndoWAM achieves an average success rate of 90.1%, outperforming the strongest baseline, GR00T-N1.7, by 43.2 percentage points. EndoWAM maintains consistently high performance across all conditions, reaching 100.0% under texture shift and exceeding 80% across all target variations. These results demonstrate that jointly modeling visual dynamics and target-aware grounding enables robust transfer to unseen lumen appearances and target configurations.

4.6 Deployment Efficiency

Table 4 compares closed-loop inference efficiency on a single NVIDIA RTX 5090. At a comparable parameter scale, EndoWAM achieves a competitive control frequency of 7.5 Hz relative to VLA policies, while retaining an explicit dynamics prior. Moreover, EndoWAM is $6.8\times$ faster than Cosmos Policy (1.1 Hz), because it predicts an action chunk from intermediate features obtained in a single denoising pass, whereas Cosmos Policy must iteratively denoise future video representations before generating actions. Combined with the improved navigation success rates reported in Table 1, these results demonstrate that EndoWAM enhances navigation accuracy while meeting the real-time control requirements of endoscopic navigation.

Table 4: **Closed-loop deployment efficiency with a single NVIDIA RTX 5090.**

Policy	Dynamics Prior	Params.	Control Freq. (Hz)
Qwen3AE	×	2.2B	12.1
GR00T-N1.7	×	1.6B	8.9
Cosmos Policy Kim et al. (2026)	✓	2.0B	1.1
EndoWAM (Ours)	✓	2.1B	**7.5**

5 Conclusion

We presented **EndoWAM**, which is, to our knowledge, the first World Action Model for generalizable robotic endoscopic navigation. EndoWAM introduces future grounding, which predicts task-relevant regions in future observations to inject target-aware supervision into intermediate world-model representations. This design enables the policy to capture future visual dynamics and target evolution, improving multi-step navigation under occlusions, visual degradation, and viewpoint changes. We further constructed **EndoMotion**, a multi-procedure robotic endoscopy dataset with aligned visual observations, action labels, future target annotations, and viewpoint augmentation. Closed-loop experiments on physical phantoms for ureteroscopy, esophagoscopy, and ERCP show that EndoWAM achieves an average success rate of 80.2%, outperforming the strongest baseline by 53.1 percentage points. It also attains a 90.1% average zero-shot success rate under unseen environment and target shifts. By generating action chunks from shared predictive representations in a single denoising pass, EndoWAM operates at 7.5 Hz, demonstrating a strong balance among navigation accuracy, generalization, and real-time control.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing of this paper, including language editing and LaTeX formatting. The authors take responsibility for the final content.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithviraj Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical AI, 2025.

Amirhosein Alian, Emilia Zari, Zeyu Wang, Enrico Franco, James P Avery, Mark Runciman, Benny Lo, Ferdinando Rodriguez y Baena, and George Mylonas. Current engineering developments for robotic systems in flexible endoscopy. *Techniques and Innovations in Gastrointestinal Endoscopy*, 25(1):67–81, 2023.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report, 2025.

Hao Bi, Hao Tan, Sheng Xie, Zhi Wang, Siteng Huang, Hao Liu, Rui Zhao, Yao Feng, Chao Xiang, Yu Rong, et al. Motus: A unified latent action world model, 2025.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. GR00T N1: An open foundation model for generalist humanoid robots, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A Vision-Language-Action flow model for general robot control, 2024.

James L Buxbaum, Martin Freeman, Stuart K Amateau, Jean M Chalhoub, Nayantara Coelho-Prabhu, Madhav Desai, Sherif E Elhanafi, Nauzer Forbes, Larissa L Fujii-Lau, Divyanshoo R Kohli, et al. American society for gastrointestinal endoscopy guideline on post-ERCP pancreatitis prevention strategies: summary and recommendations. *Gastrointestinal Endoscopy*, 97(2):153–162, 2023.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. WorldVLA: Towards autoregressive action world model, 2025.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 2023.

S. Deng, M. Yan, S. Wei, H. Ma, Y. Yang, J. Chen, Z. Zhang, T. Yang, X. Zhang, H. Cui, et al. GraspVLA: A grasping foundation model pre-trained on billion-scale synthetic action data, 2025.

Xingguang Duan, Dongsheng Xie, Runtian Zhang, Xiaotian Li, Jiali Sun, Chao Qian, Xinya Song, and Changsheng Li. A novel robotic bronchoscope system for navigation and biopsy of pulmonary lesions. *Cyborg and Bionic Systems*, 4:0013, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

H. Huang, X. Chen, Y. Chen, H. Li, X. Han, Z. Wang, T. Wang, J. Pang, and Z. Zhao. RoboGround: Robotic manipulation with grounded vision–language priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22540–22550, 2025.

Rahima Khanam and Muhammad Hussain. YOLOv11: An overview of the key architectural enhancements, 2024.

Ji Woong Kim, Tony Z Zhao, Samuel Schmidgall, Anton Deguet, Marin Krieger, Chelsea Finn, and Axel Krieger. Surgical robot transformer (SRT): Imitation learning for surgical tasks, 2024a.

Ji Woong Kim, Juo-Tung Chen, Pascal Hansen, Lucy Xiaoyang Shi, Antony Goldenberg, Samuel Schmidgall, Paul Maria Scheikl, Anton Deguet, Brandon M White, De Ru Tsai, et al. SRT-H: A hierarchical framework for autonomous surgery via language-conditioned imitation learning. *Science Robotics*, 10(104):eadt5254, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. OpenVLA: An open-source Vision-Language-Action model. In *Proceedings of the 8th Conference on Robot Learning (CoRL)*, 2024b.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning, 2026.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Lei Li, Qi Zhang, Yao Luo, Sheng Yang, Rui Wang, Feng Han, Ming Yu, Zhen Gao, Nan Xue, Xin Zhu, et al. Causal world modeling for robot control, 2026.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model, 2025a.

Shunlei Li, Jin Wang, Rui Dai, Wanyu Ma, Wing Yin Ng, Yingbai Hu, and Zheng Li. RoboNurse-VLA: Robotic scrub nurse system based on Vision-Language-Action model. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3986–3993. IEEE, 2025b.

Z. Li, L. Ren, J. Yang, Y. Zhao, X. Wu, Z. Xu, X. Bai, and H. Zhao. VIP: Vision instructed pre-training for robotic manipulation. In *Forty-second International Conference on Machine Learning (ICML)*, 2025c.

Jinsong Lin, Chi Kit Ng, Zhiyong Xiong, Zikang Pan, Yihan Hu, Tabassum Tamima, Ziyi Hao, Eddie Cheung, Jiewen Lai, Huxin Gao, and Hongliang Ren. BiliVLA: Scene-aware vision-language-action model with reinforcement learning for autonomous biliary endoscopic navigation, 2026.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2024.

Yonghao Long, An Lin, Danny H C Kwok, Lin Zhang, Zhaoshuo Yang, Kang Shi, Lei Song, Jing Fu, Hao Lin, Wei Wei, et al. Surgical embodied intelligence for generalized task autonomy in laparoscopic robot-assisted surgery. *Science Robotics*, 10(104):eadt3093, 2025.

Tao Ma, Jie Zheng, Zhi Wang, Chao Jiang, Ao Cui, Jian Liang, and Sheng Yang. Dit4dit: Jointly modeling video dynamics and actions for generalizable robot control, 2026.

Masoud Moghani, Nigel Nelson, Mohamed Ghanem, Andres Diaz-Pinto, Kush Hari, Mahdi Azizian, Ken Goldberg, Sean Huver, and Animesh Garg. SuFIA-BC: Generating high quality demonstration data for visuomotor policy learning in surgical subtasks. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025.

Chi Kit Ng, Long Bai, Guankun Wang, Yupeng Wang, Huxin Gao, Chenhan Jin, Tieyong Zeng, Hongliang Ren, et al. EndoVLA: Dual-phase Vision-Language-Action for precise autonomous tracking in endoscopy. In *9th Annual Conference on Robot Learning*, 2025.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a Vision-Language-Action model with open-world generalization, 2025.

Samuel Schmidgall, Ji Woong Kim, Alan Kuntz, Ahmed Ezzat Ghazi, and Axel Krieger. General-purpose foundation models for increased autonomy in robot-assisted surgery. *Nature Machine Intelligence*, 6(11):1275–1283, 2024.

Amritpal Singh, Wenqi Shi, and May D Wang. Autonomous soft tissue retraction using demonstration-guided reinforcement learning, 2023.

Wenxuan Song, Ziyang Zhou, Han Zhao, Jiayi Chen, Pengxiang Ding, Haodong Yan, Yuxin Huang, Feilong Tang, Donglin Wang, and Haoang Li. ReconVLA: Reconstructive vision-language-action model as effective robot perceiver, 2025.

Qingyao Tian, Huai Liao, Xinyan Huang, Jian Chen, Zihui Zhang, Bingyu Yang, Sebastien Ourselin, and Hongbin Liu. DD-VNB: A depth-based dual-loop framework for real-time visually navigated bronchoscopy. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 12979–12986. IEEE, 2024.

Ao Ye, Bo Wang, Chao Ni, Guan Huang, Guang Zhao, Hao Li, Hui Li, Jie Li, Jun Lv, Jian Liu, et al. Gigaworld-policy: An efficient action-centered world–model, 2026.

Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. In *Proceedings of the 8th Conference on Robot Learning (CoRL)*, 2024.

Jingyu Zhang, Lilu Liu, Pingyu Xiang, Qin Fang, Xiuping Nie, Honghai Ma, Jian Hu, Rong Xiong, Yue Wang, and Haojian Lu. AI co-pilot bronchoscope robot. *Nature Communications*, 15(1):241, 2024.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

Technical Appendix

A Robotic Platform and Evaluation Protocol

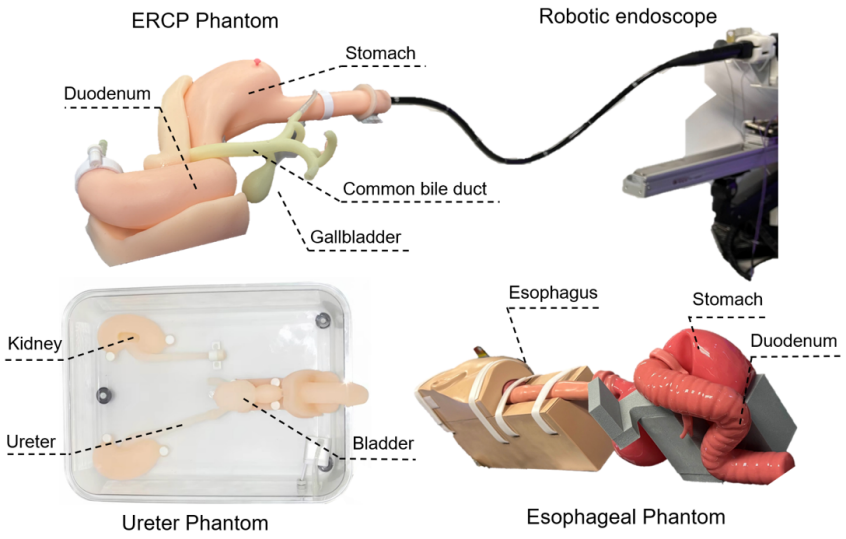


Figure 7: **Robotic platform and anatomical phantoms.** The three-DoF robotic endoscope provides two orthogonal distal-bending axes and one insertion/retraction axis. Closed-loop evaluations are conducted on physical phantoms of the ureter, esophagus, and biliary tree, corresponding to ureteroscopy, esophagoscopy, and ERCP, respectively.

Robotic platform. All policies are deployed on the same three-DoF robotic endoscope, which comprises two orthogonal distal-bending axes and one insertion/retraction axis, as illustrated in Fig. 7. Accordingly, the first two dimensions of each actuator command control distal-tip bending along the two orthogonal axes, whereas the third controls insertion or retraction.

Action execution. At control step t, the discrete action expert predicts an action chunk

$$\mathbf{a}_t = (\Delta\mathbf{q}_t, \ldots, \Delta\mathbf{q}_{t+T_a-1}), \tag{7}$$

where T_a=16 denotes the action horizon used during both training and closed-loop deployment. For $k = 0, \ldots, T_a - 1$, the actuation state is updated according to

$$\mathbf{q}_{t+k+1} = \mathbf{q}_{t+k} + \Delta\mathbf{q}_{t+k}. \tag{8}$$

Here $\mathbf{q}$ denotes the vector of three motor angles, and each command $\Delta\mathbf{q}_{t+k} \in \{-\theta, 0, +\theta\}^3$ applies a fixed angular increment θ of the corresponding motor, or leaves it stationary. Each command therefore specifies an independent discrete direction for each of the three actuation axes, consistent with the per-axis categorical action formulation in Sec. 3.3.

Anatomical phantoms. Closed-loop evaluations are performed on three physical anatomical phantoms representing the ureter, esophagus, and biliary tree. These phantoms instantiate the ureteroscopy, esophagoscopy, and ERCP procedures, respectively.

Navigation protocol. The main closed-loop evaluation follows a three-stage, long-horizon navigation protocol comprising entrance cannulation, lumen traversal, and target localization. The policy must first enter the anatomical lumen, traverse the deformable passage, and subsequently localize the instructed anatomical target. The zero-shot generalization experiments instead follow a two-stage protocol comprising lumen traversal and target localization; therefore, their success rates are not directly comparable to those of the three-stage main evaluation.

Language instructions. Each procedure is conditioned on a fixed language instruction specifying the navigation objective and instructed target. The instruction is provided to the policy at every control step, following the instruction-conditioned formulation in Eq. 2.

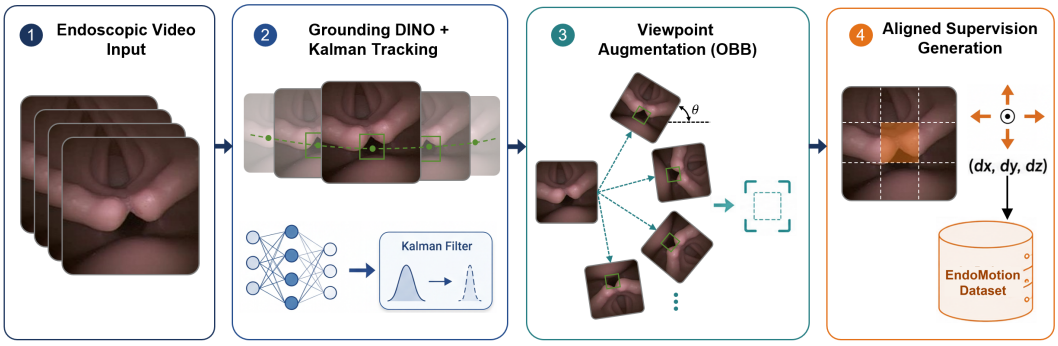


Figure 8: **EndoMotion dataset creation pipeline.** Recorded endoscopic videos are processed by a fine-tuned Grounding DINO and a Kalman filter to obtain temporally consistent target tracks. Each trajectory and its corresponding oriented bounding-box annotations are jointly transformed through viewpoint augmentation at eight camera-roll orientations. The transformed target locations are subsequently converted into aligned three-DoF discrete action labels and future target-region supervision.

Success criteria. A trial is considered successful if the endoscope reaches the instructed target and maintains it within the field of view until the rollout terminates. The rollout horizon is fixed to 675 control steps for all three procedures. Lumen loss, wall impaction, and failure to complete the navigation task within the prescribed horizon are categorized as failures, with the last case recorded as timeout. Representative failure modes are presented in Appendix G.

Evaluation protocol. Each policy is evaluated over 32 trials per procedure, yielding 96 trials for the main closed-loop comparison reported in Table 1. For the zero-shot generalization experiments in Table 3, each method is evaluated over 32 trials per condition.

For each procedure, the initial viewpoint is randomly sampled, and uniformly spaced camera-roll offsets spanning the full 360° range are applied. The evaluation viewpoints are deliberately unaligned with the eight roll angles used for viewpoint augmentation during training, ensuring that every trial evaluates generalization to an unseen viewpoint.

B Dataset Creation Pipeline

Target localization and tracking. For each recorded trajectory, we apply a fine-tuned Grounding DINO Liu et al. (2024) to every frame using the instructed anatomical target as the text prompt. A Kalman filter is then applied to the frame-wise detections to obtain temporally consistent oriented bounding-box (OBB) tracks. Each OBB is parameterized as (c_x, c_y, w, h, θ), where (c_x, c_y) denotes the target center, (w, h) denotes its spatial extent, and θ denotes its orientation. Kalman filtering suppresses transient detection noise and bridges short detector dropouts, producing temporally coherent target annotations whenever the target remains observable.

Discrete action-label generation. Each target track is converted into a per-frame three-DoF action label $\mathbf{a}_t \in \{-\theta, 0, +\theta\}^3$. For the two distal-bending axes, the target center is represented in the transformed frame and assigned to one of three spatial zones along each image axis. A fixed hysteresis margin is introduced around the zone boundaries to prevent small tracking fluctuations from causing rapid label transitions. The resulting horizontal and vertical zones are mapped to the corresponding discrete bending commands $(\Delta x, \Delta y)$.

The insertion/retraction label Δz is generated using target visibility, target position, and apparent target scale. Forward insertion is assigned only when the target is visible and simultaneously centered along both bending axes, so that the endoscope advances exclusively when it is already aligned with the lumen. Retraction is assigned when no target is tracked, which occurs on the small fraction of frames left unannotated by the tracker. All remaining frames receive a hold command: whenever the target is visible but off-center, the policy is labeled to re-center it through bending alone before resuming advancement, and the same command is issued in the terminal

case, where a sufficiently large and centered target indicates that the endoscope has reached the localization objective and should stabilize. Together, these rules produce an aligned three-axis label $(\Delta x, \Delta y, \Delta z)$ for every video frame.

Future-grounding supervision. The target tracks additionally provide supervision for future grounding. For each valid future frame, the target region defined by its tracked OBB is cropped from the image and resized to the prescribed target-region resolution. The cropped region is subsequently encoded by the frozen video VAE to obtain the target-region latent $\mathbf{z}_n^{\mathrm{tar}}$ used by the Grounding DiT.

Frames for which no valid target track is available are excluded from the future-grounding objective through the validity indicator u_n. These frames remain available for the action and video objectives, ensuring that temporary target absence does not remove the corresponding transition from policy training.

Viewpoint augmentation. To improve robustness to camera-roll variations, each trajectory is transformed at eight roll orientations,

$$0^\circ, 45^\circ, \ldots, 315^\circ,$$

with the original orientation included. Each frame is rotated about its optical center on an expanded canvas, followed by centered cropping and resizing to the target resolution. The identical geometric transformation is applied analytically to the corresponding OBB annotations.

Action labels and future target-region supervision are regenerated after transformation. Consequently, each augmented orientation is treated as an independently aligned training sequence rather than as an image-only copy of the original trajectory. This procedure expands EndoMotion from 515 trajectories to $4{,}120$ viewpoint-augmented sequences.

C EndoMotion Dataset

Table 5 reports the per-procedure statistics of EndoMotion. The real dataset contains 515 teleoperated trajectories and $979{,}564$ frames across ureteroscopy, esophagoscopy, and ERCP. The three procedures contain comparable numbers of trajectories but differ substantially in trajectory duration. Ureteroscopy and ERCP contribute $406{,}612$ and $407{,}617$ frames, respectively, whereas esophagoscopy contributes $165{,}335$ frames.

Table 5: **EndoMotion dataset statistics.** The annotation pipeline produces a dense three-DoF action label for every frame and an OBB target whenever a valid target track is available. Viewpoint augmentation transforms each trajectory and its annotations at eight camera-roll orientations separated by 45°.

Procedure	Training sequences	Action labels	OBB targets
Ureteroscopy	151	406,612	404,711
Esophagoscopy	163	165,335	164,524
ERCP	201	407,617	398,801
Total (real)	515	979,564	968,036
+ viewpoint aug.	**4,120**	**7,836,512**	**7,651,320**

Every frame is associated with a three-DoF discrete action label. Valid OBB annotations, and therefore future target-region supervision, are available for $968{,}036$ frames, corresponding to 98.8% of the real dataset. Frames without valid target annotations remain part of the action and video objectives but are masked out of the future-grounding objective.

After viewpoint augmentation at eight camera-roll orientations, the dataset contains $4{,}120$ sequences and $7{,}836{,}512$ action-labeled frames. The number of action labels scales exactly by a factor of eight because every transformed frame is retained. By contrast, the number of valid OBB targets increases to $7{,}651{,}320$, slightly below the theoretical eight-fold expansion. This difference occurs because targets located near the image boundary may move outside the valid cropped field of view under some camera-roll transformations.

Esophagoscopy constitutes the smallest portion of EndoMotion in terms of frame count and also yields the lowest closed-loop success rate in the main evaluation. Although this correlation does not establish causality, it suggests that increasing the diversity and duration of esophagoscopy trajectories may further improve procedure-specific performance.

D Model and Training Configuration

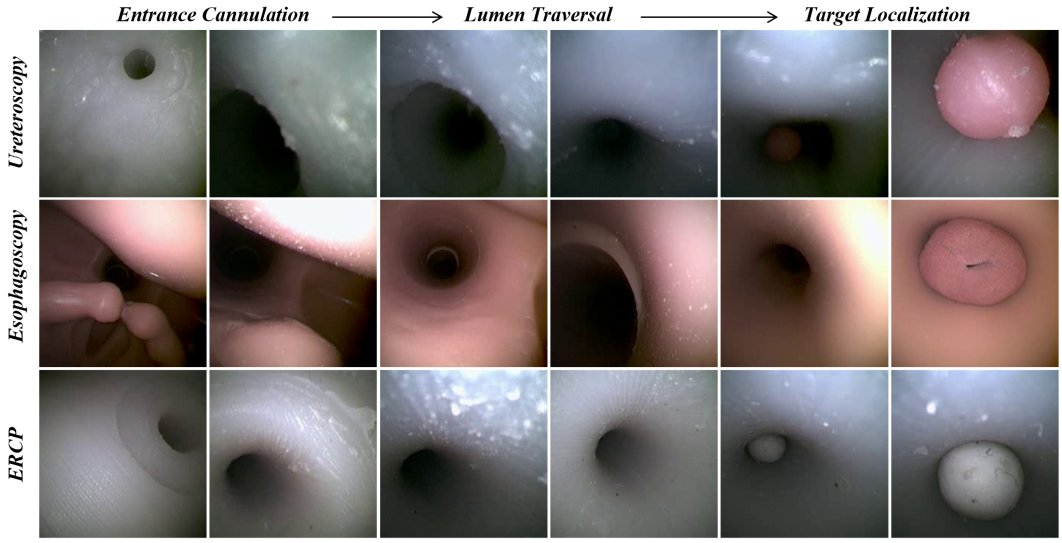


Figure 9: **Additional closed-loop demonstrations under unseen viewpoints.** Representative EndoWAM rollouts across ureteroscopy, esophagoscopy, and ERCP. Each row illustrates the three-stage navigation protocol consisting of entrance cannulation, lumen traversal, and target localization. The evaluation viewpoints are induced by camera-roll offsets that are unaligned with the eight orientations used for viewpoint augmentation during training.

Backbone adaptation. EndoWAM is initialized from Cosmos-Predict2.5-2B Agarwal et al. (2025), which comprises a pretrained video DiT, a text encoder, and a video VAE. We apply LoRA Hu et al. (2022) with rank 32, scaling factor α=64, and dropout 0.02 to the query, key, value, output, and feed-forward projections of the Video DiT. The text encoder and video VAE remain frozen throughout training.

Following the formulation in Sec. 3.2, the shared predictive representation H is extracted from the block-17 activation after a single denoising pass and flattened into a token sequence. The same representation jointly conditions the discrete action expert and the Grounding DiT.

Discrete action expert. The discrete action expert contains a 6-layer transformer with hidden dimension 768 and 12 attention heads. It maintains one learnable action query for each step of the action horizon. The resulting T_a=16 query tokens are processed using bidirectional self-attention and cross-attention to the shared representation H.

Bidirectional self-attention allows each query to incorporate the temporal context of the complete action chunk, while cross-attention injects predictive visual information from the Video DiT. An MLP classification head subsequently predicts per-axis categorical logits over $\{-\theta, 0, +\theta\}$ for each of the three actuation axes. All action classes are assigned equal weights in the reported experiments.

Grounding DiT. The Grounding DiT comprises three AdaLN-Zero transformer blocks with hidden dimension 1024 and 16 attention heads. It reconstructs the VAE latents of future target regions conditioned on a projected and resampled representation of H. Target-region latents are denoised under a cosine noise schedule with K=1000 diffusion steps. During each optimization step, the grounding objective is averaged over R=4 independently sampled noise levels and noise realizations.

Optimization. We optimize EndoWAM with AdamW using a cosine learning-rate schedule and linear warmup. The LoRA-adapted Video DiT uses a learning rate of 5×10^{-6}, whereas the discrete action expert and Grounding DiT use 1×10^{-4}. Training employs bf16 mixed precision, gradient accumulation, and DeepSpeed ZeRO Stage 2 across four GPUs.

All reported models, including EndoWAM, the controlled grounding variants, the ablation models, and the external baselines, are trained for $100{,}000$ optimization steps. The loss weights in the EndoWAM joint objective are set to

$$\lambda_v = \lambda_r = 1,$$

where λ_v and λ_r weight the video and future-grounding objectives, respectively. EndoWAM is trained on the complete viewpoint-augmented EndoMotion dataset using an input resolution of 224×224 per frame.

Controlled grounding variants. The controlled variants introduced in Sec. 3.4 share the same Cosmos-Predict2.5-2B backbone, LoRA adaptation, discrete action expert, viewpoint-augmented EndoMotion dataset, action horizon, and optimization configuration as EndoWAM. They differ only in how task-relevant target information is incorporated into the policy. Whenever an auxiliary grounding objective is used, its loss weight is set to $\lambda_r{=}1$, consistent with the full model.

EndoWAM-CG. EndoWAM-CG replaces the Grounding DiT with a regression head that predicts the oriented bounding box of the target in each future frame directly from the shared predictive representation H. Each future target is represented as

$$\mathbf{b}_n = (c_x, c_y, w, h, \sin 2\theta, \cos 2\theta),$$

together with a target-presence prediction. The geometric and orientation parameters are optimized using smooth-L_1 loss on frames with valid target annotations, while binary cross-entropy supervises target presence over all future frames. The regression head is used only during training and removed at deployment. This variant evaluates whether low-dimensional future target geometry provides sufficient grounding supervision for the shared predictive representation.

EndoWAM-FG-Gen. EndoWAM-FG-Gen incorporates future target regions into the generation target of the Video DiT. Specifically, the world model jointly models the latent representations of future observations and their corresponding target regions through its generative pathway. Unlike the full EndoWAM model, it does not employ a separate Grounding DiT to reconstruct future target-region latents. Instead, grounding supervision is introduced directly through the world model's generative objective. At deployment, future observations and target regions are not decoded, and the discrete action expert generates the action chunk from the shared predictive representation obtained after a single denoising pass.

Ablation configurations. For the *w/o Grounding DiT* ablation, we remove the Grounding DiT and the corresponding future-grounding objective $\mathcal{L}_{\text{ground}}$, while retaining the Video DiT, discrete action expert, video objective, action objective, and viewpoint-augmented training data. For the *w/o viewpoint augmentation* ablation, we retain the complete EndoWAM architecture and joint objective but train only on the original 515 trajectories, without applying the eight camera-roll transformations. All remaining model, optimization, and evaluation settings are identical to those of the full model.

E Baseline Training Configurations

All external baselines receive the same language instructions and endoscopic observations as EndoWAM and are trained on the same viewpoint-augmented EndoMotion dataset for $100{,}000$ optimization steps. The action horizon and observation resolution are matched whenever supported by the corresponding architecture. These settings control for differences in training data and evaluation protocol, allowing the comparison to primarily reflect policy architecture and action-generation mechanism.

Diffusion Policy Chi et al. (2023). Diffusion Policy serves as a non-VLA visuomotor baseline. It uses a CNN-based visual encoder and a U-Net diffusion head to predict a 16-step action chunk. It does not incorporate a pretrained vision-language backbone, predictive video dynamics, or future-grounding supervision.

$\pi_{0.5}$ **Physical Intelligence et al. (2025).** $\pi_{0.5}$ is a large-scale pretrained vision-language-action policy. We fine-tune the released pretrained model on EndoMotion following its official training configuration, while matching the action horizon and observation resolution to the EndoWAM evaluation setting.

GR00T-N1.7 Bjorck et al. (2025). GR00T-N1.7 is a pretrained generalist vision-language-action policy. The model is adapted to the three-DoF endoscopic action space and fine-tuned on the same EndoMotion training mixture, with the action horizon and optimization budget matched to those used by EndoWAM.

Qwen3DiT and Qwen3AE. Qwen3DiT and Qwen3AE share the same Qwen3-VL-2B Bai et al. (2025) vision-language backbone and LoRA adaptation configuration. They differ only in their action-generation heads.

Qwen3DiT uses a diffusion-transformer action head to generate continuous action trajectories through flow matching. Qwen3AE instead adopts the same discretized action-decoding architecture used by EndoWAM, predicting per-axis categorical logits over $\{-\theta, 0, +\theta\}$. Neither model receives future-grounding supervision or a predictive video-dynamics objective. Their comparison therefore isolates the effect of diffusion-based and discretized action parameterizations under a shared vision-language backbone.

F Additional Demonstrations

Figure 9 provides additional closed-loop rollouts across the three procedures. During entrance cannulation, the policy aligns the endoscope with the target lumen opening and advances into the anatomical passage. During lumen traversal, it maintains a valid intraluminal view while correcting for changes in lumen position and camera orientation. During target localization, the policy approaches the instructed anatomical target and stabilizes it within the field of view.

Together with the quantitative results in Table 1, these trajectories demonstrate that EndoWAM can execute the complete long-horizon navigation sequence under camera viewpoints not observed during training. The rollouts additionally contain transient visual degradation, including darkness, partial occlusion, and specular highlights. Qualitatively, the policy can preserve or re-establish target-directed navigation in several such cases. The Grounding DiT ablation in Sec. 4.4 provides quantitative evidence that future target-region supervision is important for this behavior.

As shown in Fig. 11, EndoWAM consistently progresses from lumen traversal to target localization across all six zero-shot conditions. Under cross-environment shifts, the policy preserves target-directed navigation despite unseen variations in lumen reflectance, geometry, and texture. Under cross-target shifts, it successfully approaches and stabilizes on targets with novel shapes, sizes, and colors. Together with the results in Table 3, these examples demonstrate that EndoWAM generalizes beyond the specific lumen appearances and target configurations observed in EndoMotion.

G Failure-Case Analysis

Figure 10 visualizes three representative behaviors observed in unsuccessful trials. These behaviors provide a more fine-grained interpretation of the protocol-level failure categories defined in Sec. 4.1.

Target loss. In the first example, the instructed target initially appears within the field of view but progressively moves outside the observable region as the endoscope deviates from the target-directed path. The policy subsequently fails to recover the target before the rollout terminates. When the corresponding lumen is also lost, this behavior is recorded as lumen loss under the evaluation protocol.

Step-limit exceeded. In the second example, the policy successfully traverses the lumen and approaches the target but remains near an intermediate configuration without producing sufficient corrective motion. This prolonged action stagnation prevents stable target localization before the maximum rollout length of 675 control steps and is therefore recorded as timeout.

Wrong-target acquisition. In the third example, the intended target is lost during navigation and the policy subsequently approaches a different lumen opening with a visually similar appearance.

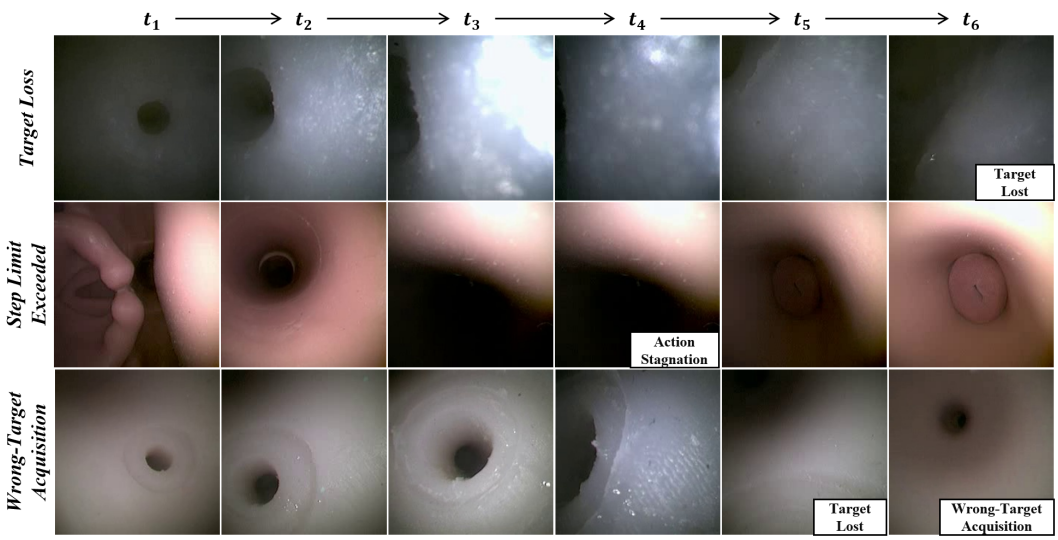


Figure 10: **Representative failure behaviors.** Top: the policy loses the intended target after a progressive viewpoint deviation and fails to recover it. Middle: the policy approaches the target but exhibits prolonged action stagnation, causing the rollout to exceed the maximum step limit. Bottom: after losing the intended target, the policy converges to an incorrect lumen opening, resulting in wrong-target acquisition.

Because the instructed target is not reached and stabilized before the rollout terminates, this behavior is also scored as an unsuccessful trial. Under the protocol-level taxonomy, it is recorded as timeout unless it is accompanied by lumen loss or wall impaction.

These examples indicate that failures can arise not only from complete loss of the intraluminal view, but also from insufficient corrective motion and ambiguity between visually similar anatomical structures. Future work could address these cases through explicit uncertainty estimation, recovery-state supervision, and longer temporal context.

H Limitations and Discussion

EndoWAM demonstrates that predictive visual dynamics and future grounding can enable accurate, generalizable, and real-time closed-loop navigation across multiple endoscopic procedures. While the current results establish the effectiveness of this formulation, they also highlight several directions for extending the framework toward broader and more clinically realistic deployment.

Dataset scale and procedural diversity. EndoMotion contains 515 real trajectories and 0.98M real frames, which are expanded to $4{,}120$ sequences and 7.84M frames through viewpoint augmentation. Despite this moderate data scale, EndoWAM achieves strong closed-loop performance and zero-shot generalization, suggesting that the pretrained video prior and future-grounding objective provide effective inductive biases for data-limited endoscopic control. Extending EndoMotion to additional anatomical structures, endoscopic procedures, device platforms, and operators would further increase the diversity of visual observations and anatomical dynamics represented during training. Incorporating multi-device and multi-center data may additionally improve robustness across acquisition conditions and hardware configurations.

Scope of validation. The present study focuses on closed-loop evaluation using physical anatomical phantoms. This setting provides repeatable experimental conditions, physical endoscope–environment interactions, and controlled variations in viewpoint, anatomy, and visual appearance, enabling systematic comparison across methods. Evaluation in ex vivo and in vivo settings would introduce additional physiological variability, including bleeding, secretions, peristalsis, patient-specific anatomy, and more complex tissue responses to contact. Extending EndoWAM to these

settings represents an important next step for assessing the robustness, safety, and translational potential of predictive and target-grounded endoscopic control.

Failure recovery. Recovery behavior currently emerges implicitly from the learned predictive representation rather than being optimized through an explicit recovery objective. As illustrated in Fig. 10, the policy may still lose the intended target, stagnate near an intermediate configuration, or converge to a visually similar but incorrect target. Future grounding provides a promising foundation for addressing these failure modes by modeling the short-horizon evolution of task-relevant regions. Future work could incorporate recovery demonstrations, longer-term temporal memory, uncertainty-aware action selection, and auxiliary supervision for target identity and target visibility. Safety-aware recovery mechanisms could further trigger controlled retraction or exploratory viewpoint adjustment when predictive confidence is low.

Overall, the modular structure of EndoWAM allows these extensions to be incorporated without altering its core predictive and target-grounded formulation. Scaling the training data, expanding validation to more realistic environments, and explicitly modeling uncertainty and recovery could further strengthen the applicability of EndoWAM to diverse long-horizon endoscopic navigation tasks.

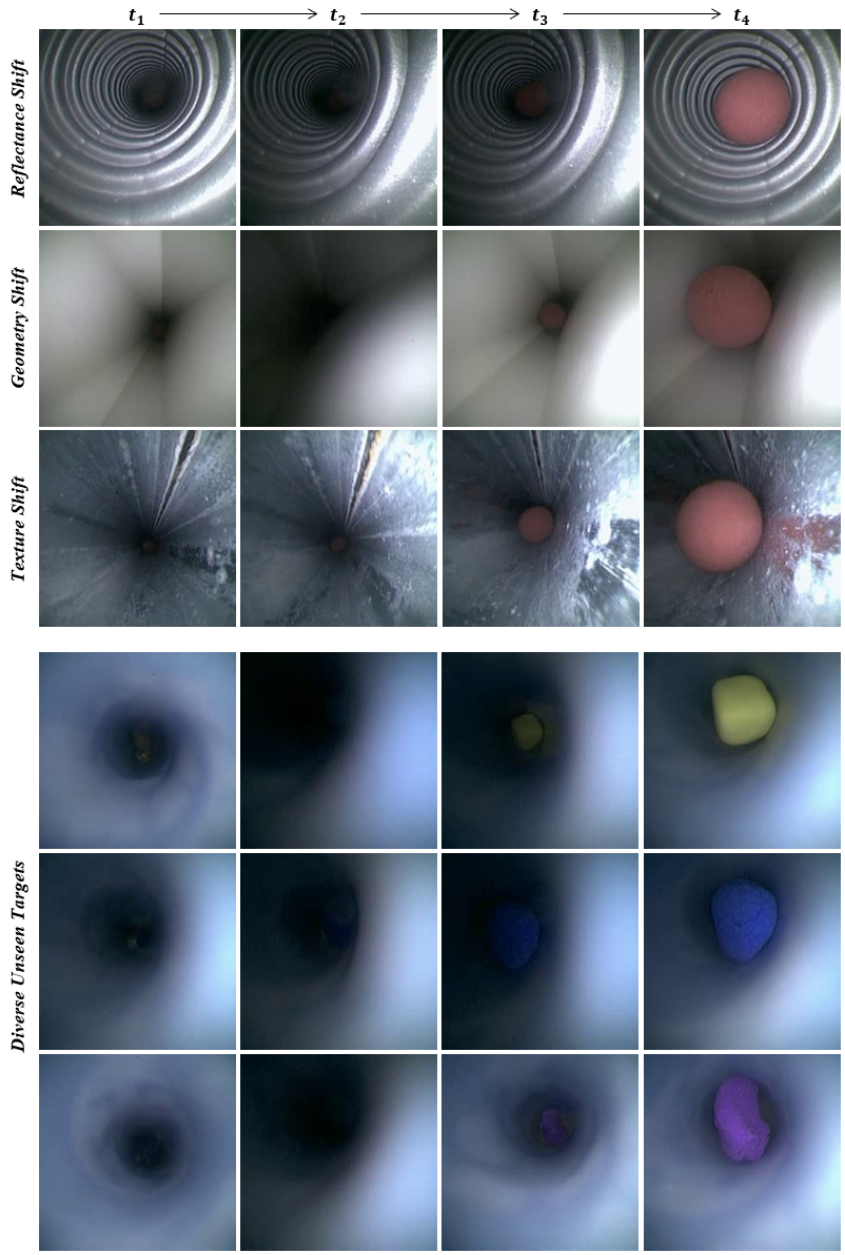


Figure 11: **Zero-shot generalization under cross-environment and cross-target shifts.** Representative successful EndoWAM rollouts under conditions unseen during training. The top three rows show cross-environment shifts in lumen reflectance, geometry, and texture, while the bottom three rows show cross-target shifts involving diverse unseen target appearances. Each sequence illustrates progression from lumen traversal to target localization.