HIERASURG: WORLD MODELS FOR SURGICAL DATA SCIENCE

Diego Biagini [1,2] Azade Farshad[1,2] Nassir Navab[1,2]

[1]Technical University of Munich [2]Munich Center for Machine Learning (MCML)

Code @

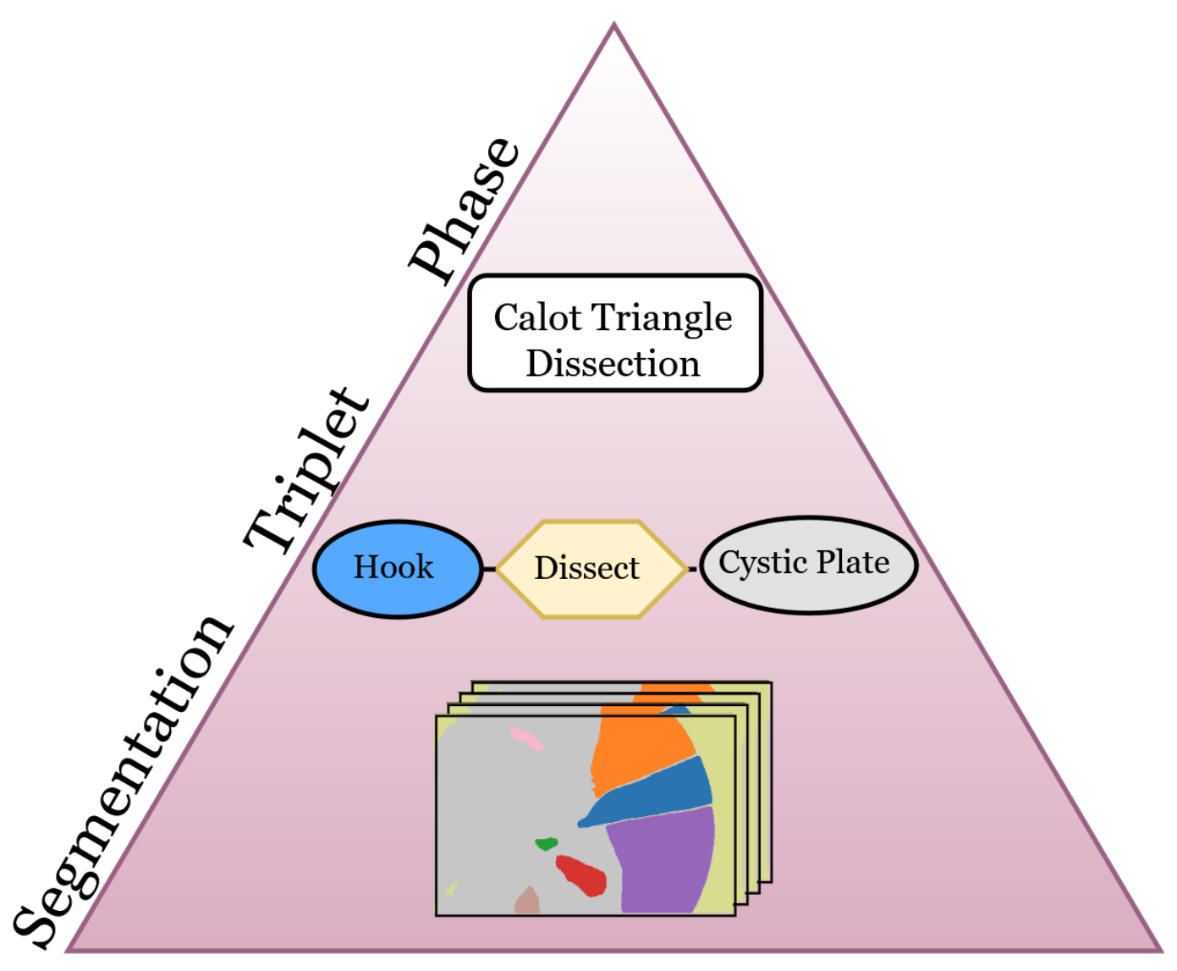
MICCAI2025

ICVSS2025
Sicily ~ 6-12 July 2025
International Computer Vision Summer School

Abstract

Synthetic data generation in data-scarce settings, like internal surgery, greatly benefits from the implicit approach enabled by learning-based methods. We first present **HieraSurg**, a video generation framework consisting of two specialized diffusion models that can generate realistic surgical videos. While visual quality is great, faithfully replicating physics, interactions and domain peculiarities requires a more profound understanding; which we believe ought to be fulfilled by simulation.

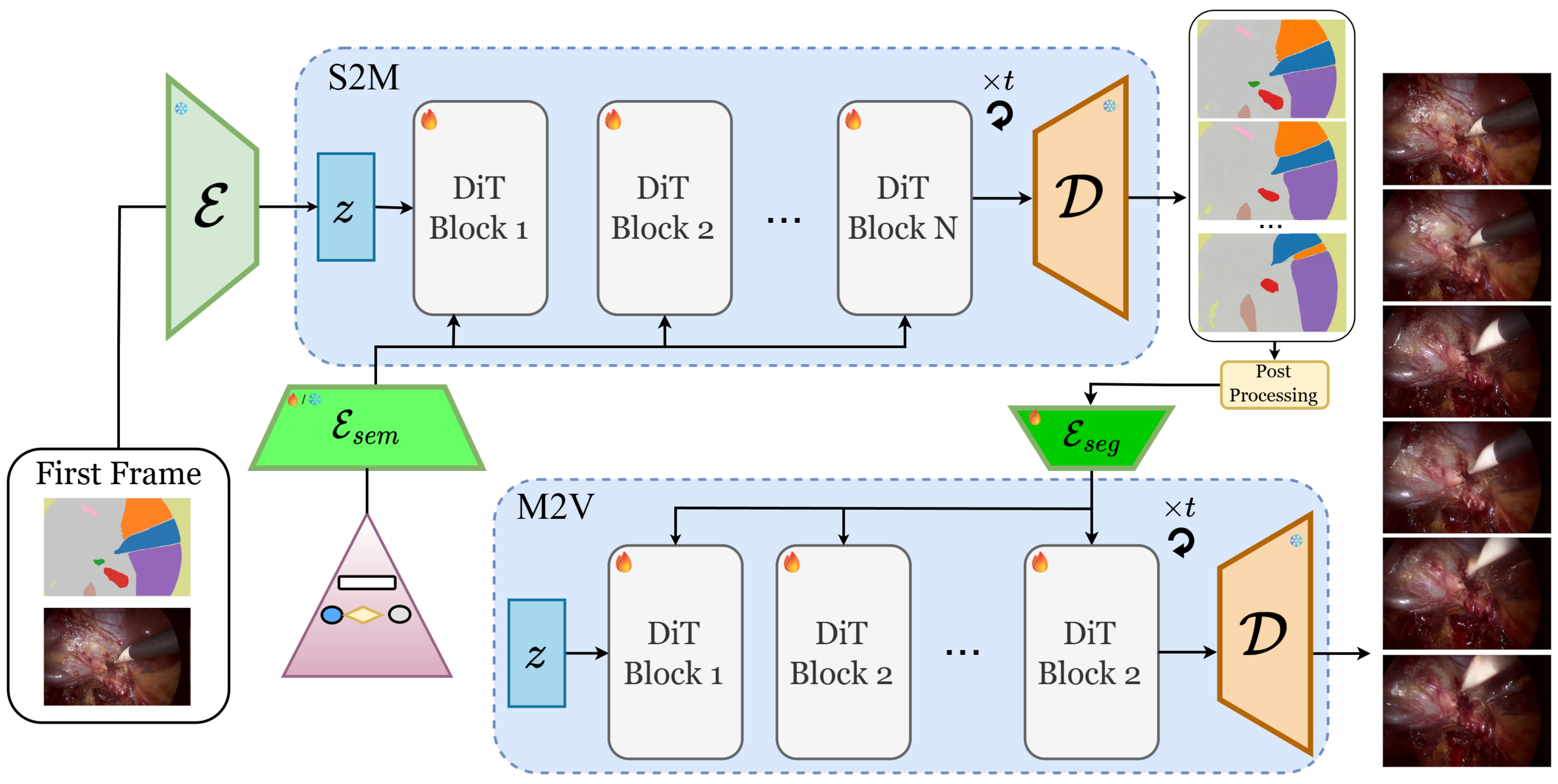
Semantics in Surgical Scenes



Semantic information is modeled hierarchically, defining different levels of abstractions:

- **Surgical Phase**: long term (minutes) scene descriptor
- **Action Triplet**: short-term (5-10 seconds) interaction between tools and body parts
- **Segmentation Maps**: frame-level information

HieraSurg Architecture

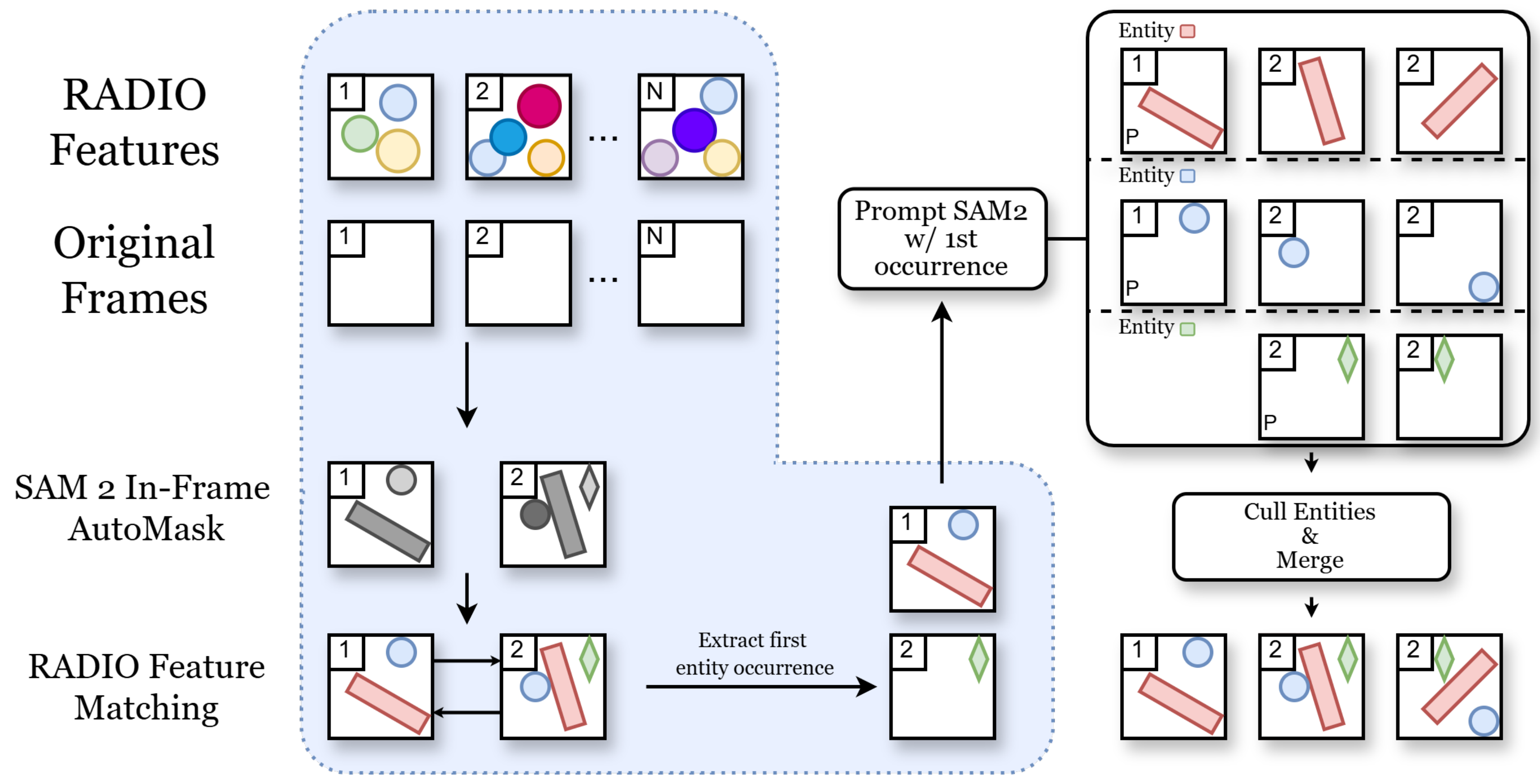


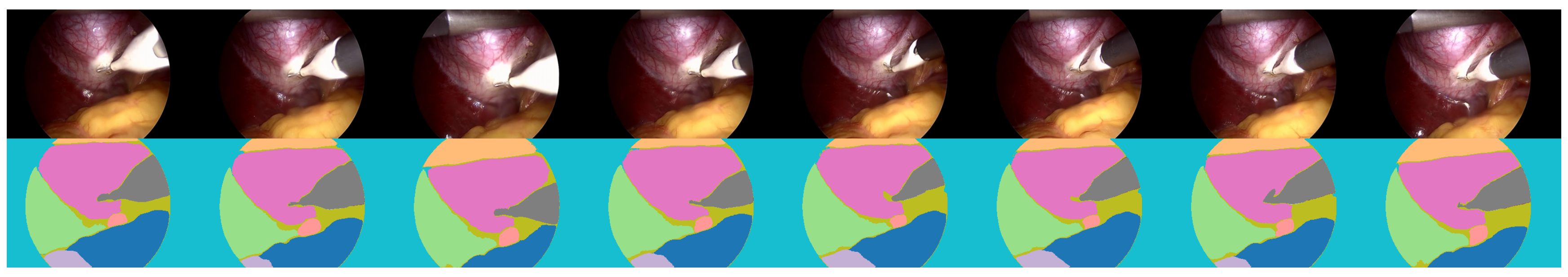
HieraSurg is a pair of diffusion models consisting of:

- **HieraSurg-S2M** (Semantic to Map): generates a plausible evolution of the surgical scene in the domain of segmentation maps from phase and triplet information
- **HieraSurg-M2V** (Map to Video): completes the task by bringing segmentation maps to video-space.

Extraction of Semantic Maps from surgical videos

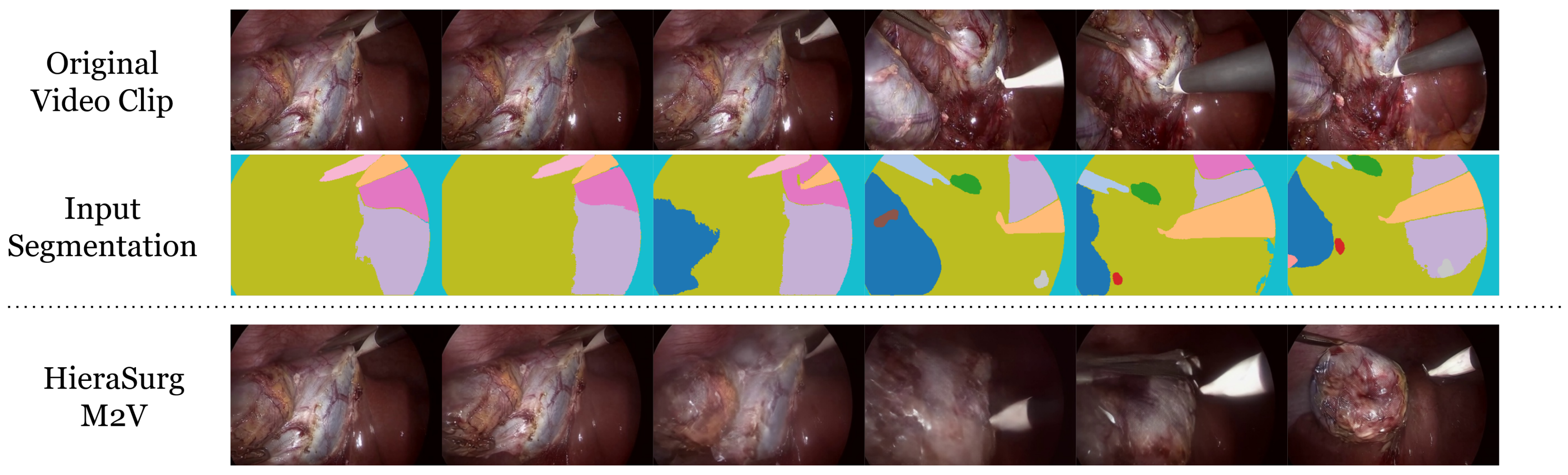
To train both components, an automated labelling pipeline based on SAM2 and RADIO feature matching is used, allowing the extraction of panoptic segmentation maps from unlabeled surgical videos.



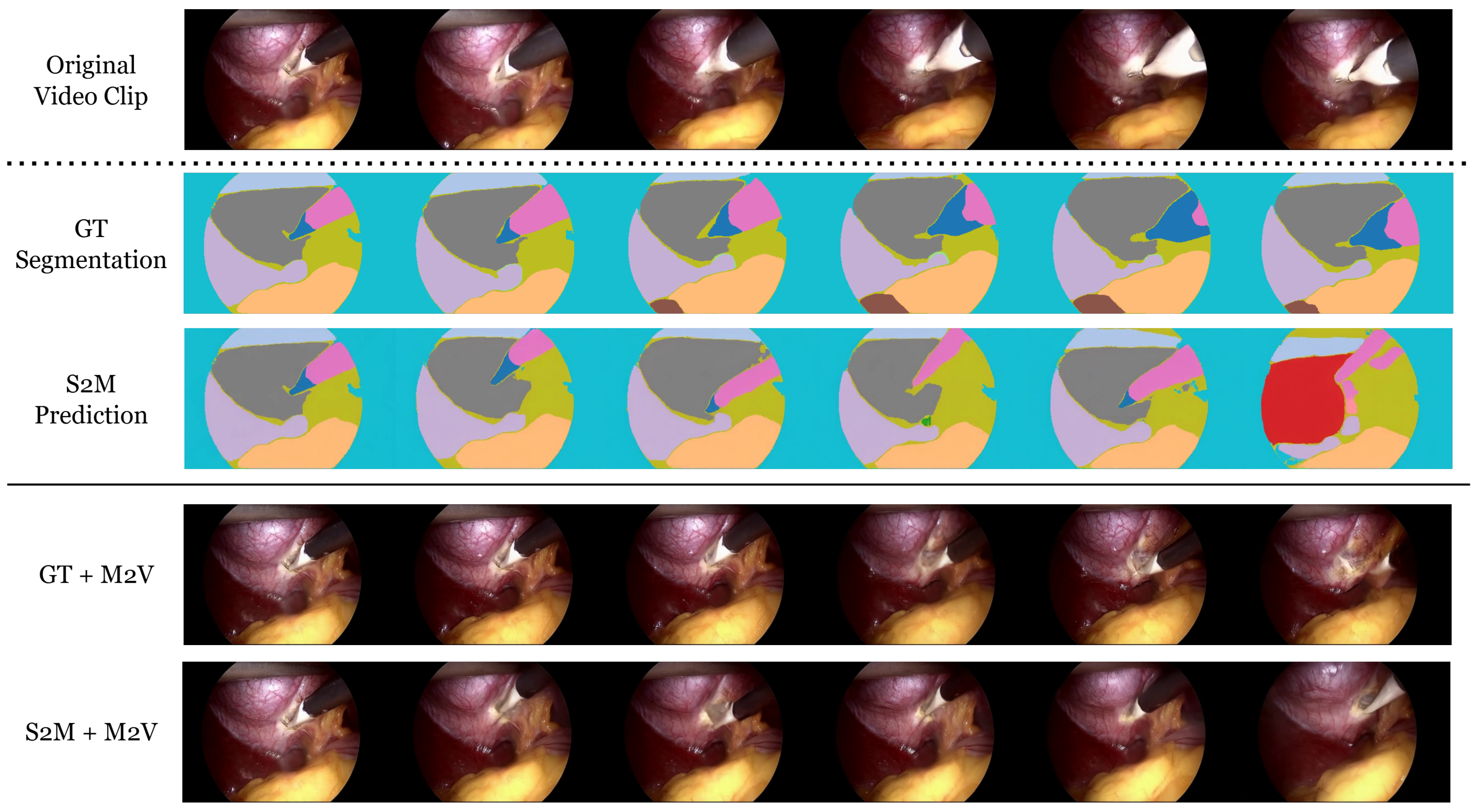


Experimental Results

We obtain HieraSurg-M2V by finetuning CogVideoX-2B[1] on the Cholec80[2] dataset, both at 1fps (16 frames) and 8fps (48 frames), after running the semantic map extraction procedure on 65 videos.



We separately train HieraSurg-S2M on predicting 16 future segmentation maps at 1fps. We train on the subset CholecT45[3], which provides phase and triplet information.
Finally we evaluate the full pipeline, i.e. when we let S2M generate future predictions of the segmentation maps.



Marrying Simulation and Surgical Videos

Learning World Models through Latent Actions

References

[1] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng, D. Yin, Yuxuan.Zhang, W. Wang, Y. Cheng, B. Xu, X. Gu, Y. Dong, and J. Tang, "CogVideoX: Text-to-video diffusion models with an expert transformer,"

[2] T. Andru, S. Sherif, M. Didier, M. Jacques, D. M. Michel, and P. Nicolas, "Endonet: A deep architecture for recognition tasks on laparoscopic videos," *IEEE Transactions on Medical Imaging*, vol. 36, 02 2016.

[3] C. I. Nwoye, T. Yu, C. Gonzalez, B. Seeliger, P. Mascagni, D. Mutter, J. Marescaux, and N. Padoy, "Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos," *Medical Image Analysis*, vol. 78, p. 102433, 2022.