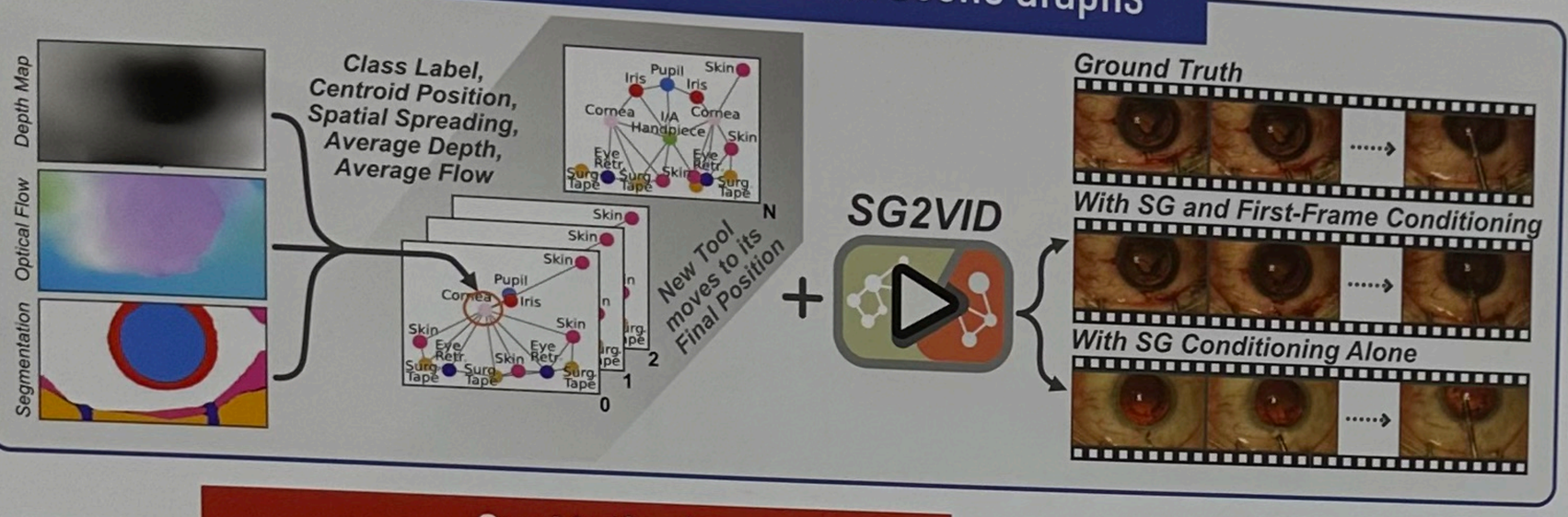
SG2VID: Scene Graphs Enable Fine-Grained Control for Video Synthesis

Ssharvien Kumar Sivakumar – Yannik Frisch – Ghazal Ghazaei – Anirban Mukhopadhyay

MICCAI 2025
28th International Conference on Medical Image Computing and Computer Assisted Intervention
23-27 September 2025
Daejeon Convention Center
Daejeon REPUBLIC OF KOREA

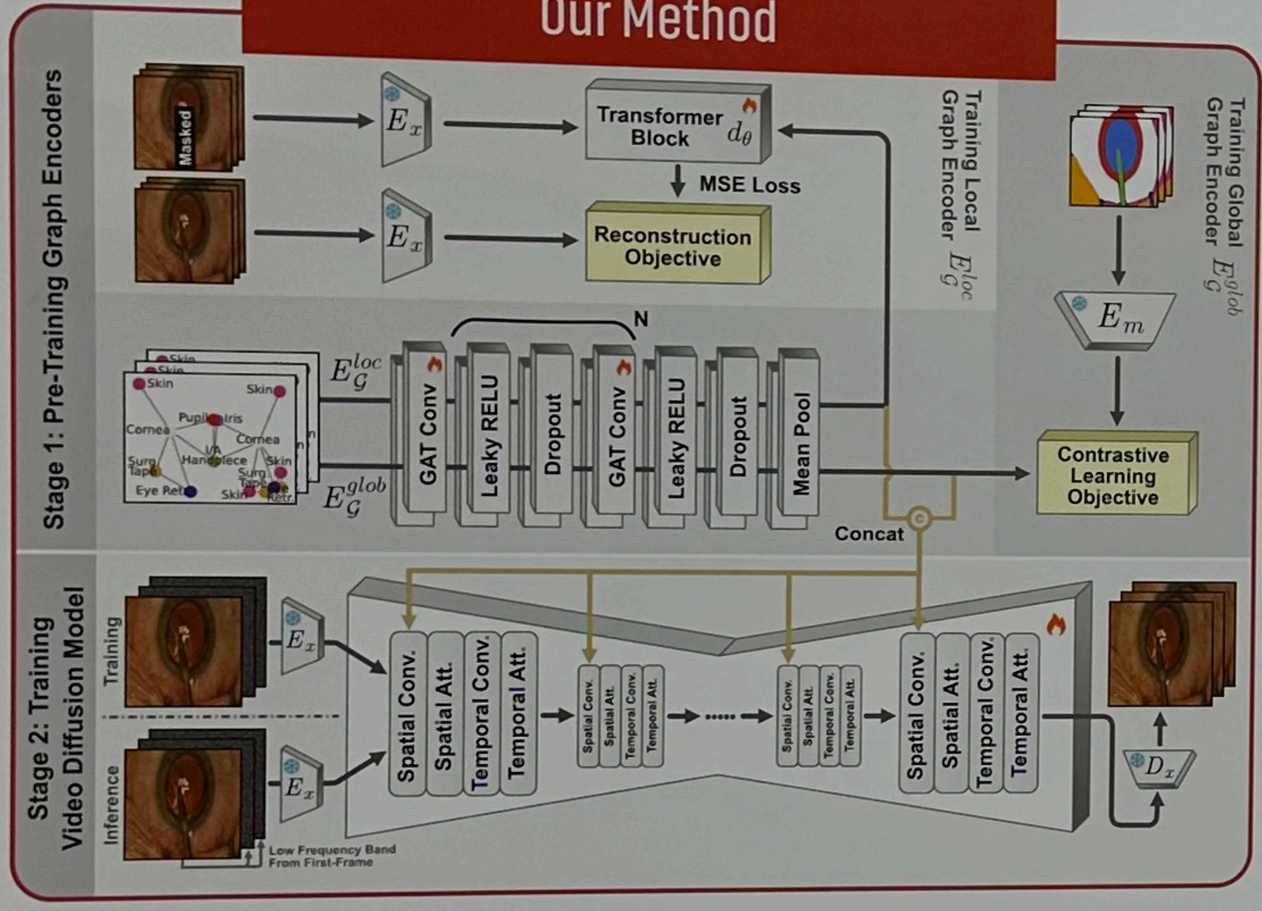
Video Synthesis from Scene Graphs



Problem Statement

- Render-based simulators lack photorealism and rely on hard-coded behaviours
- Generative model-based alternatives struggle to provide conditioning mechanisms that balance **precise** video synthesis with **fine-grained, human-interactable** control over scenes

Our Method



Controllability and Generalizability

Synthesise Rare Irregularities

Easily interact & edit Scene Graphs to simulate sudden pupil contraction [1]

Zero-Shot Style Transfer

Demonstrates generalization by generating videos from unseen image domains [2]

Quantitative Assessment

- SG2VID outperforms baselines in visual fidelity while more faithfully adhering to the conditioning signal.

Method	Cataract-1k FVD↓	Cataract-1k FID↓	Cataract-1k LPIPS↑	CATARACTS FVD↓	CATARACTS FID↓	CATARACTS LPIPS↑	Cholec80 FVD↓	Cholec80 FID↓	Cholec80 LPIPS↑
StyleGAN-V	442.6	118.2	0.286	618.7	94.8	0.382	1544.1	200.5	0.378
Endora	265.9	30.2	0.377	649.5	45.9	0.454	533.8	47.0	0.525
MedSora	901.8	137.6	0.324	952.1	112.1	0.403	1297.5	153.6	0.406
*LVDM	1656.6	186.0	0.534	1178.9	113.7	0.559	1507.0	110.8	0.668
MOFA	722.2	89.9	0.361	713.4	88.3	0.460	651.1	72.4	0.506
SG2VID (Ours)	77.0	15.5	0.397	523.8	40.9	0.444	457.3	16.4	0.532
SG2VID-XIMG (Ours)	278.4	33.3	0.409	535.7	39.8	0.465	560.1	25.1	0.533

Method	Cataract-1k BB IoU↑	Cataract-1k F1↑	CATARACTS BB IoU↑	CATARACTS F1↑	Cholec80 BB IoU↑	Cholec80 F1↑
*LVDM	0.193	0.164	0.225	0.149	0.42	0.28
MOFA	0.456	0.449	0.425	0.275	0.599	0.45
SG2VID (Ours)	0.624	0.634	0.493	0.379	0.623	0.476
SG2VID-XIMG (Ours)	0.440	0.460	0.389	0.318	0.530	0.389
Mask R-CNN on Annotated Subset	0.731	0.745	0.636	0.585	0.887	0.967

Qualitative Samples

Conclusion

- Scene Graphs offer a succinct representation to mediate precise synthesis and fine-grained human control
- Potential applications in **privacy-preserving data sharing**, **unpaired video domain transfer** and realistic surgical simulation

[1] Ghamsarian, Negin, et al. "Cataract-1k dataset for deep-learning-assisted analysis of cataract surgery videos." *Scientific data* 11.1 (2024)
[2] Hu, Ming, et al. "Ophnet: A large-scale video benchmark for ophthalmic surgical workflow understanding." *ECCV* (2024)

ssharvien_kumar.sivakumar@tu-darmstadt.de

software campus
Bundesministerium für Bildung und Forschung

LinkedIn | Project Page | Code