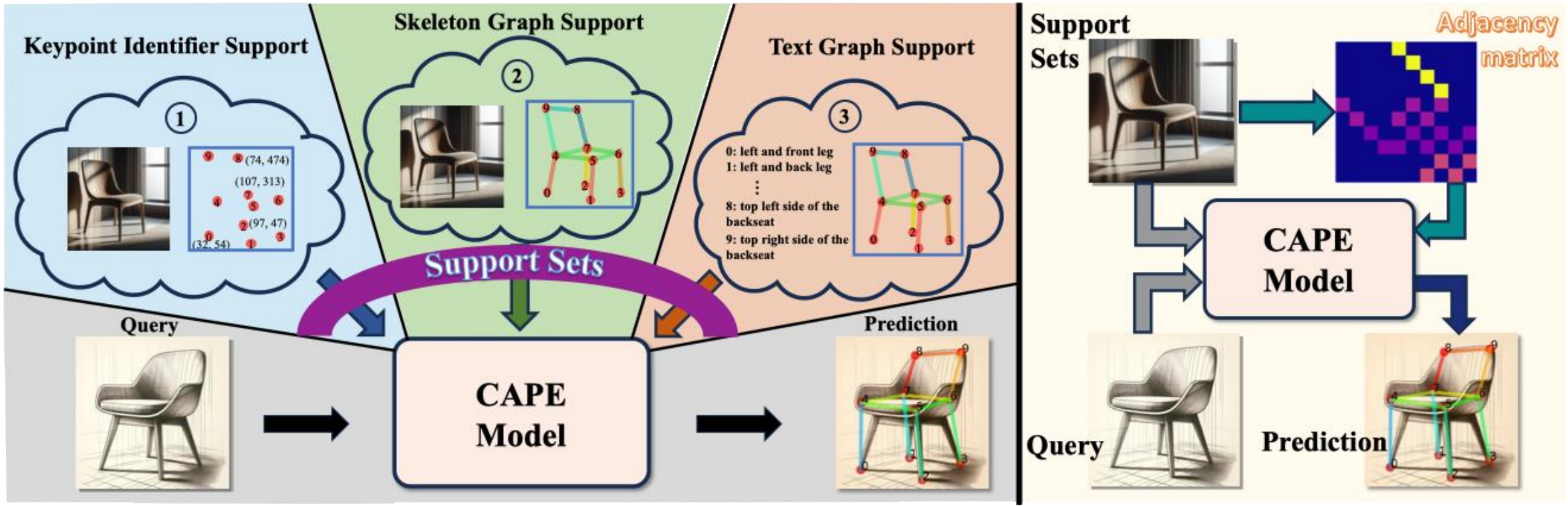
GenCape: Structure-Inductive Generative Modeling for Category-Agnostic Pose Estimation

Jiyong Rao, Yu Wang, Shengjie Zhao

School of Computer Science and Technology, Tongji University

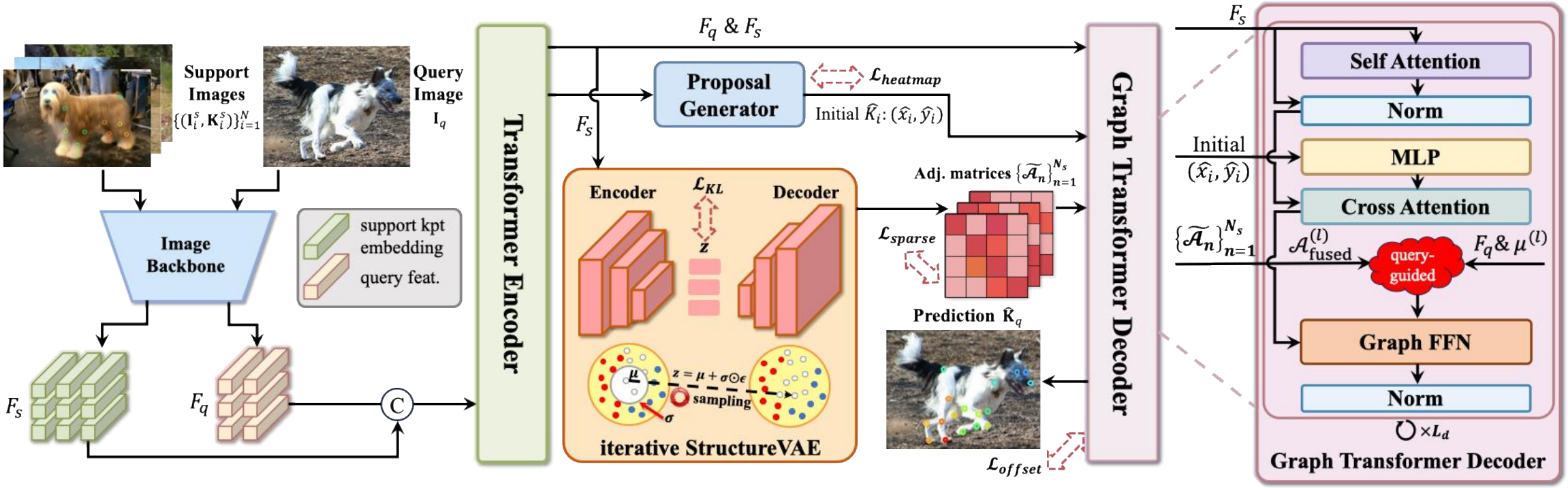
ICLR 2026

Limitation of Category-Agnostic Pose Estimation Models



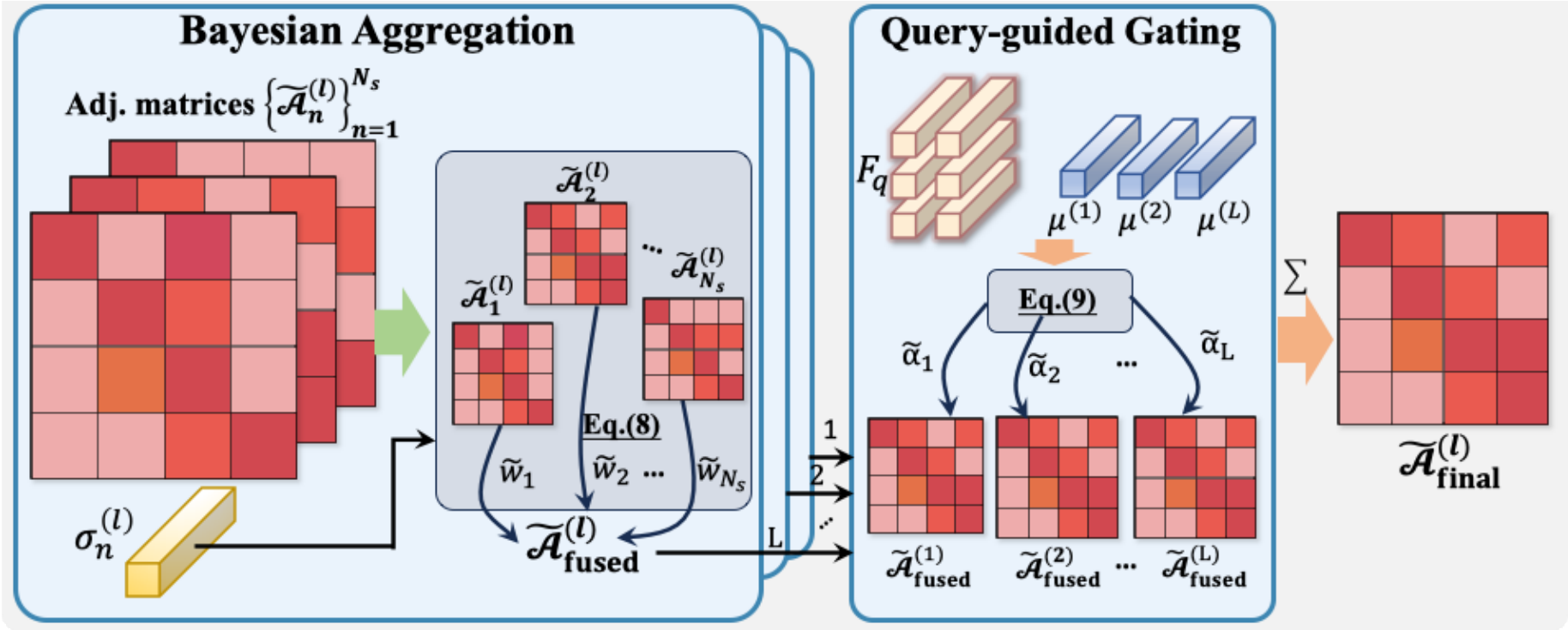
- Previous CAPE methods either treat keypoints as isolated entities or rely on **manually** defined skeleton priors or auxiliary textual descriptions.
- We recognize that **instance-specific structural dependencies** among keypoints should be inferred directly from support images in a **data-driven manner**, rather than hand-crafted priors.

General Framework



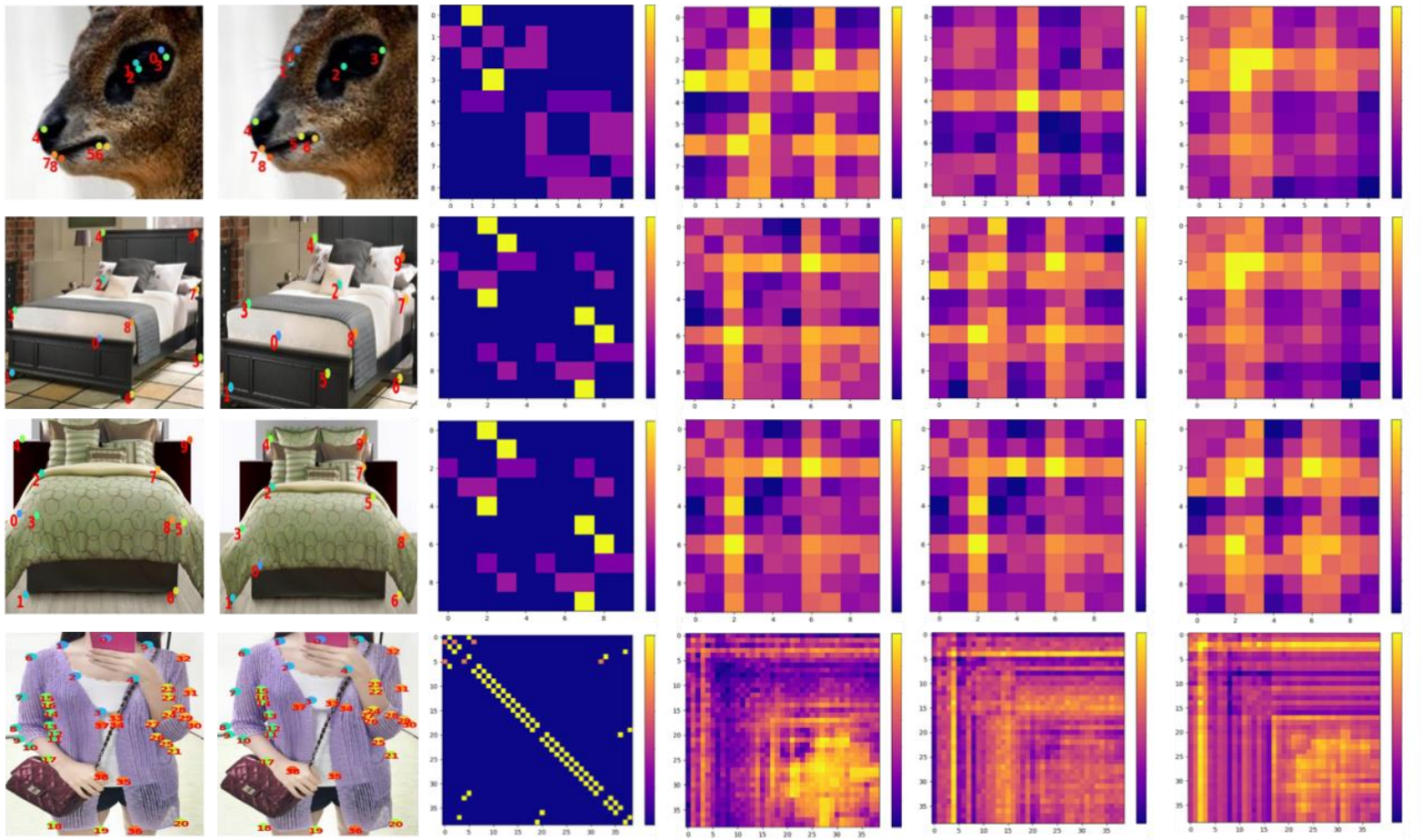
- Iterative Structure-aware Variational Autoencoder to infer latent, instance-specific skeletons solely from image-based support sets.
- Probabilistic encoder that parameterizes a latent graph distribution.
- Decoder that constructs the adjacency matrix from the latent space.

Compositional Graph Transfer

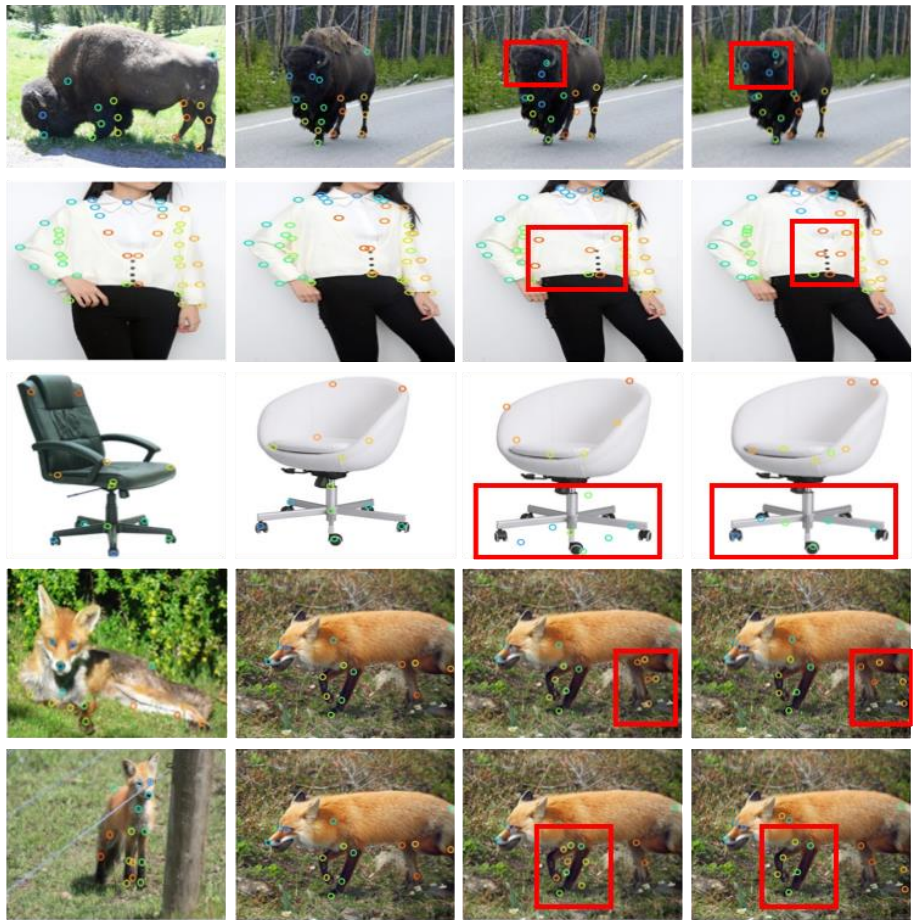


- Compositional Graph Transfer mechanism that dynamically aggregates multiple structural hypotheses into a unified, query-conditioned graph through attention-guided fusion.
- Bayesian confidence-weighted aggregation strategy.
- Query-guided gating.

Results - Qualitative



(a) Adjacency Matrix Visualization



(b) Qualitative Results

Results - Quantitative

Type	Method	Support	Split 1	Split 2	Split 3	Split 4	Split 5	Avg.
Image-support	POMNet Xu et al. (2022a)	Image	84.23	78.25	78.17	78.68	79.17	79.70
	CapeFormer Shi et al. (2023)	Image	89.45	84.88	83.59	83.53	85.09	85.31
	ESCAPE Nguyen et al. (2024)	Image	86.89	82.55	81.25	81.72	81.32	82.74
	MetaPoint+ Chen et al. (2024)	Image	90.43	85.59	84.52	84.34	85.96	86.17
	CapeFormer-T Shi et al. (2023)	Image	89.48	86.69	85.31	84.79	84.97	86.25
	SDPNet (HRNet-32) Ren et al. (2024)	Image	91.54	86.72	85.49	85.77	87.26	87.36
	SCAPE Liang et al. (2024)	Image	91.67	86.87	87.29	85.01	86.92	87.55
Text-support	CLAMP Zhang et al. (2023)	Text	72.37	-	-	-	-	-
	X-Pose Yang et al. (2024)	Image\Text	89.07	85.05	85.26	85.52	85.79	86.14
	PPM+CPT Peng et al. (2024)	Image + Text	91.03	88.06	84.48	86.73	87.40	87.54
	CapeX-S Rusanovsky et al. (2025)	Image + Text + Graph	95.17	88.88	87.72	88.24	91.81	90.37
Graph-support	GraphCape-T Hirschorn & Avidan (2024)	Image + Graph	91.19	87.81	85.68	85.87	85.61	87.23
	GenCape-T (Ours)	Image (Graph)	**92.05**	**88.69**	**86.89**	**85.88**	**87.02**	**88.09**
	GraphCape-S Hirschorn & Avidan (2024)	Image + Graph	94.73	89.79	**90.69**	88.09	90.11	90.68
	GenCape-S (Ours)	Image (Graph)	**95.23**	**90.60**	89.46	**89.32**	**90.43**	**91.01**

Conclusion

- We introduce GenCape, a generative framework for CAPE that infers keypoint relationships solely from visual inputs
- GenCape integrates an i-VAE to progressively infer instance-specific keypoint relationships, alongside a CGT module that aggregates multiple latent graph hypotheses into query-aware structural cues.
- State-of-the-art performance
- Our code will be publicly available !

Our code will be publicly available !