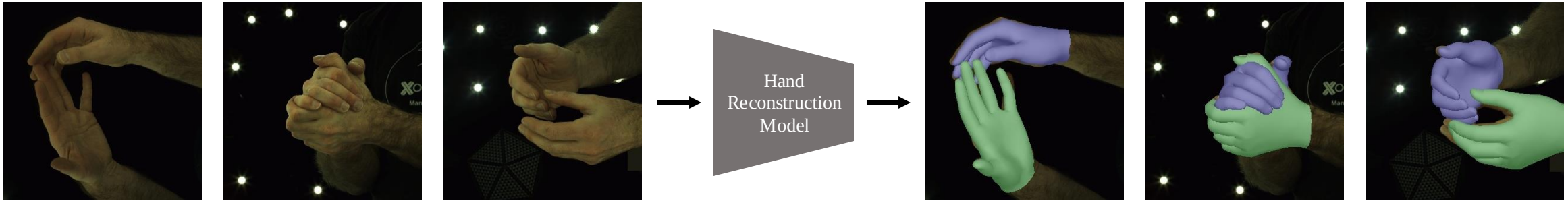
Project Page

EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO 2024

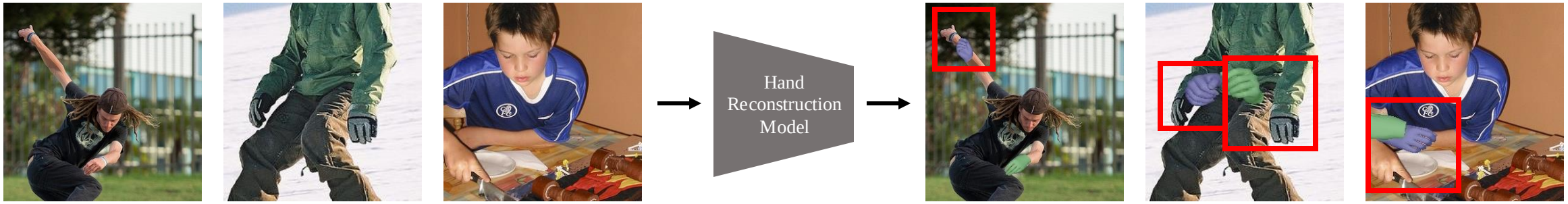
AttentionHand: Text-driven Controllable Hand Image Generation for 3D Hand Reconstruction in the Wild

Junho Park[1,2*], Kyeongbo Kong[3*], Suk-Ju Kang[1†]
[1]Sogang University, [2]LG Electronics, [3]Pusan National University

Motivation



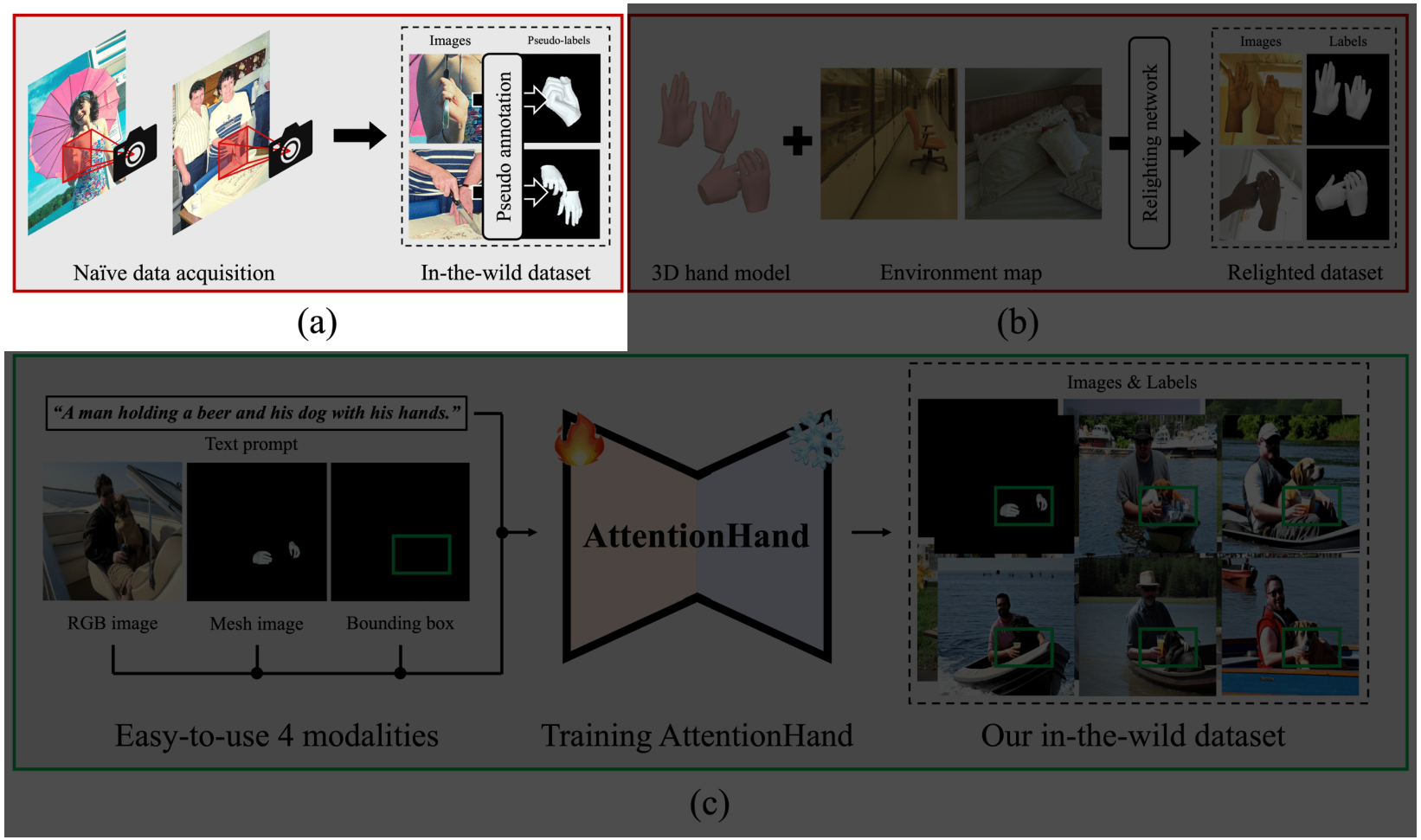
Well-reconstructed in in-the-lab scenes,



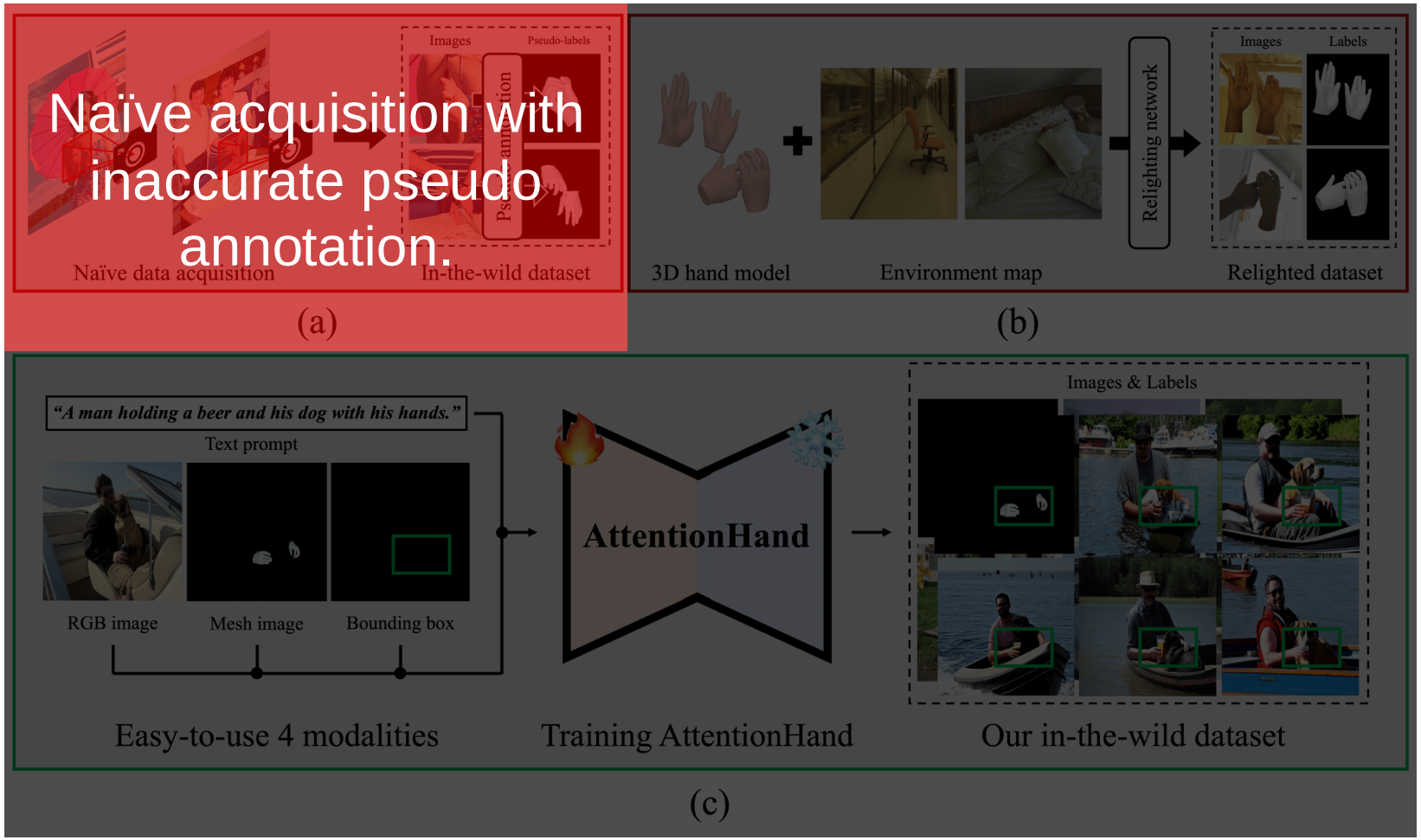
But get trouble in in-the-wild scenes.

Due to insufficiency of in-the-wild 3D hand datasets.

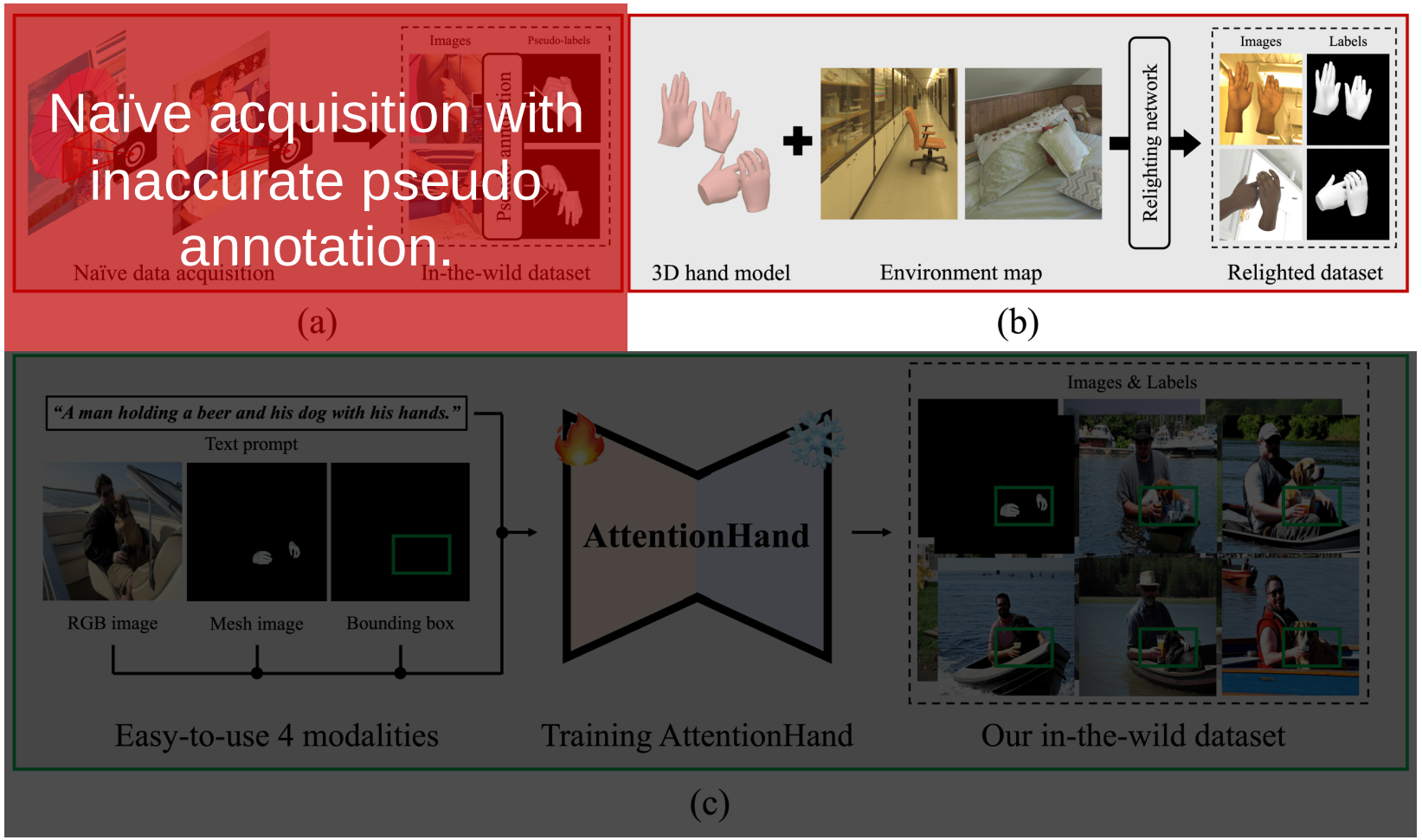
Motivation


Images
Pseudo-labels
Pseudo annotation
Naïve data acquisition
In-the-wild dataset
(a)
Relighting network
Images
Labels
3D hand model
Environment map
Relighted dataset
(b)
"A man holding a beer and his dog with his hands."
Text prompt
RGB image
Mesh image
Bounding box
AttentionHand
Images & Labels
Easy-to-use 4 modalities
Training AttentionHand
Our in-the-wild dataset
(c)

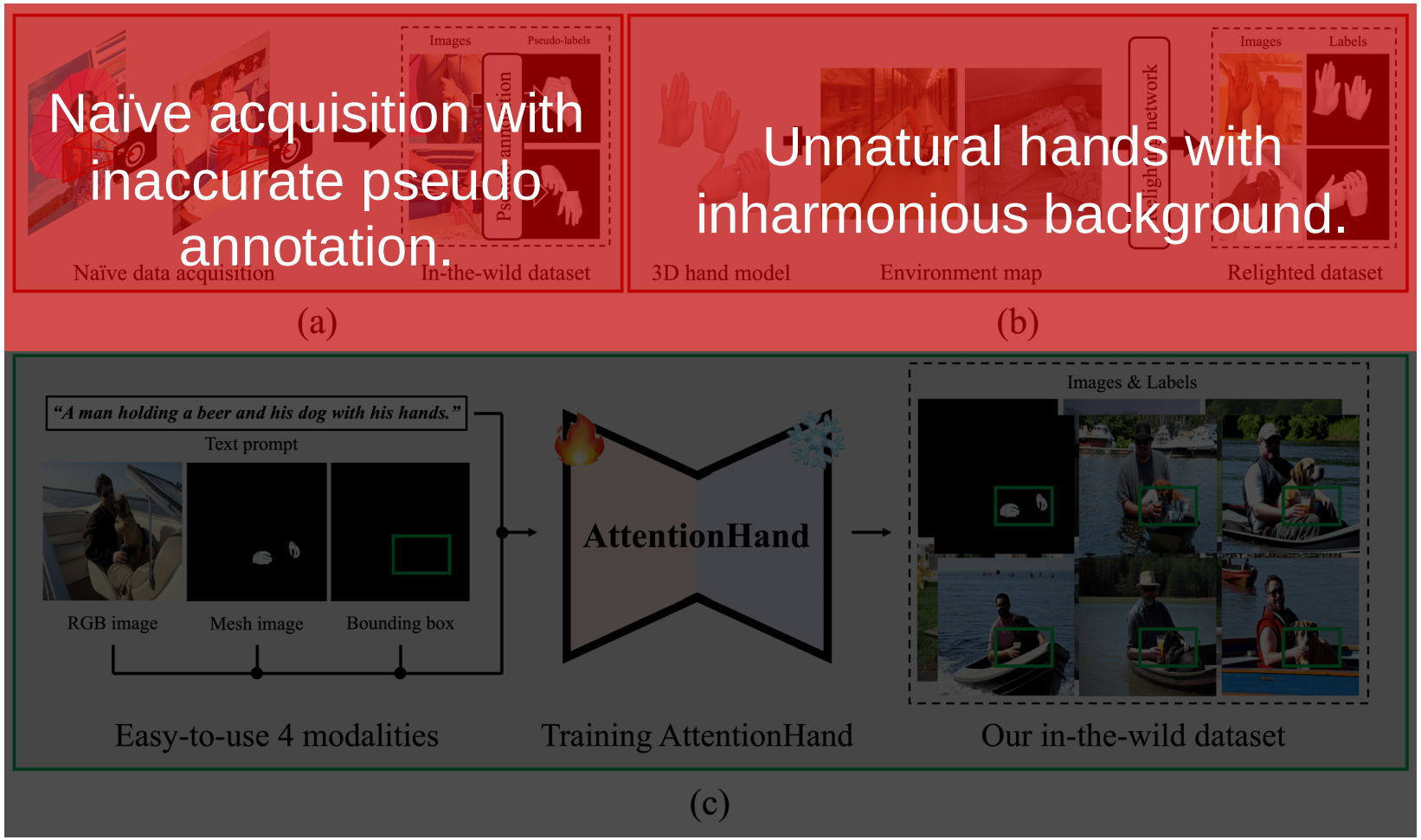
Motivation


Naïve acquisition with inaccurate pseudo annotation.
Images
Pseudo-labels
Pseudo annotation
Naïve data acquisition
In-the-wild dataset
(a)
Relighting network
Images
Labels
3D hand model
Environment map
Relighted dataset
(b)
"A man holding a beer and his dog with his hands."
Text prompt
RGB image
Mesh image
Bounding box
AttentionHand
Images & Labels
Easy-to-use 4 modalities
Training AttentionHand
Our in-the-wild dataset
(c)

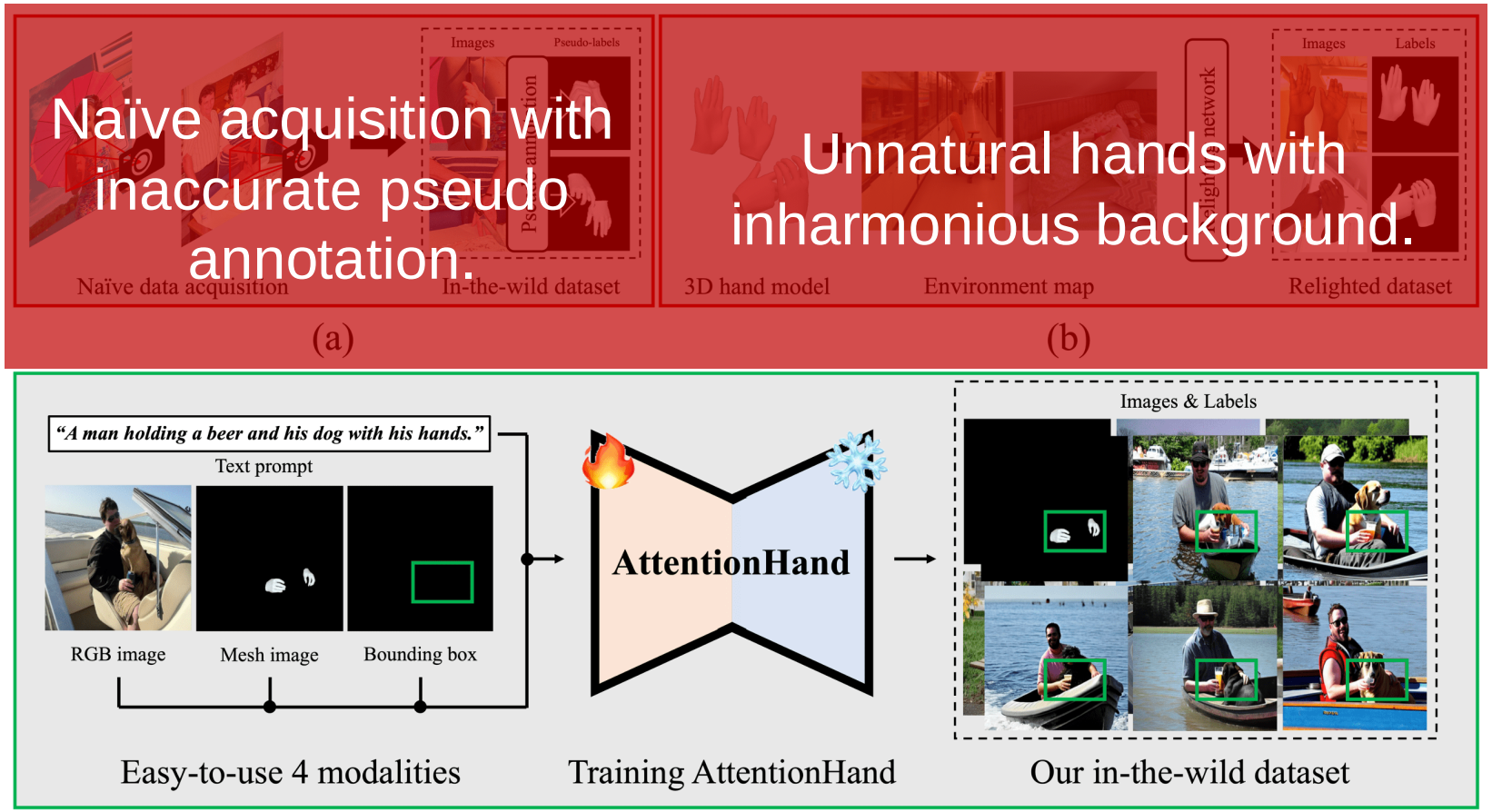
Motivation


Naïve acquisition with inaccurate pseudo annotation.
Images
Pseudo-labels
Pseudo annotation
Naïve data acquisition
In-the-wild dataset
(a)
Images
Labels
Relighting network
3D hand model
Environment map
Relighted dataset
(b)
"A man holding a beer and his dog with his hands."
Text prompt
RGB image
Mesh image
Bounding box
AttentionHand
Images & Labels
Easy-to-use 4 modalities
Training AttentionHand
Our in-the-wild dataset
(c)

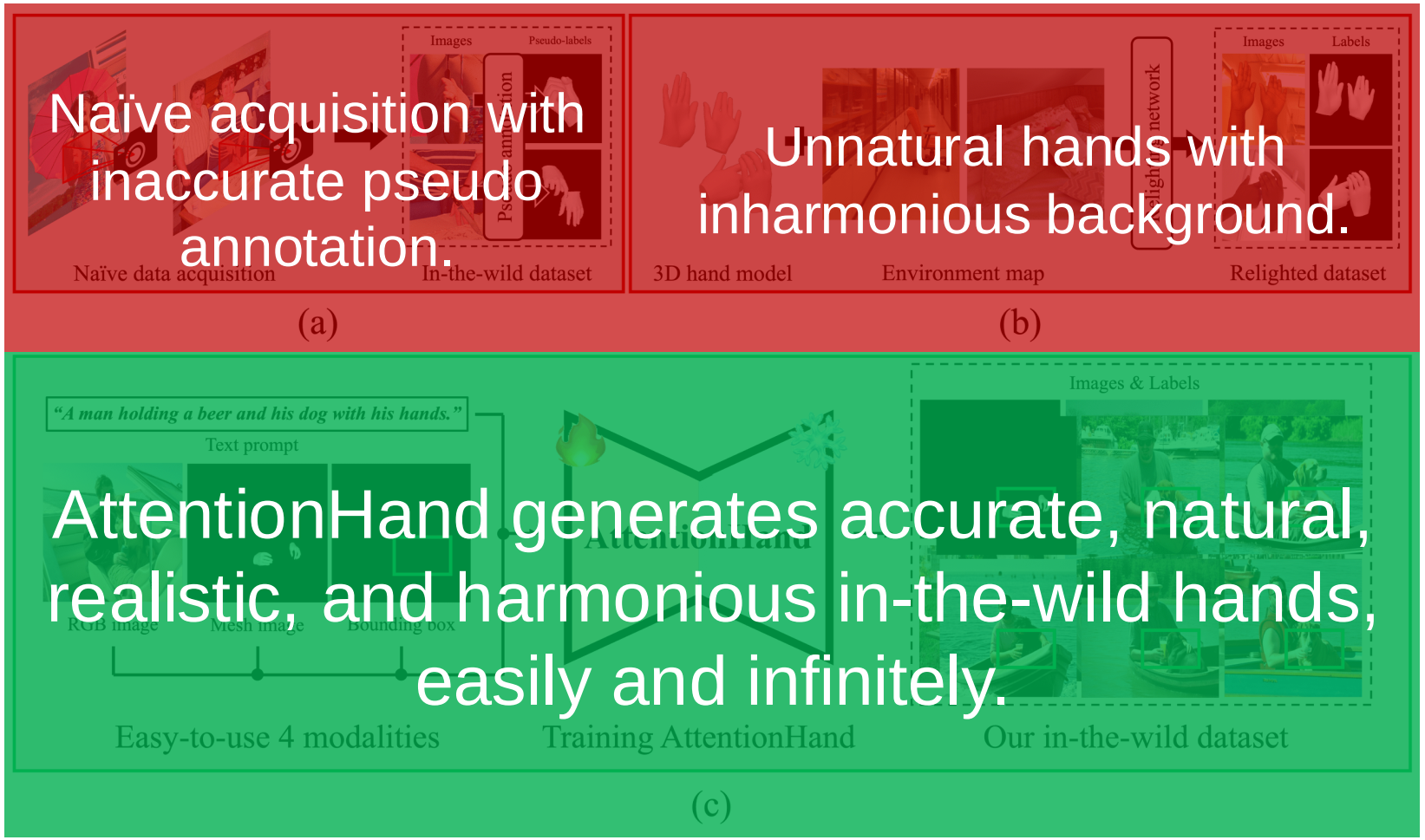
Motivation


Naïve acquisition with inaccurate pseudo annotation.
Images
Pseudo-labels
Naïve data acquisition
In-the-wild dataset
(a)
Unnatural hands with inharmonious background.
Images
Labels
3D hand model
Environment map
Relighted dataset
(b)
"A man holding a beer and his dog with his hands."
Text prompt
RGB image
Mesh image
Bounding box
AttentionHand
Images & Labels
Easy-to-use 4 modalities
Training AttentionHand
Our in-the-wild dataset
(c)

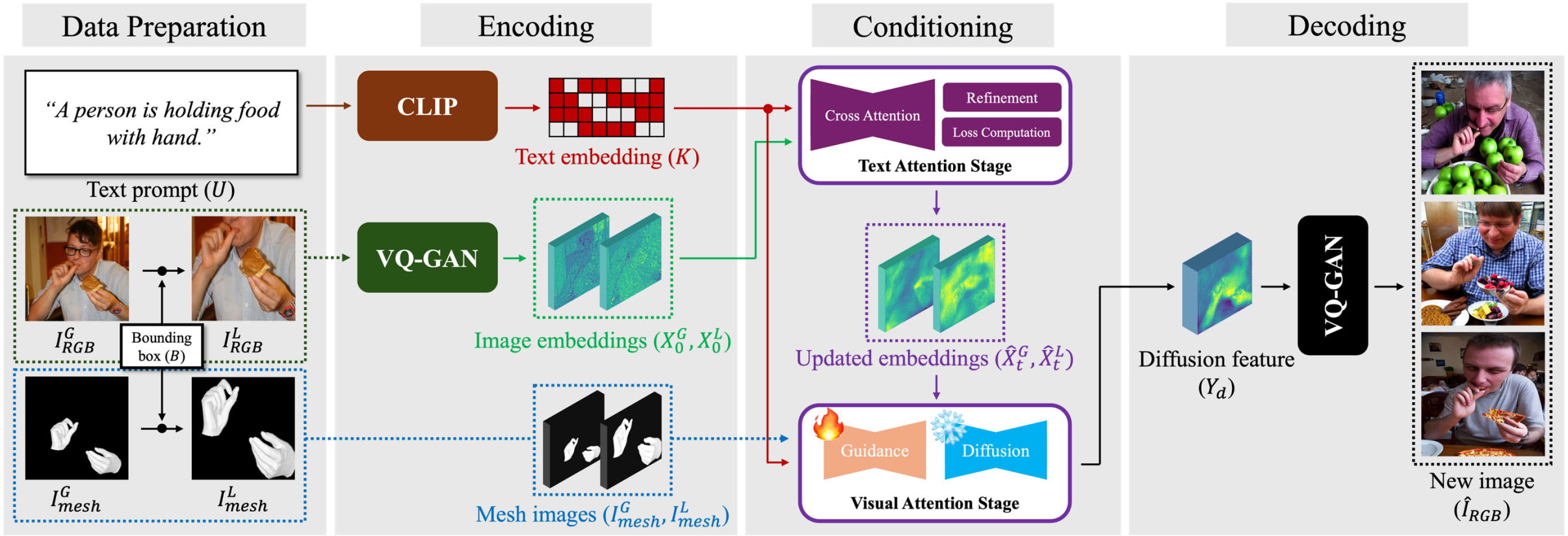
Motivation


Naïve acquisition with inaccurate pseudo annotation.
Images
Pseudo-labels
Naïve data acquisition
In-the-wild dataset
(a)
Unnatural hands with inharmonious background.
Images
Labels
3D hand model
Environment map
Relighted dataset
(b)
"A man holding a beer and his dog with his hands."
Text prompt
RGB image
Mesh image
Bounding box
AttentionHand
Images & Labels
Easy-to-use 4 modalities
Training AttentionHand
Our in-the-wild dataset
(c)

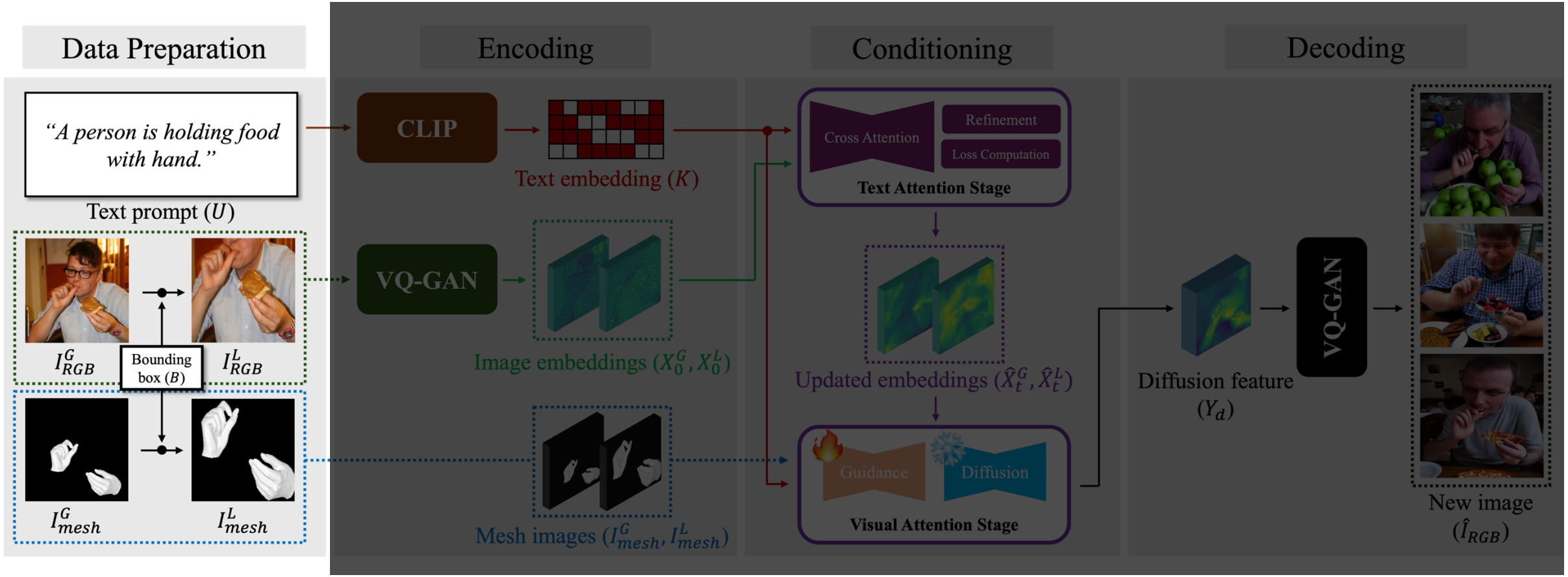
Motivation


Naïve acquisition with inaccurate pseudo annotation.
Images
Pseudo-labels
Naïve data acquisition
In-the-wild dataset
(a)
Unnatural hands with inharmonious background.
Images
Labels
3D hand model
Environment map
Relighted dataset
(b)
AttentionHand generates accurate, natural, realistic, and harmonious in-the-wild hands, easily and infinitely.
"A man holding a beer and his dog with his hands."
Text prompt
RGB image
Mesh image
Bounding box
Easy-to-use 4 modalities
Training AttentionHand
Images & Labels
Our in-the-wild dataset
(c)

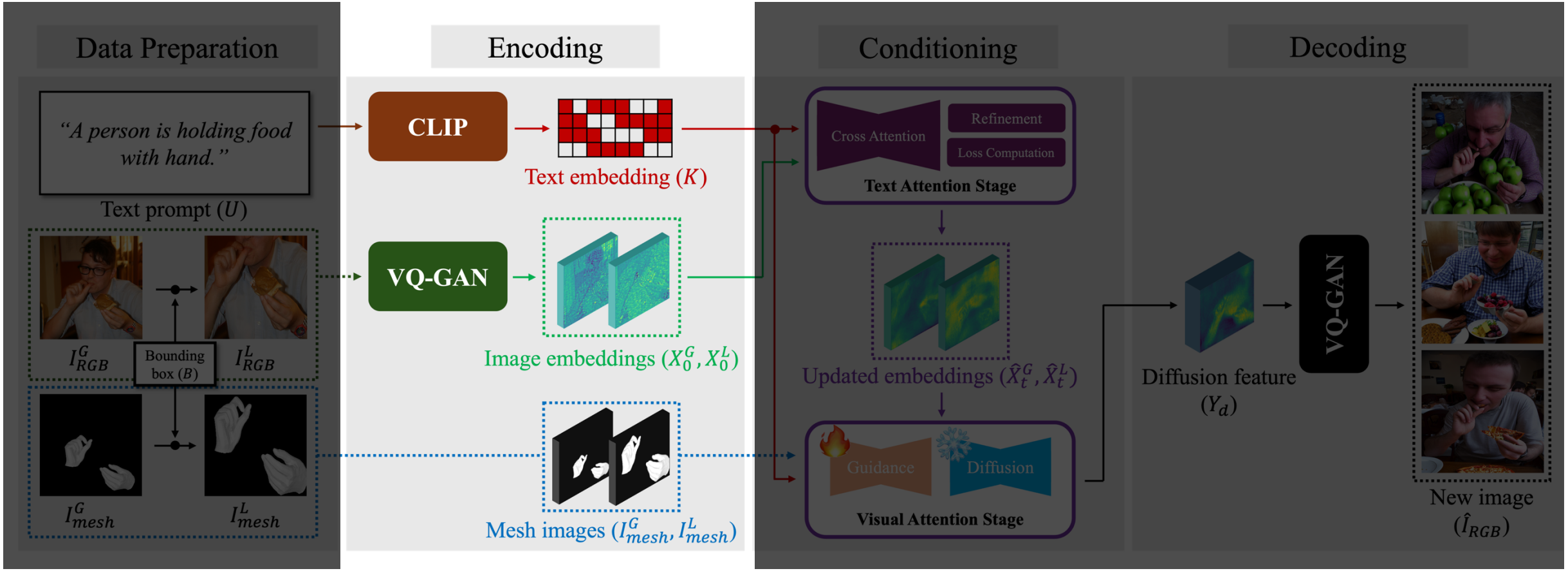
Methodology Overview



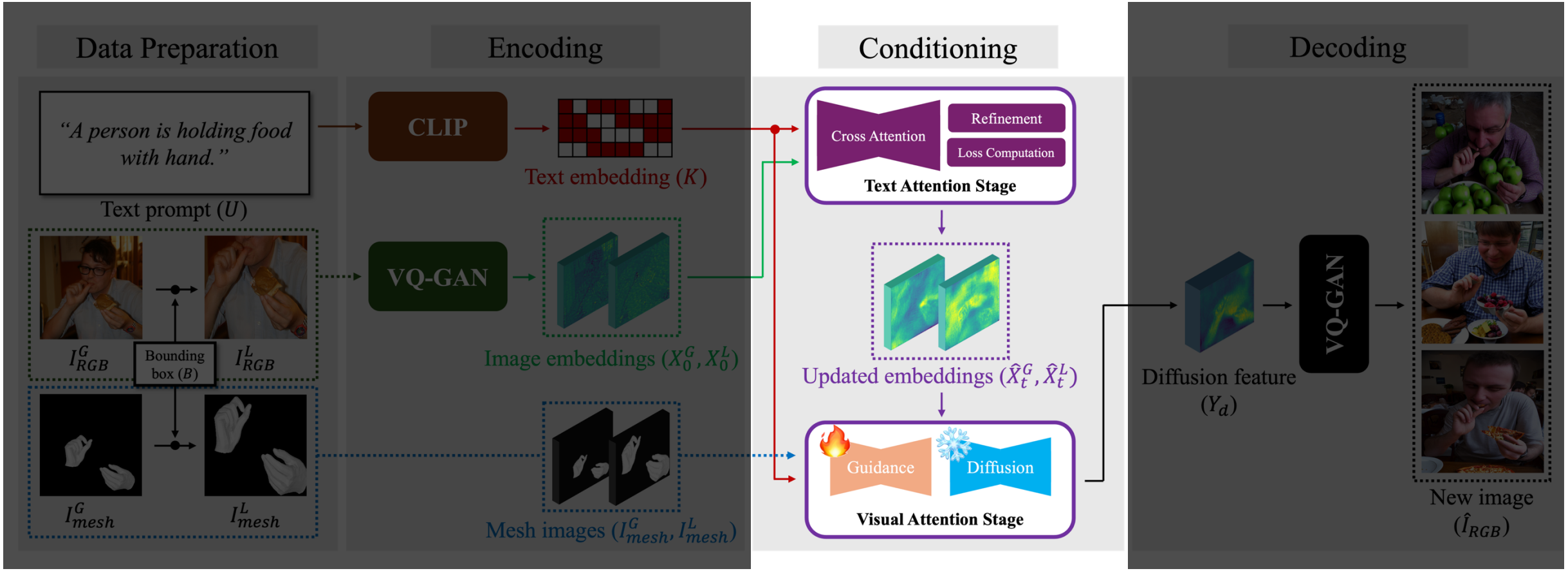
Methodology Overview



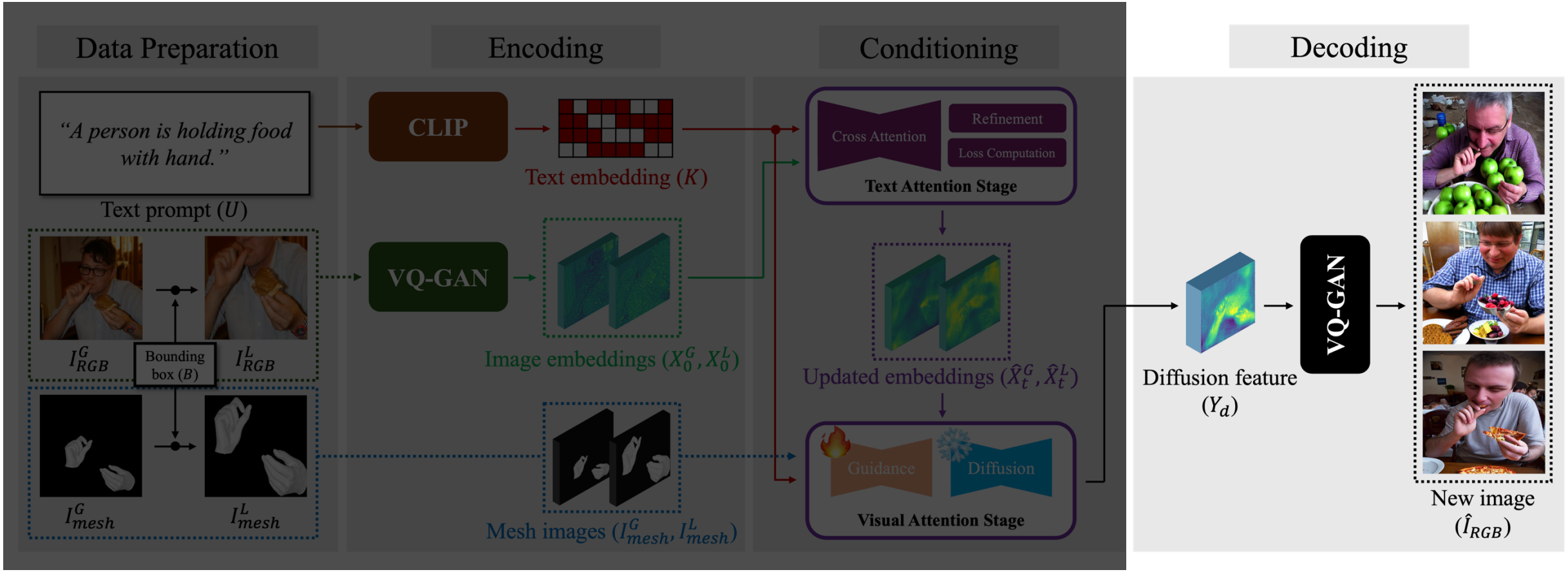
Methodology Overview



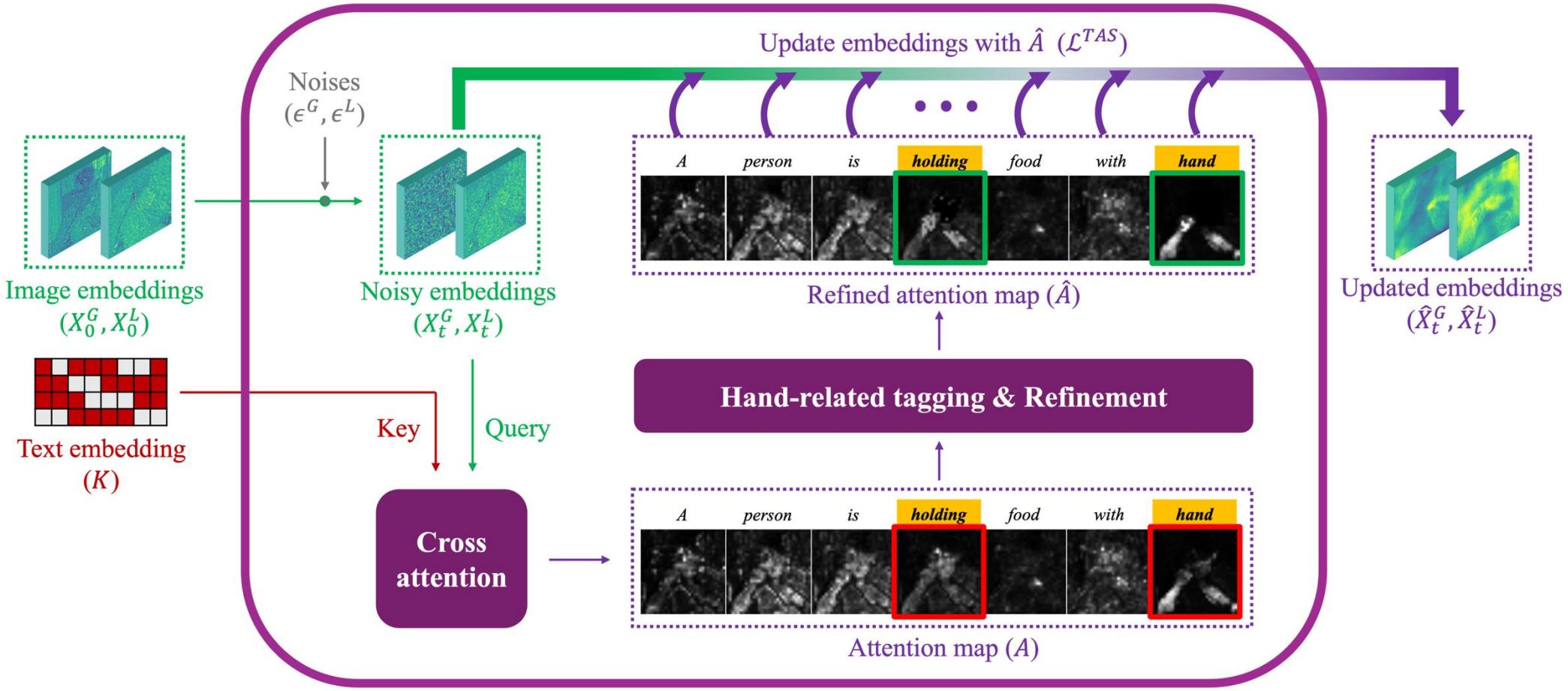
Methodology Overview



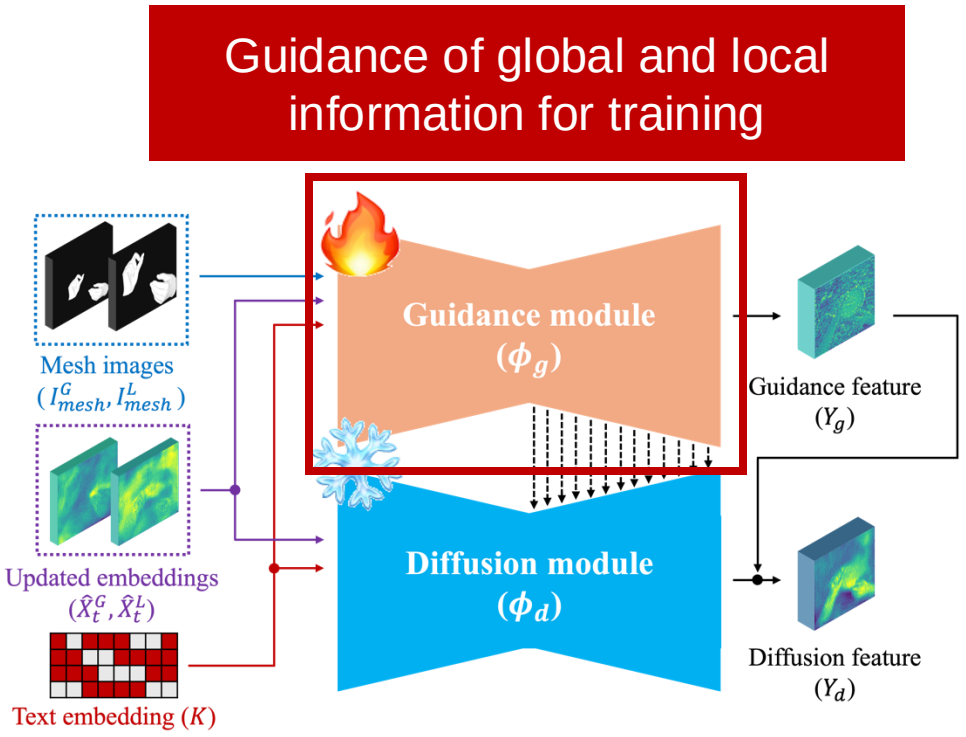
Methodology Overview

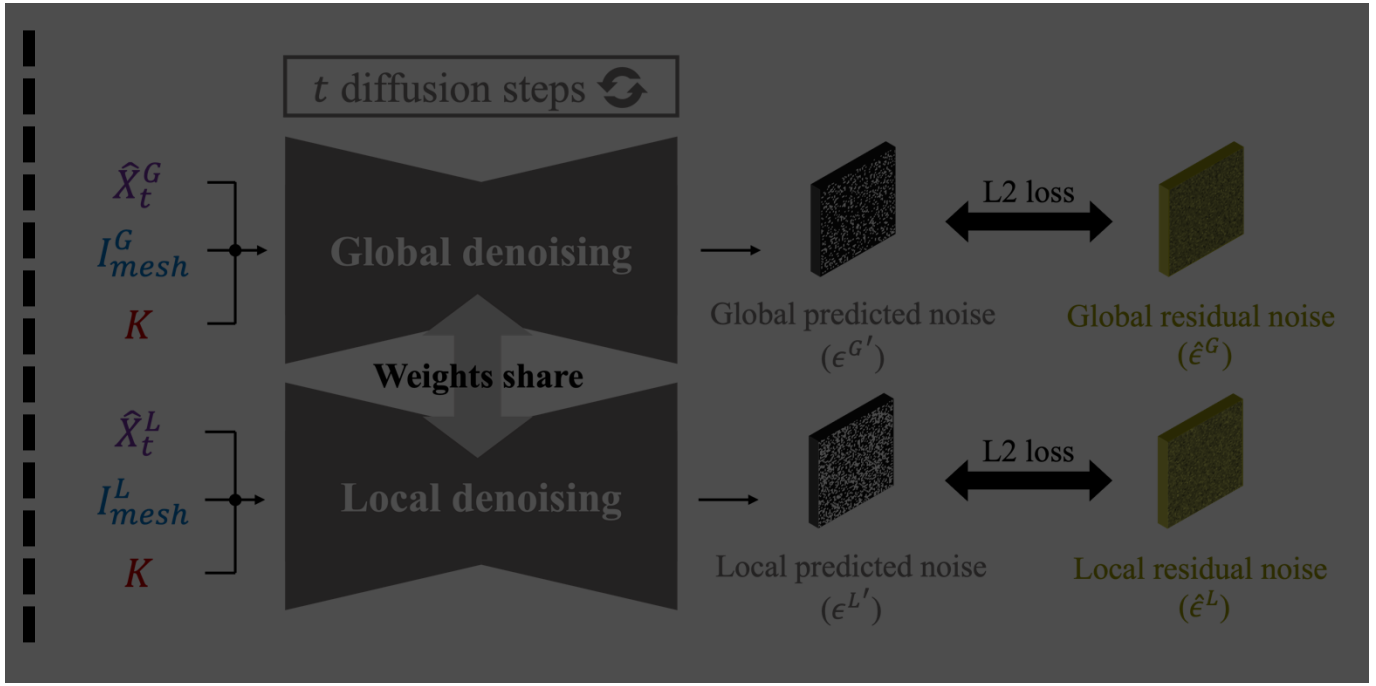


Methodology : Text Attention Stage

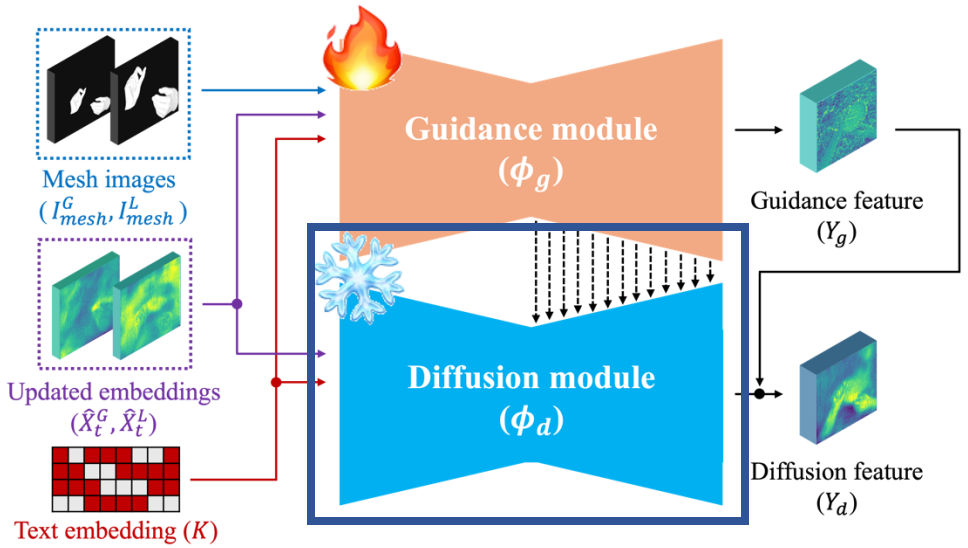


Methodology : Visual Attention Stage

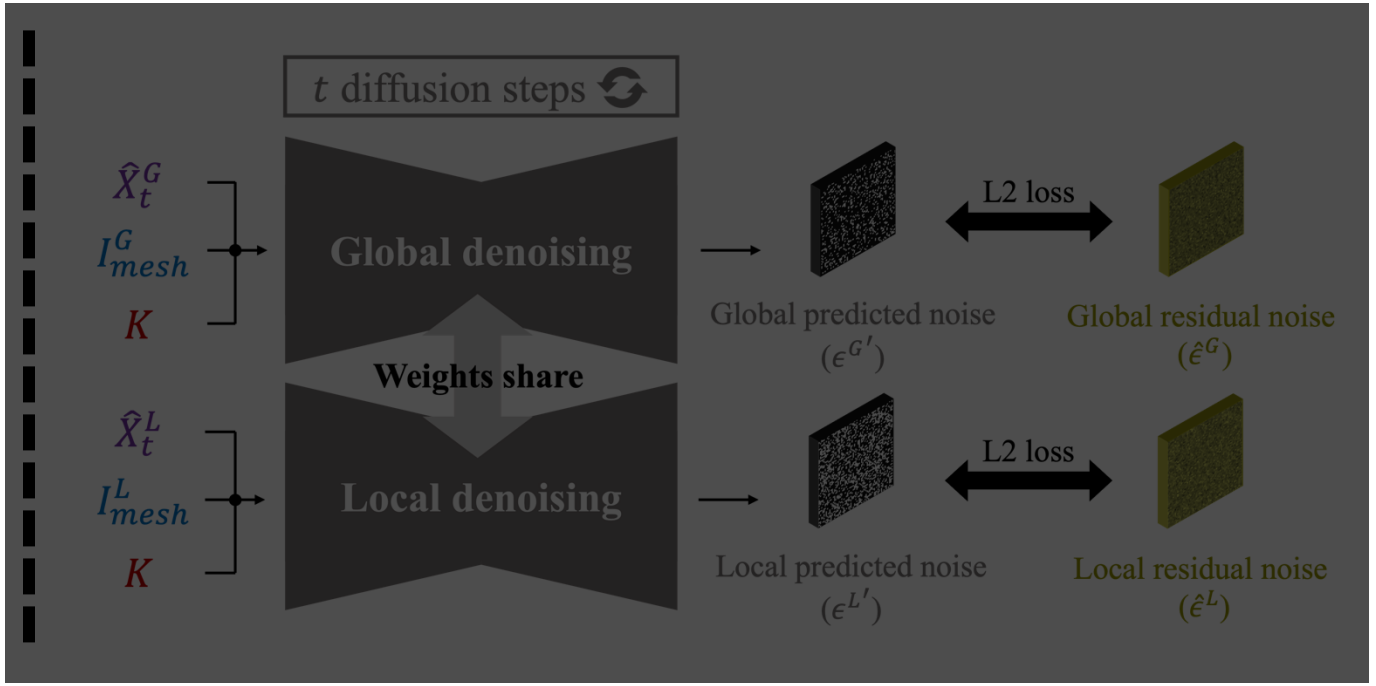




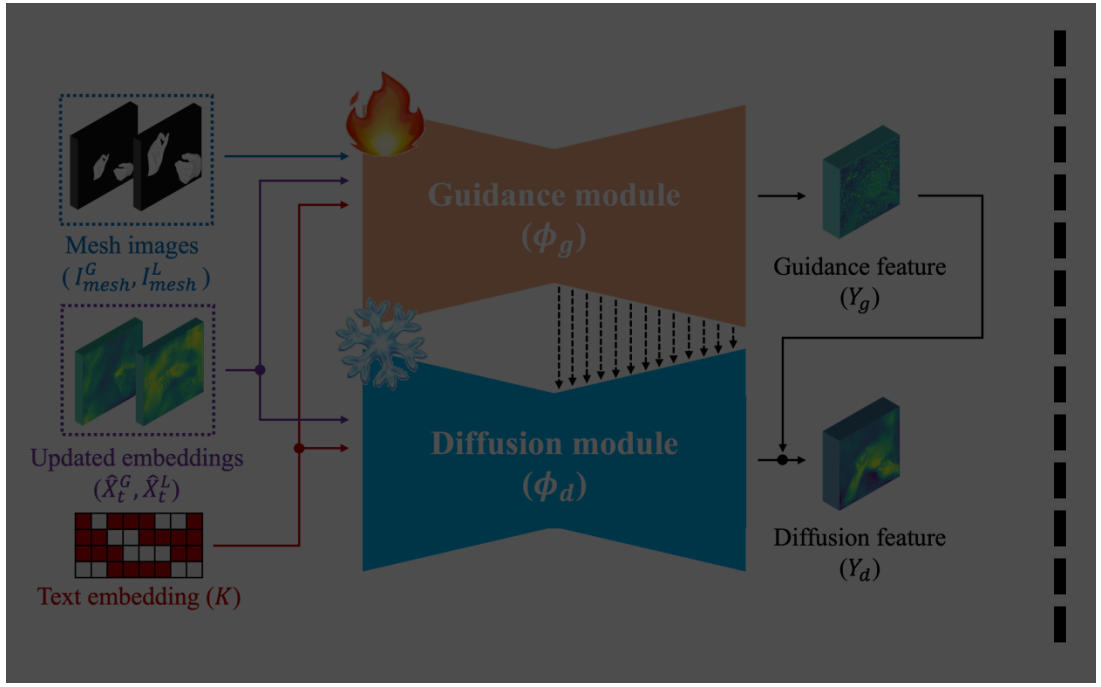
Methodology : Visual Attention Stage

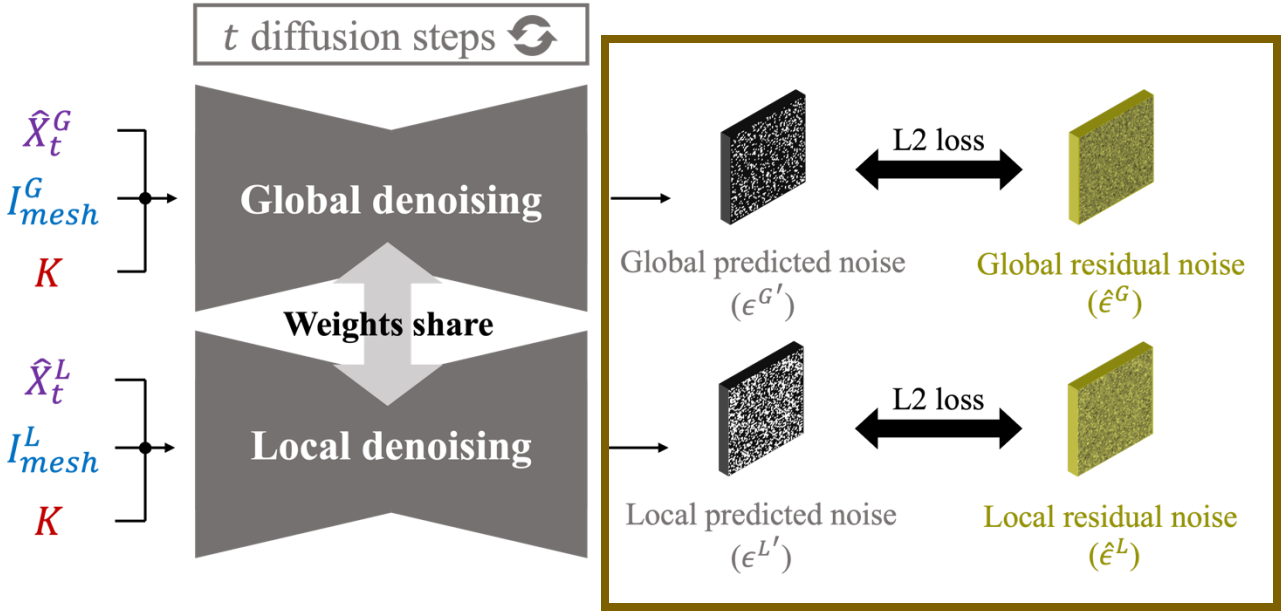


Leverage of image generation capability of Stable Diffusion



Methodology : Visual Attention Stage





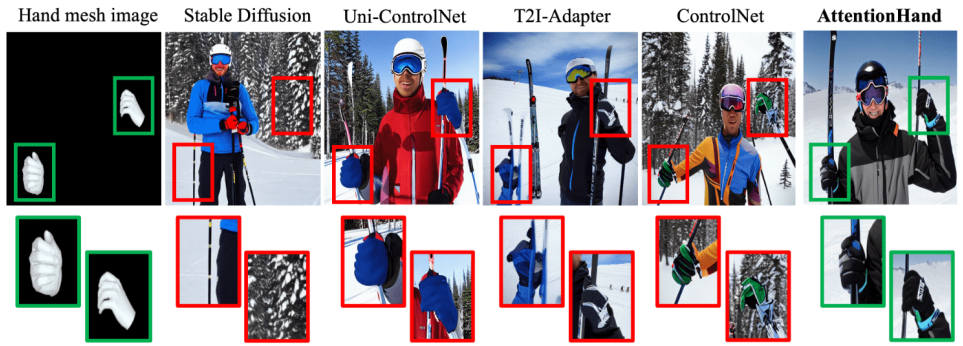
Simultaneous prediction of global and local noises

Comparisons with State-of-the-arts

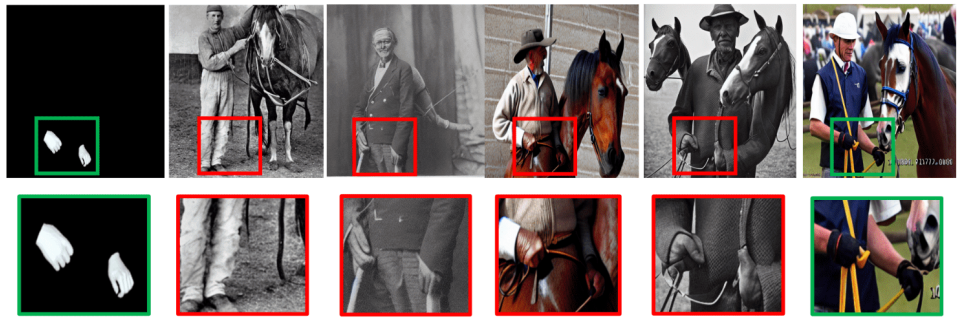
	FID↓	KID↓	FID-H↓	KID-H↓	Hand Conf.↑	MSE-2D↓	MSE-3D↓	User Pref.(%)↑
Stable Diffusion [1]	40.52	0.00684	50.78	0.02554	0.651	2.932	4.591	5.864
Uni-ControlNet [2]	30.34	0.00744	37.77	0.02004	0.855	2.105	3.039	8.796
T2I-Adapter [3]	22.00	0.00761	32.08	0.01568	0.914	1.546	2.451	19.676
ControlNet [4]	21.67	0.00658	40.32	0.02098	0.810	1.252	2.182	7.948
AttentionHand (w/o TAS)	21.27	0.00331	28.56	0.01390	0.955	1.211	2.042	20.734
AttentionHand (w/ TAS)	**20.71**	**0.00301**	**27.09**	**0.01287**	**0.965**	**1.026**	**1.986**	**36.905**

[1] Rombach, Robin *et al.*, "High-resolution image synthesis with latent diffusion models," CVPR 2022.
[2] Zhao, Shihao *et al.*, "Uni-ControlNet: All-in-one control to text-to-image diffusion models," NeurIPS 2023.
[3] Mou, Chong *et al.*, "T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," arXiv 2023.
[4] Zhang, Lvmin *et al.*, "Adding conditional control to text-to-image diffusion models," ICCV 2023.

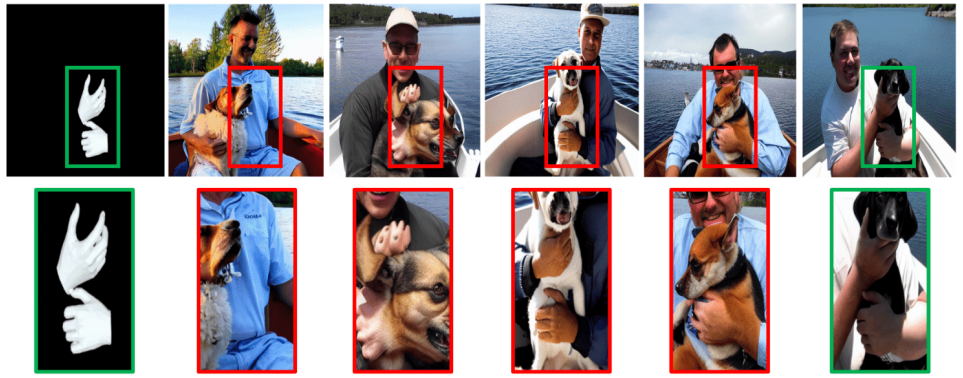
Comparisons with State-of-the-arts



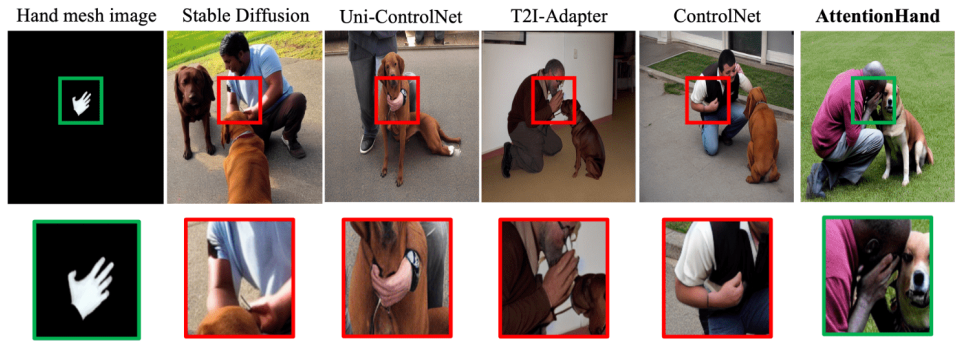
"A man is holding skis and ski poles."



"A man is holding the bridles of a horse."

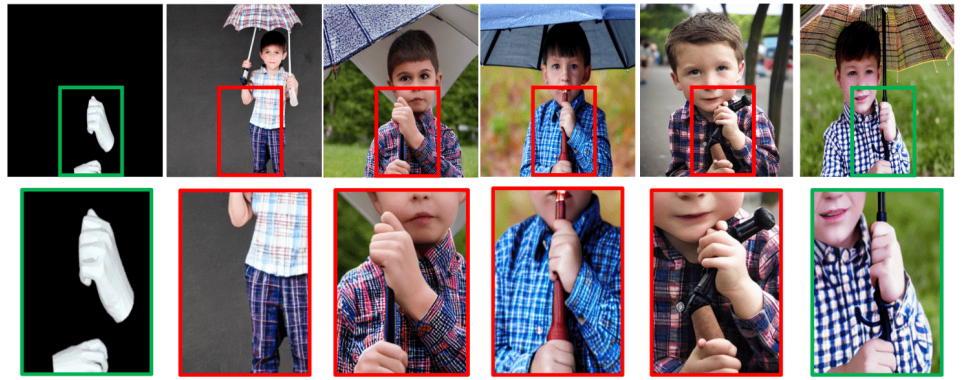


"A man on a boat is holding his dog in his lap."



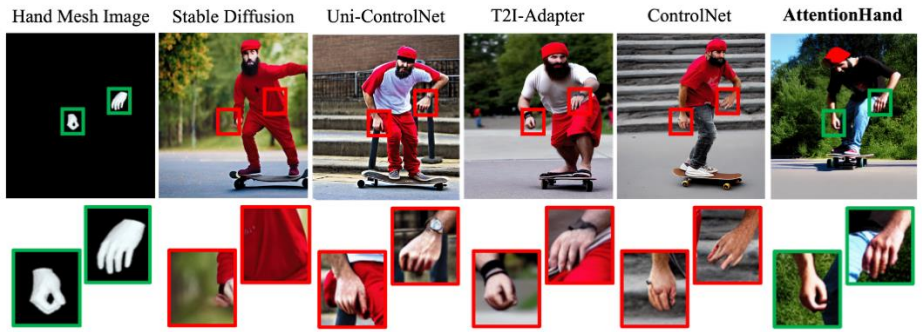
"A man is stroking a brown dog with his hand."

"A person is putting his hands in front of bowls with different kinds of food."



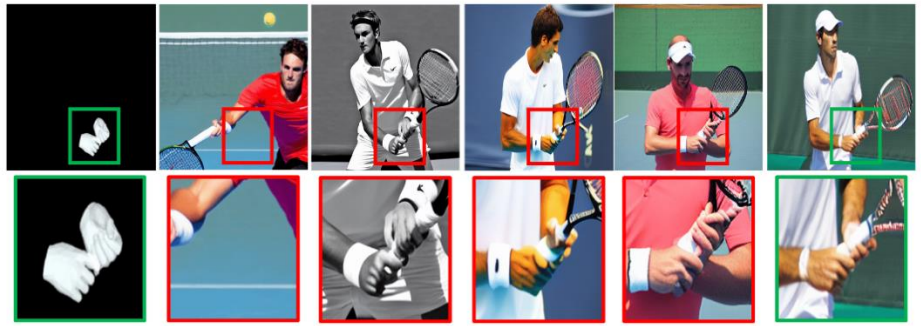
"A boy in a plaid shirt is holding an umbrella with his hands."

Comparisons with State-of-the-arts



"A bearded man in a red cap is riding on a skateboard."

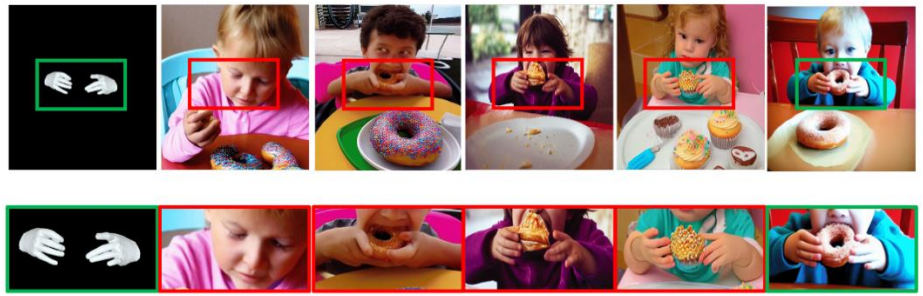
"Man is eating a bite of pizza with wall behind."



"A tennis player in action holding his racket with his hands."

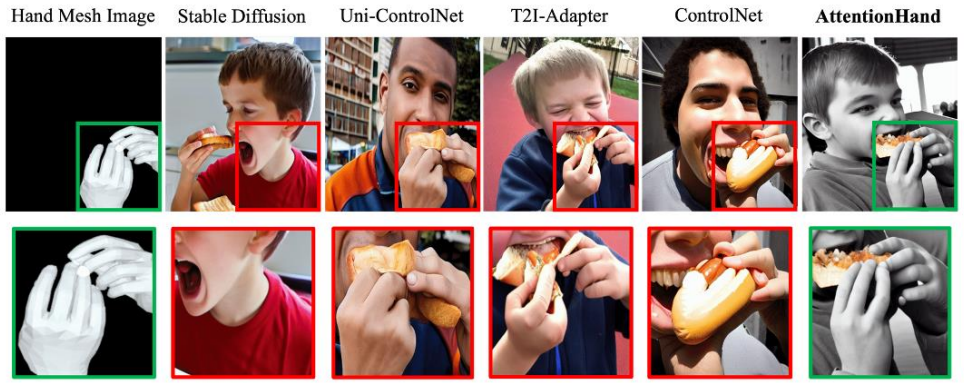
"A man in a hat is holding two cell phones with his hands."

"A man is holding a camera up over his left shoulder."



"A kid is eating a doughnut at a table."

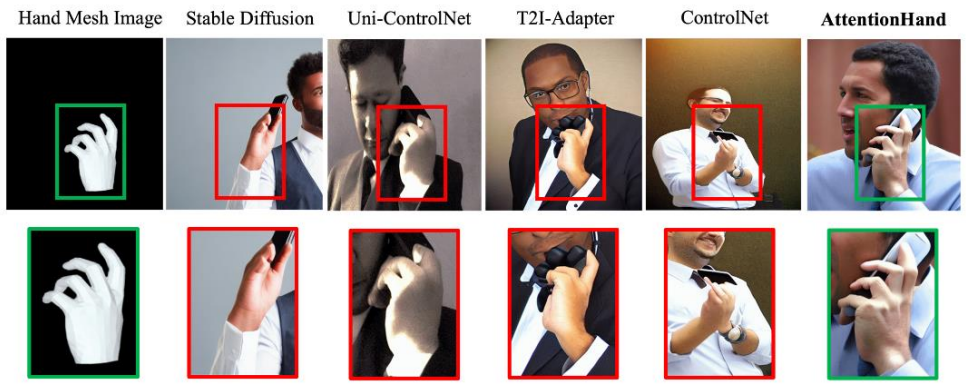
Comparisons with State-of-the-arts



"A young man is eating a sandwich."

"A boy is playing video games in his bedroom."

"A man is holding cell phone with hand."



"A man with a formal outfit is taking on a cell phone."

"A young man is shaping dough into balls with his hands."

"A man with headphones is taking his skateboard on his back."

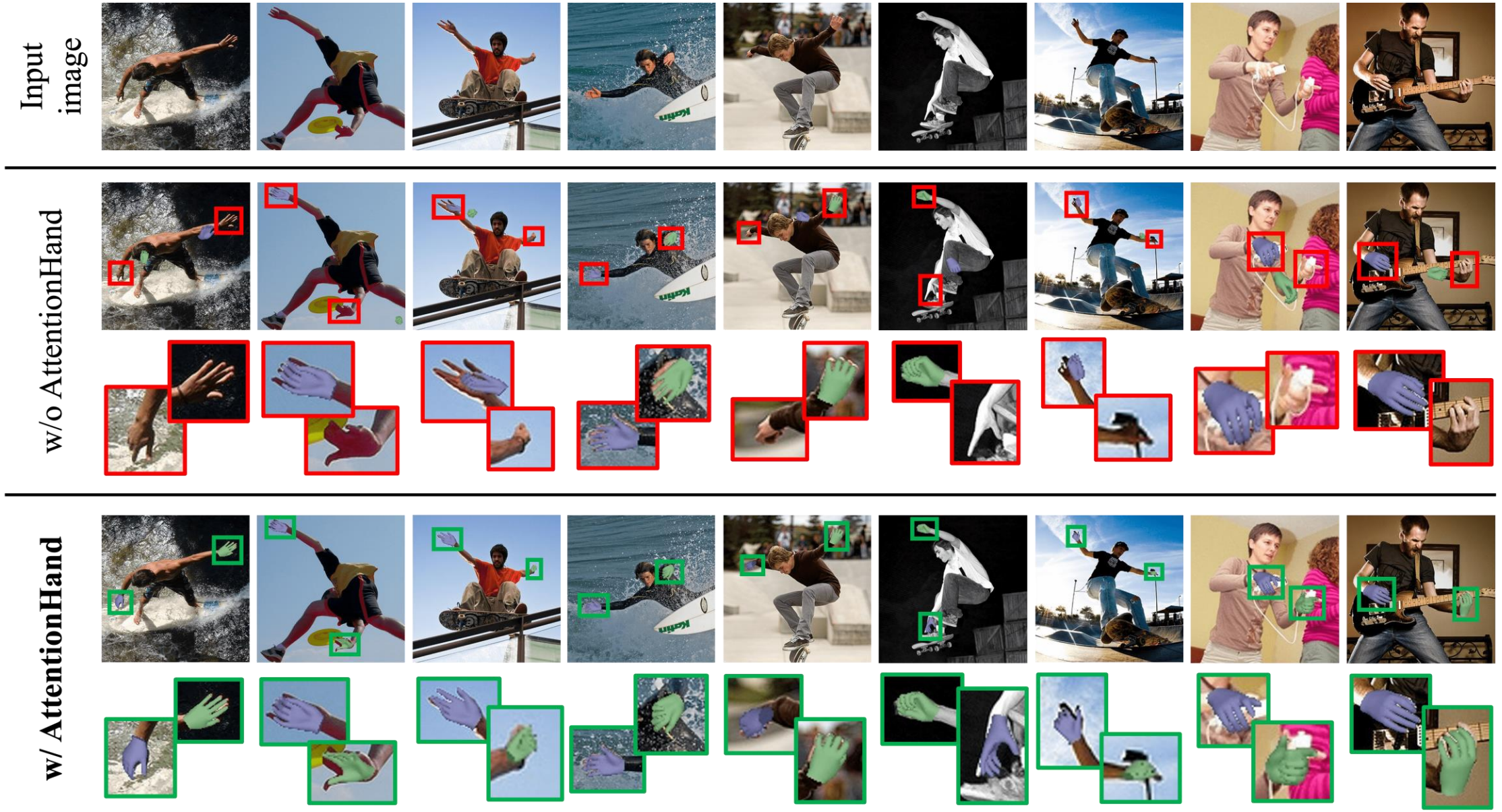
Comparisons with State-of-the-arts

	In-the-wild Scenes						In-the-lab Scenes		
	HIC [10]			Re:InterHand [11]			InterHand2.6M [12]		
	MPVPE↓	RRVE↓	MRRPE↓	MPVPE↓	RRVE↓	MRRPE↓	MPVPE↓	RRVE↓	MRRPE↓
IHMR [5]	38.57	45.51	119.64	30.90	45.55	98.45	16.94	21.98	33.39
IHMR + AttentionHand	**36.73**$_{-1.84}$	**44.10**$_{-1.41}$	**94.63**$_{-25.01}$	**29.11**$_{-1.79}$	**43.12**$_{-2.43}$	**87.07**$_{-11.38}$	**15.09**$_{-1.85}$	**20.55**$_{-1.43}$	**32.21**$_{-1.18}$
InterShape [6]	27.66	34.69	110.25	27.87	38.56	80.04	12.97	17.35	31.56
InterShape + AttentionHand	**25.04**$_{-2.62}$	**33.33**$_{-1.36}$	**80.17**$_{-30.08}$	**26.44**$_{-1.43}$	**36.54**$_{-2.02}$	**61.41**$_{-18.63}$	**11.90**$_{-1.07}$	**16.22**$_{-1.13}$	**30.04**$_{-1.52}$
IntagHand [7]	23.07	28.74	52.46	25.90	30.05	42.22	12.34	17.32	29.31
IntagHand + AttentionHand	**21.87**$_{-1.20}$	**27.09**$_{-1.65}$	**47.11**$_{-5.35}$	**23.39**$_{-2.51}$	**28.77**$_{-1.28}$	**33.98**$_{-8.24}$	**11.42**$_{-0.92}$	**15.81**$_{-1.51}$	**29.18**$_{-0.13}$
DIR [8]	21.89	26.11	43.11	21.82	29.66	37.01	10.26	17.11	28.98
DIR + AttentionHand	**20.66**$_{-1.23}$	**25.87**$_{-0.24}$	**40.54**$_{-2.57}$	**19.91**$_{-1.91}$	**26.67**$_{-2.99}$	**35.05**$_{-1.96}$	**10.09**$_{-0.17}$	**16.99**$_{-0.12}$	**28.02**$_{-0.96}$
InterWild [9]	15.30	21.35	31.26	13.99	20.07	22.38	11.52	19.77	26.87
InterWild + AttentionHand	**14.74**$_{-0.56}$	**21.10**$_{-0.25}$	**29.26**$_{-2.00}$	**13.95**$_{-0.04}$	**19.94**$_{-0.13}$	**22.05**$_{-0.33}$	**10.62**$_{-0.90}$	**19.09**$_{-0.68}$	**25.74**$_{-1.13}$

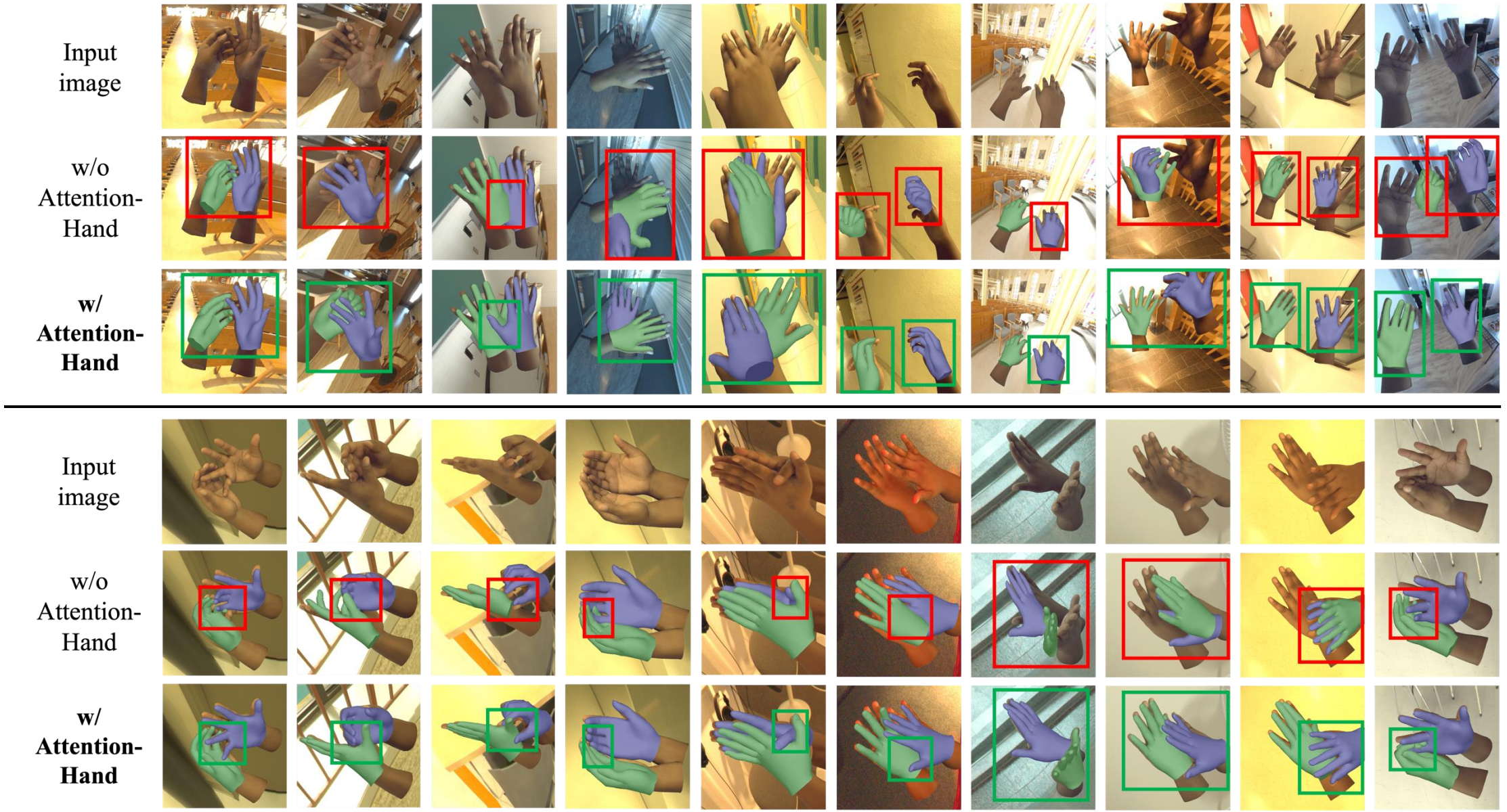
[5] Rong, Yu *et al.*, "Monocular 3D reconstruction of interacting hands via collision-aware factorized refinements," 3DV 2021.
[6] Zhang, Baowen *et al.*, "Interacting two-hand 3D pose and shape reconstruction from single color image," ICCV 2021.
[7] Li, Mengcheng *et al.*, "Interacting attention graph for single image two-hand reconstruction," CVPR 2022.
[8] Ren, Pengfei *et al.*, "Decoupled iterative refinement framework for interacting hands reconstruction from a single RGB image," ICCV 2023.
[9] Moon, Gyeongsik, "Bringing inputs to shared domains for 3D interacting hands recovery in the wild," CVPR 2023.
[10] Tzionas, Dimitrios *et al.*, "Capturing hands in action using discriminative salient points and physics simulation," IJCV 2016.
[11] Moon, Gyeongsik *et al.*, "A dataset of relighted 3d interacting hands," NeurIPS 2023.
[12] Moon, Gyeongsik *et al.*, "Interhand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image," ECCV 2020.

Comparisons with State-of-the-arts

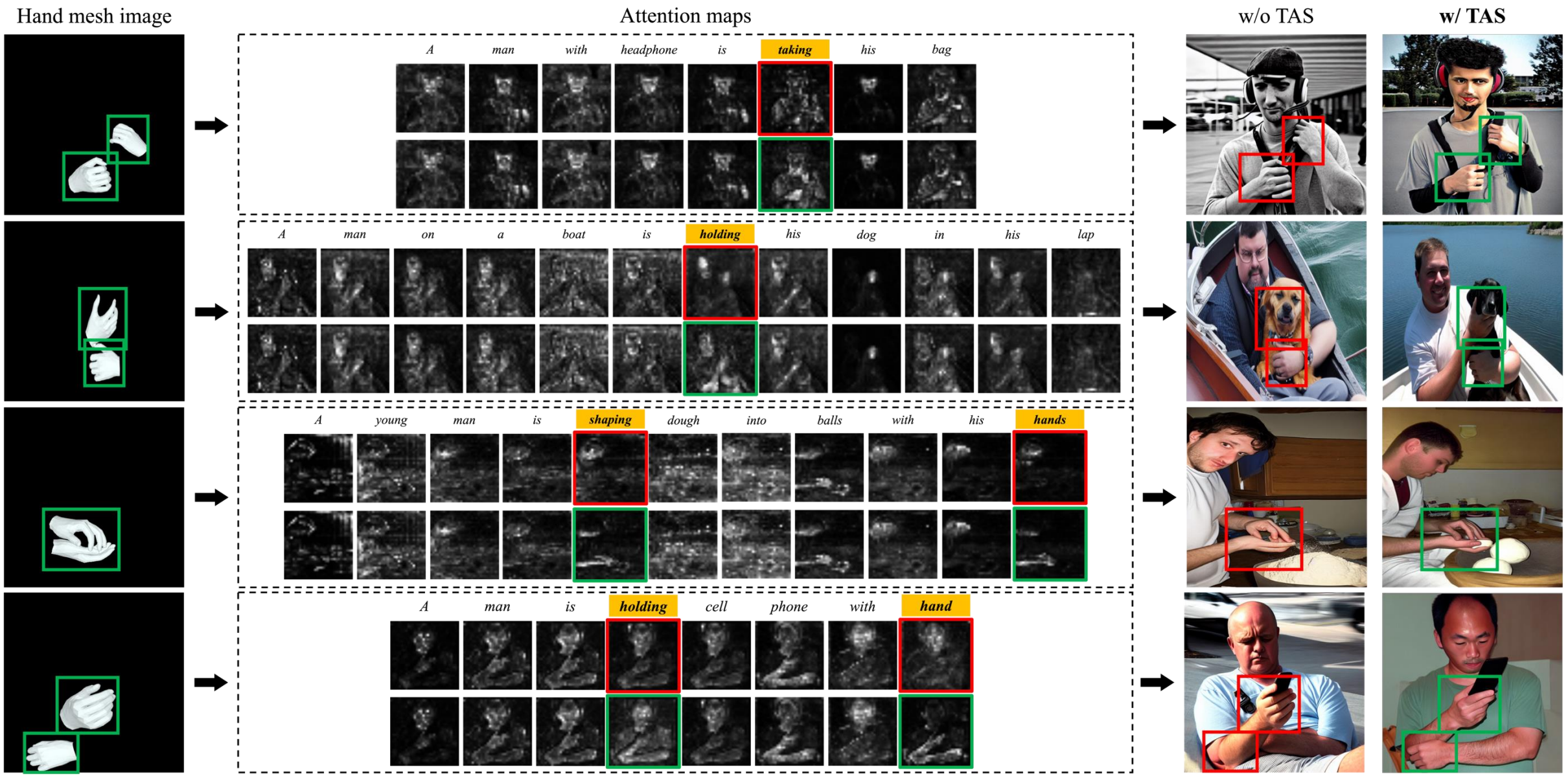
Comparisons with State-of-the-arts



Comparisons with State-of-the-arts

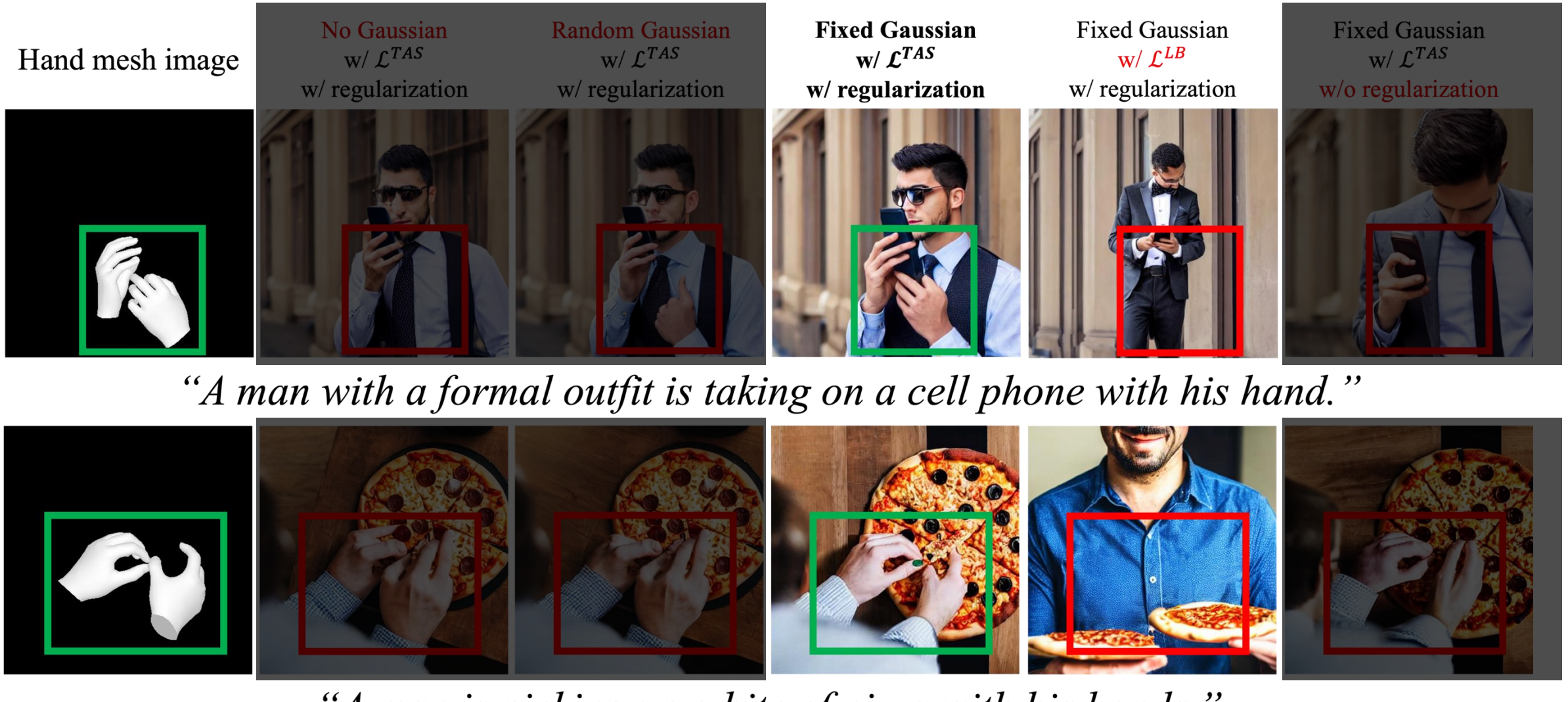


Exploration of Text Attention Stage

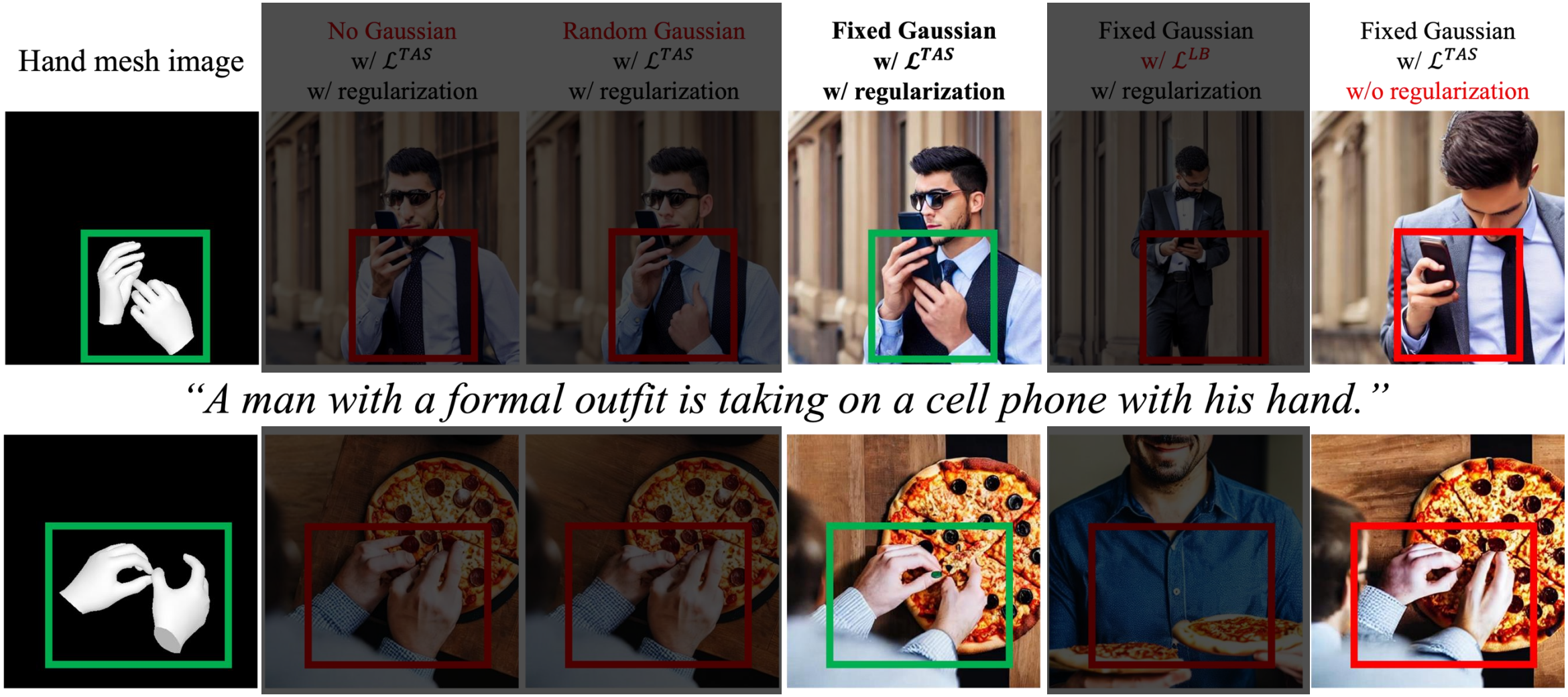


Exploration of Text Attention Stage

Exploration of Text Attention Stage

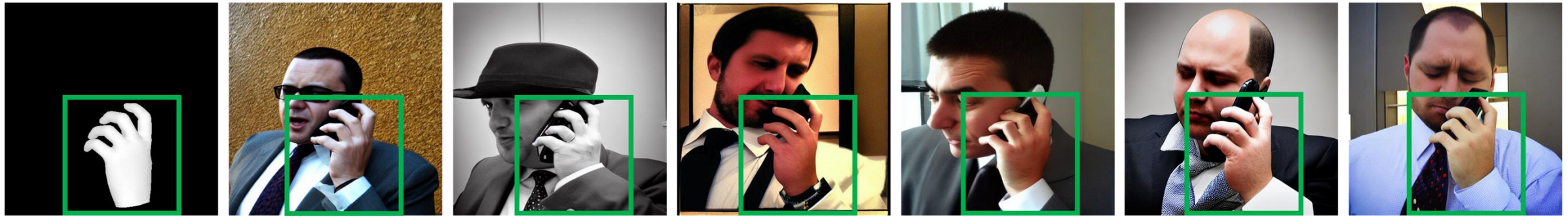


Exploration of Text Attention Stage



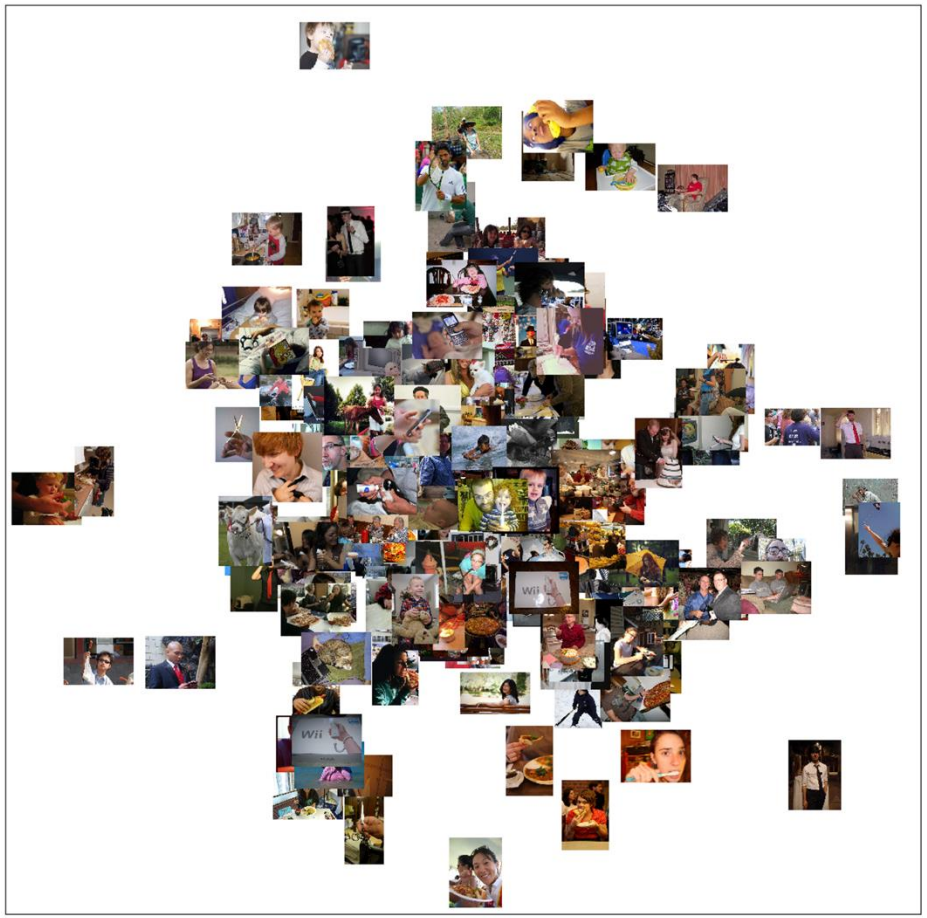
Robustness of Generated Dataset

"A young man is eating a sandwich."



"A man with a formal outfit is taking on a cell phone."

Robustness of Generated Dataset



MSCOCO

AttentionHand

Project Page

EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO 2024

Thank you

📌 See you at ***Poster #249*** 📌

IHS 서강대학교 SOGANG UNIVERSITY

LG Electronics

부산대학교 PUSAN NATIONAL UNIVERSITY