NEP: Autoregressive Image Editing via Next Editing Token Prediction

Huimin Wu[1], Xiaojian Ma[1], Haozhe Zhao[2], Yanpeng Zhao[1], Qing Li[1]

[1]BIGAI [2]Peking University

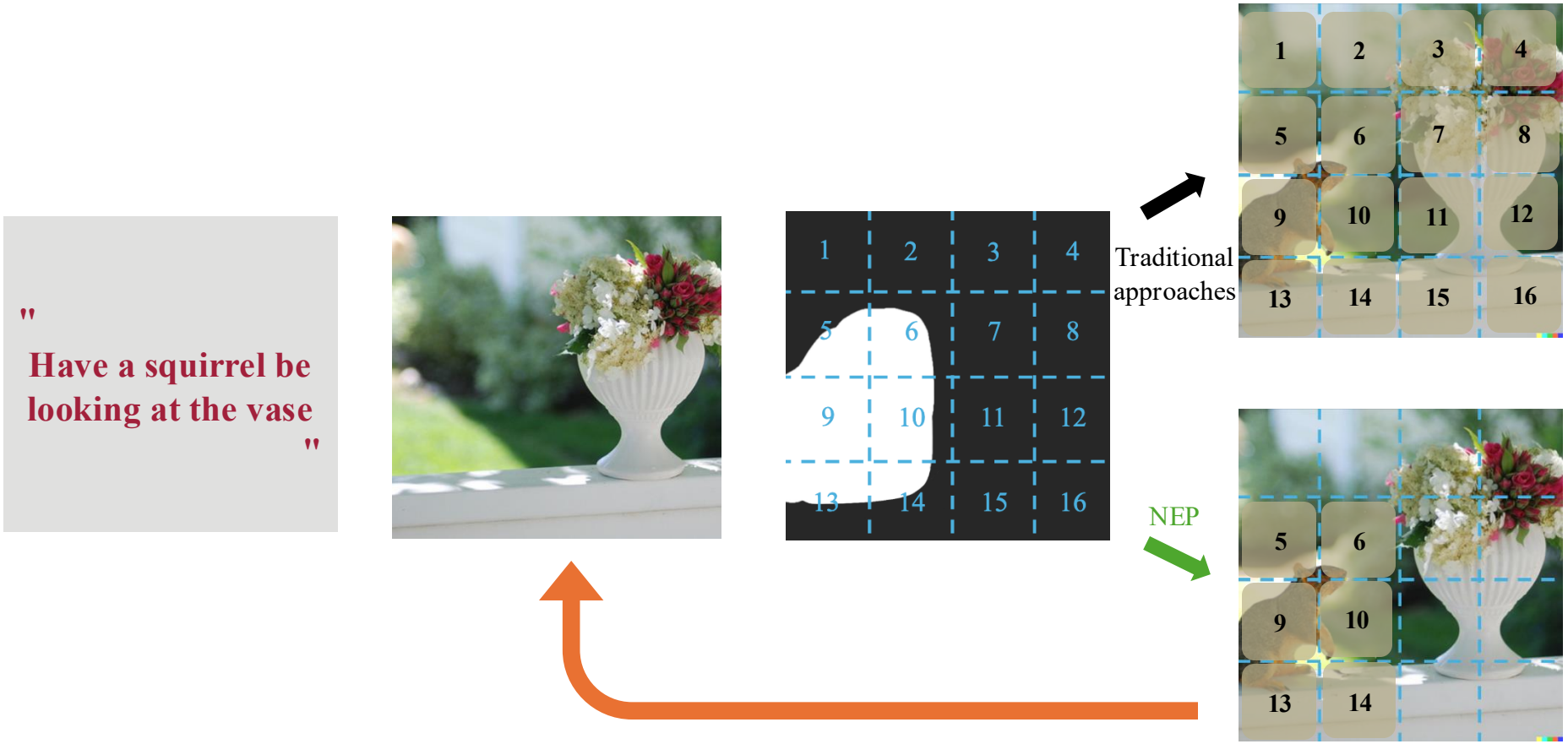
Text-driven Image Editing

"
Have a squirrel be looking at the vase
"

We propose NEP, a targeted and efficient autoregressive image editing method.

It allows **regeneration solely at edited regions**.

It makes **self-improvement** during the image generation process possible.



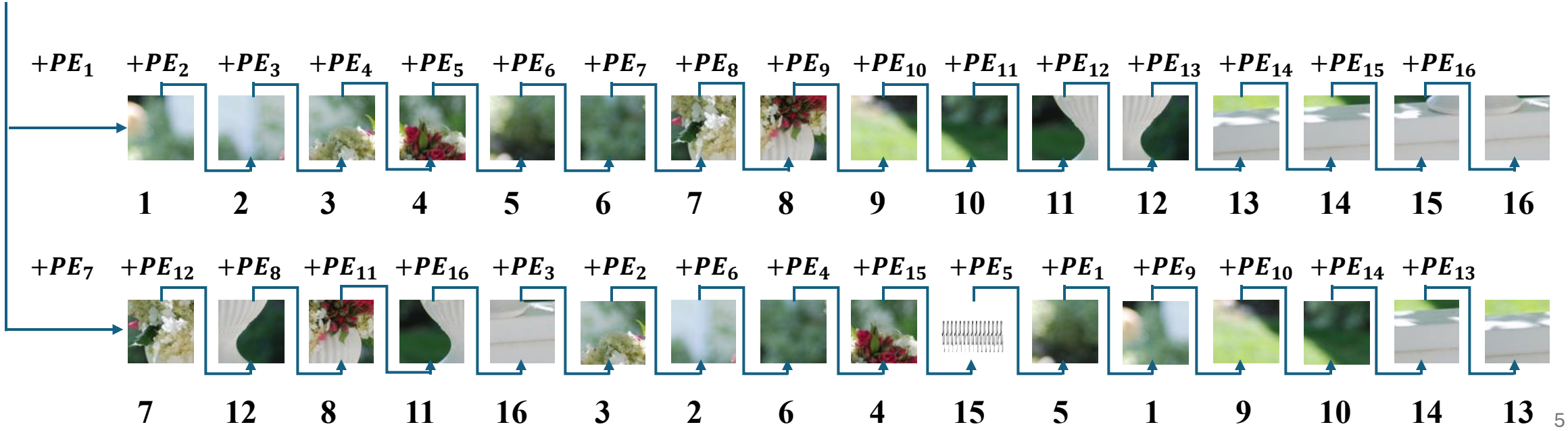
Two-stage training regime for NEP

1) NEP Pre-training
2) NEP fine-tuning

Two-stage training regime for NEP

1) NEP Pre-training
2) NEP fine-tuning

"
A flower vase is sitting on a porch stand
"



Two-stage training regime for NEP

1) NEP Pre-training
2) NEP fine-tuning

"Pancake with brown topping and ice cream."

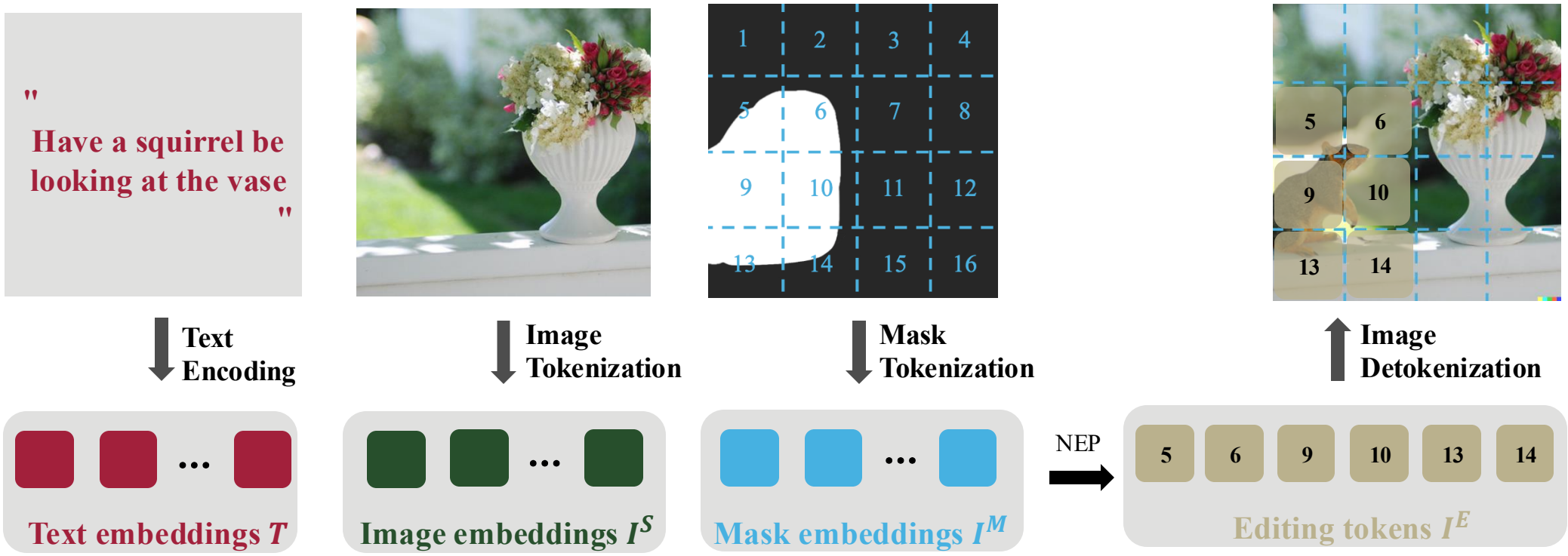
"A black refrigerator with its cabinet door closed in a kitchen."

"Club sandwich with fries and mustard."

"Red surfboard on grass in backyard."

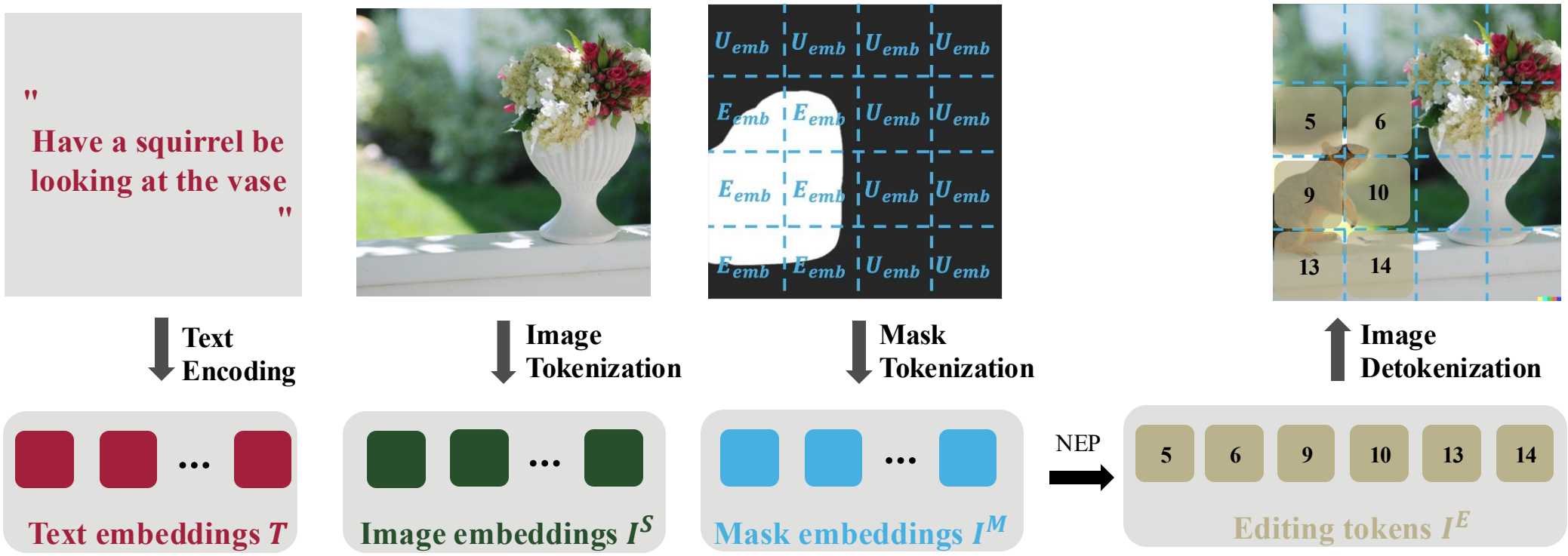
Two-stage training regime for NEP

1) NEP Pre-training
2) NEP fine-tuning



Two-stage training regime for NEP

1) NEP Pre-training
2) NEP fine-tuning



Two-stage training regime for NEP

Settings	Methods	L1↓	L2↓	CLIP-I↑	DINO↑
	Global Description-guided				
Single-turn	SD-SDEdit	0.1014	0.0278	0.8526	0.7726
	Null Text Inversion	0.0749	0.0197	0.8827	0.8206
	GLIDE	3.4973	115.8347	0.9487	0.9206
	Blended Diffusion	3.5631	119.2813	0.9291	0.8644
	Instruction-guided				
	HIVE	0.1092	0.0380	0.8519	0.7500
	InstructPix2Pix (IP2P)	0.1141	0.0371	0.8512	0.7437
	IP2P w/ MagicBrush	0.0625	0.0203	0.9332	0.8987
	UltraEdit	0.0575	0.0172	0.9307	0.8982
	FireEdit	0.0701	0.0238	0.9131	0.8619
	AnySD	0.1114	0.0439	0.8676	0.7680
	Ours	**0.0547**	**0.0163**	**0.9350**	**0.9044**
	Global Description-guided				
Multi-turn	SD-SDEdit	0.1616	0.0602	0.7933	0.6212
	Null Text Inversion	0.1057	0.0335	0.8468	0.7529
	GLIDE	11.7487	1079.5997	0.9094	0.8494
	Blended Diffusion	14.5439	1510.2271	0.8782	0.7690
	Instruction-guided				
	HIVE	0.1521	0.0557	0.8004	0.6463
	InstructPix2Pix (IP2P)	0.1345	0.0460	0.8304	0.7018
	IP2P w/ MagicBrush	0.0964	0.0353	0.8924	0.8273
	UltraEdit	0.0745	**0.0236**	0.9045	0.8505
	FireEdit	0.0911	0.0326	0.8819	0.8010
	AnySD	0.0748	0.0273	**0.9152**	**0.8623**
	Ours	**0.0707**	0.0269	0.9107	0.8493

Two-stage training regime for NEP

Input	Mask	Ground truth	IP2P	MagicBrush	UltraEdit	AnySD	NEP (ours)

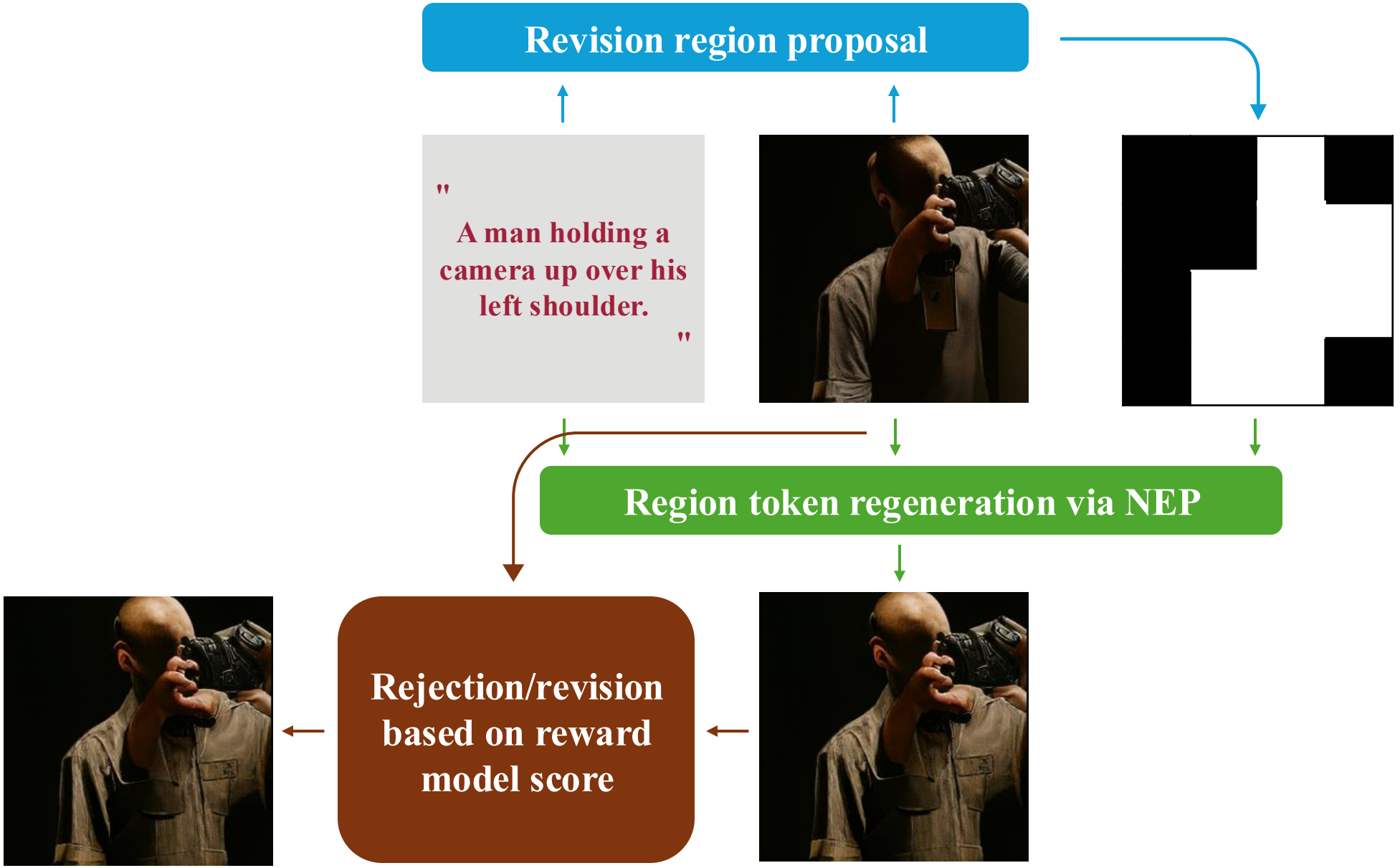
"Could it be a glass of wine on the table?"

"let there be sports on TV"

"let the woman wear a designer gown"

Test-time Scaling of NEP for Improving Image Quality

Test-time Scaling of NEP for Improving Image Quality



Test-time Scaling of NEP for Improving Image Quality

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

"A boat floating on a lake through a brick bridge."

"A passenger plane that is parked on the runway."

"A kitchen filled with lots of counter top space."

THANK YOU!