The Power of Context: How Multimodality Improves Image Super-Resolution

CVPR 2025

Kangfu Mei* [1,2], Hossein Talebi[1], Mojtaba Ardakani[1],
Vishal M. Patel[2], Peyman Milanfar[1], Mauricio Delbracio[1]
[1] Google, [2] Johns Hopkins University
Project Page: https://mmsr.kfmei.com/

Presented by Zejia Fan
2025.10.12

Background

- Super resolution with diffusion needs semantic guidance

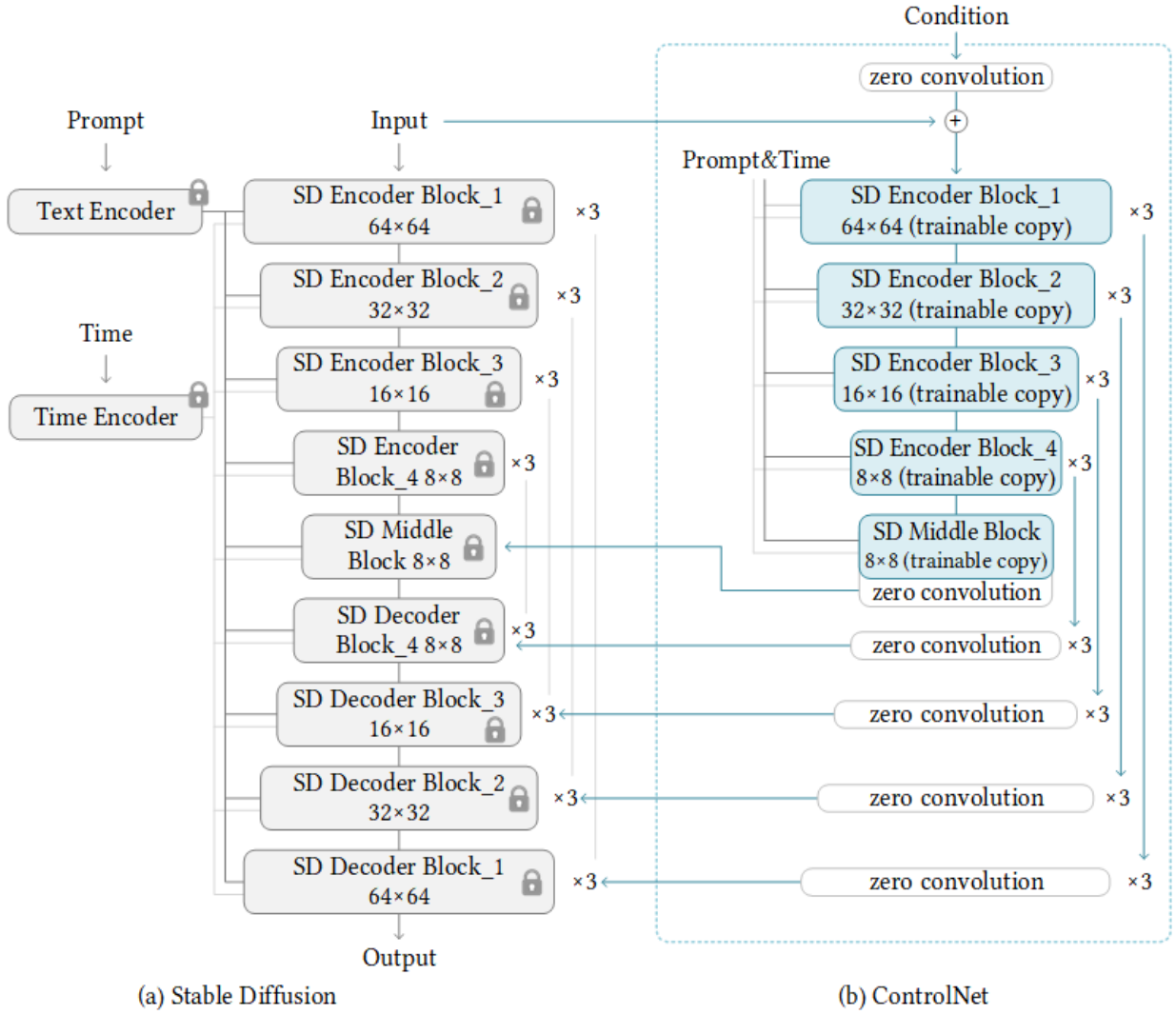
DIT4SR

HR

LR

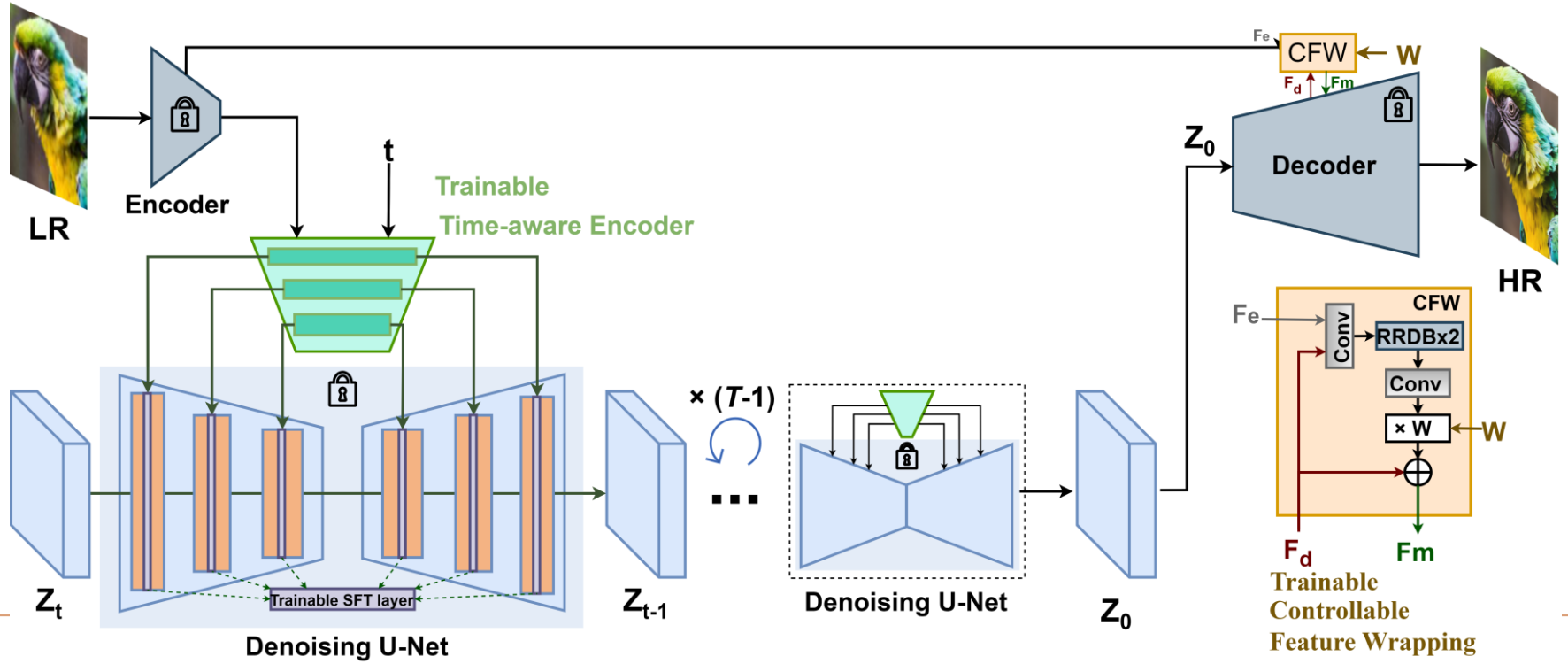
Background

- SR with diffusion
- Add LR as condition
  - SD-upsampler with additional LR channel added directly
  - ControlNet



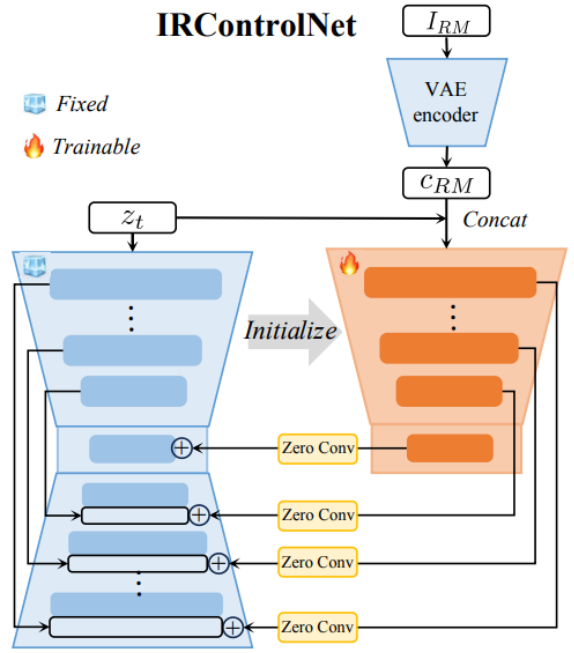
Background

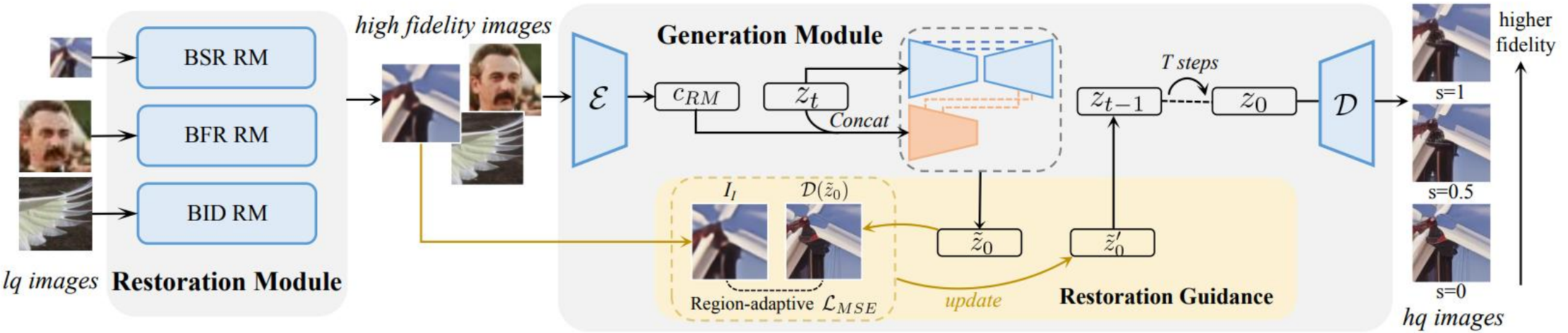
- StableSR
  - IJCV 2024



Background

- DiffBIR
  - ECCV 2024





Background

- SUPIR
  - CVPR 2024

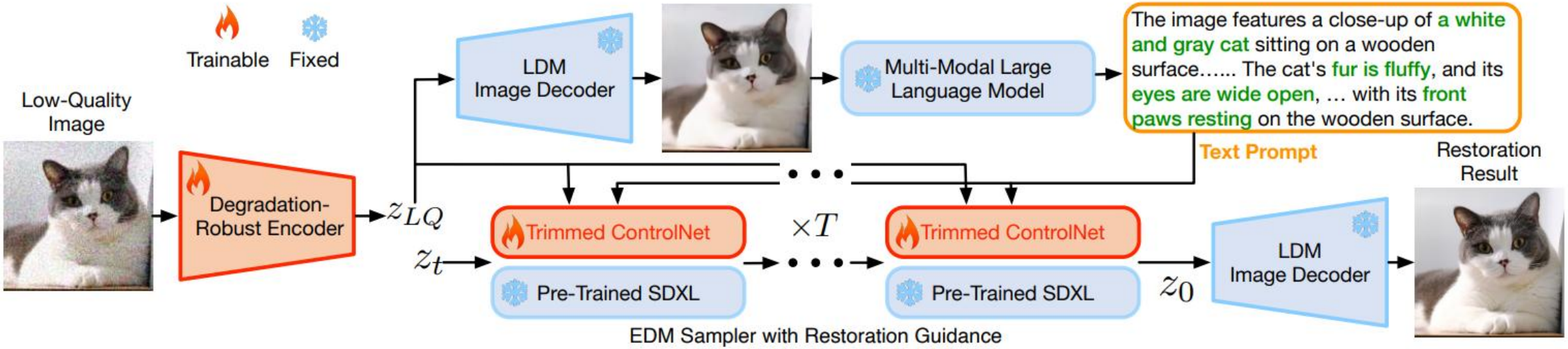
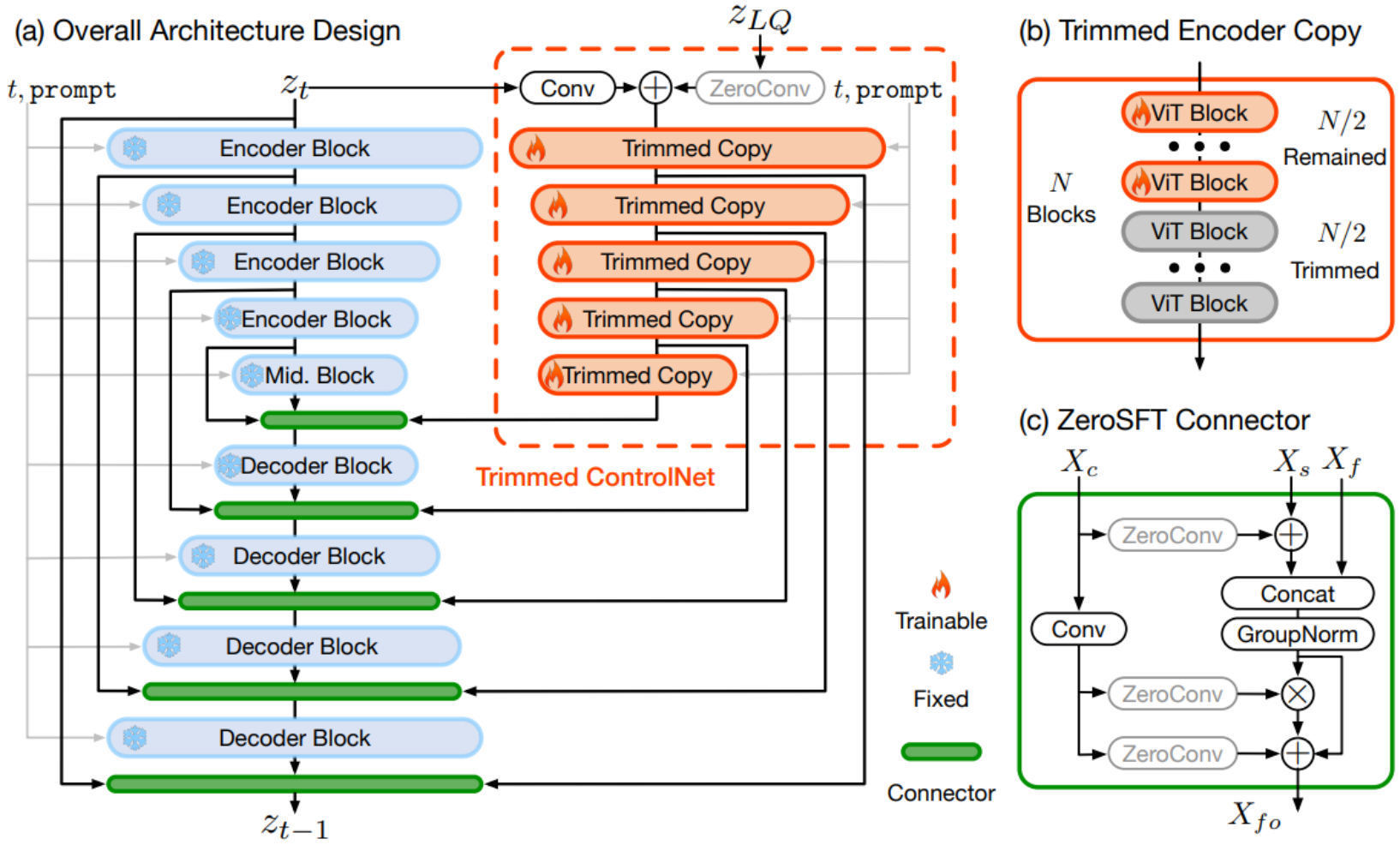


Figure 2. This figure briefly shows the workflow of the proposed SUPIR model.

Background

- SUPIR
  - CVPR 2024



Background

- SUPIR
  - CVPR 2204

LR

result

Background

- SUPIR
  - CVPR 2204

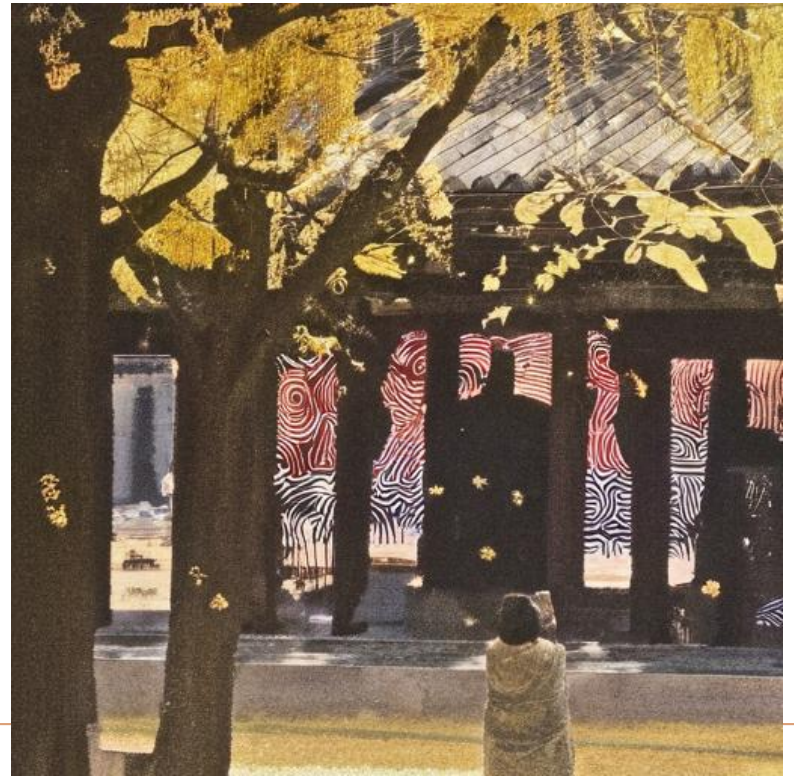
LR

result

Background

- SUPIR
  - CVPR 2204

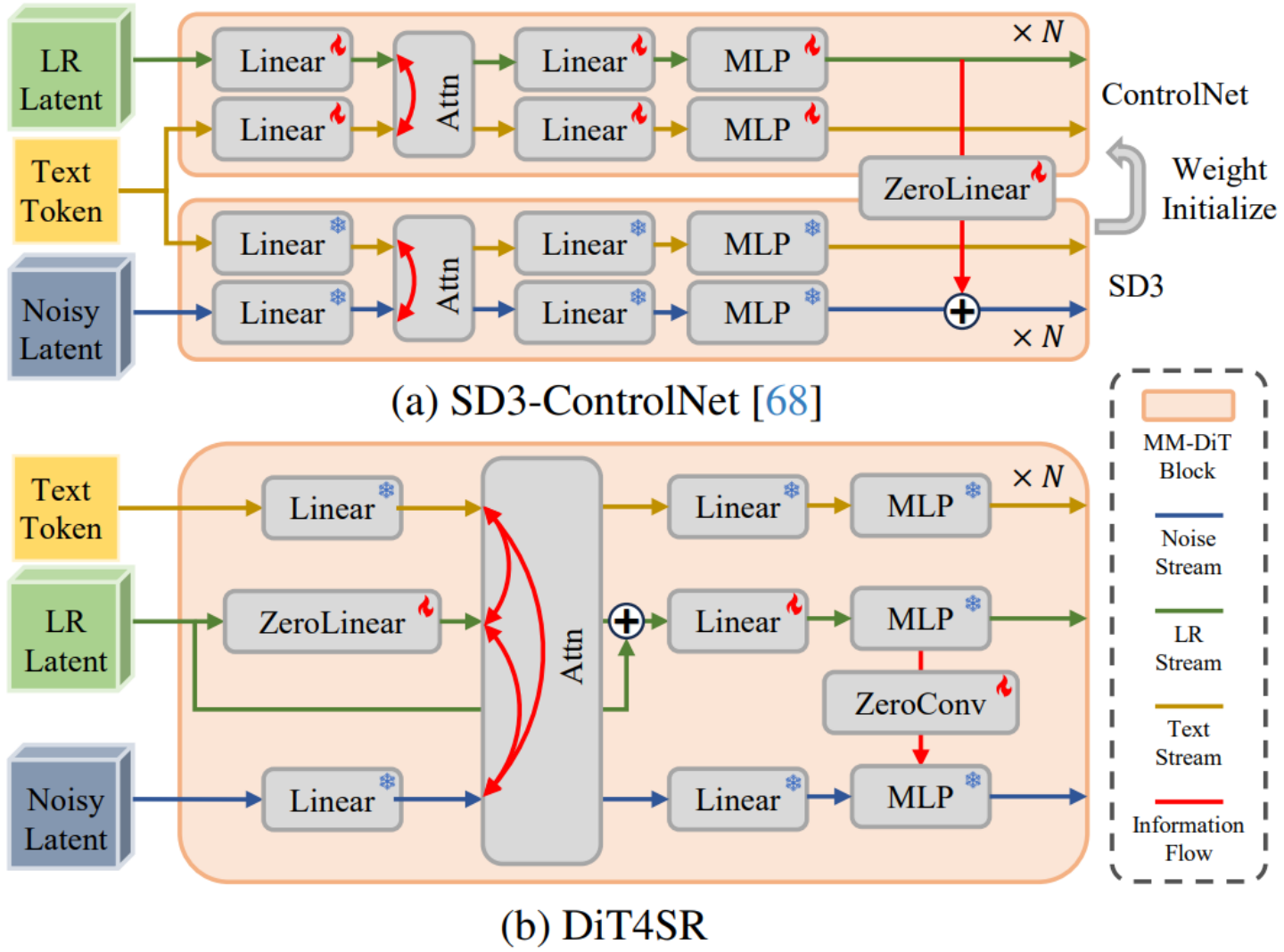
LR



result

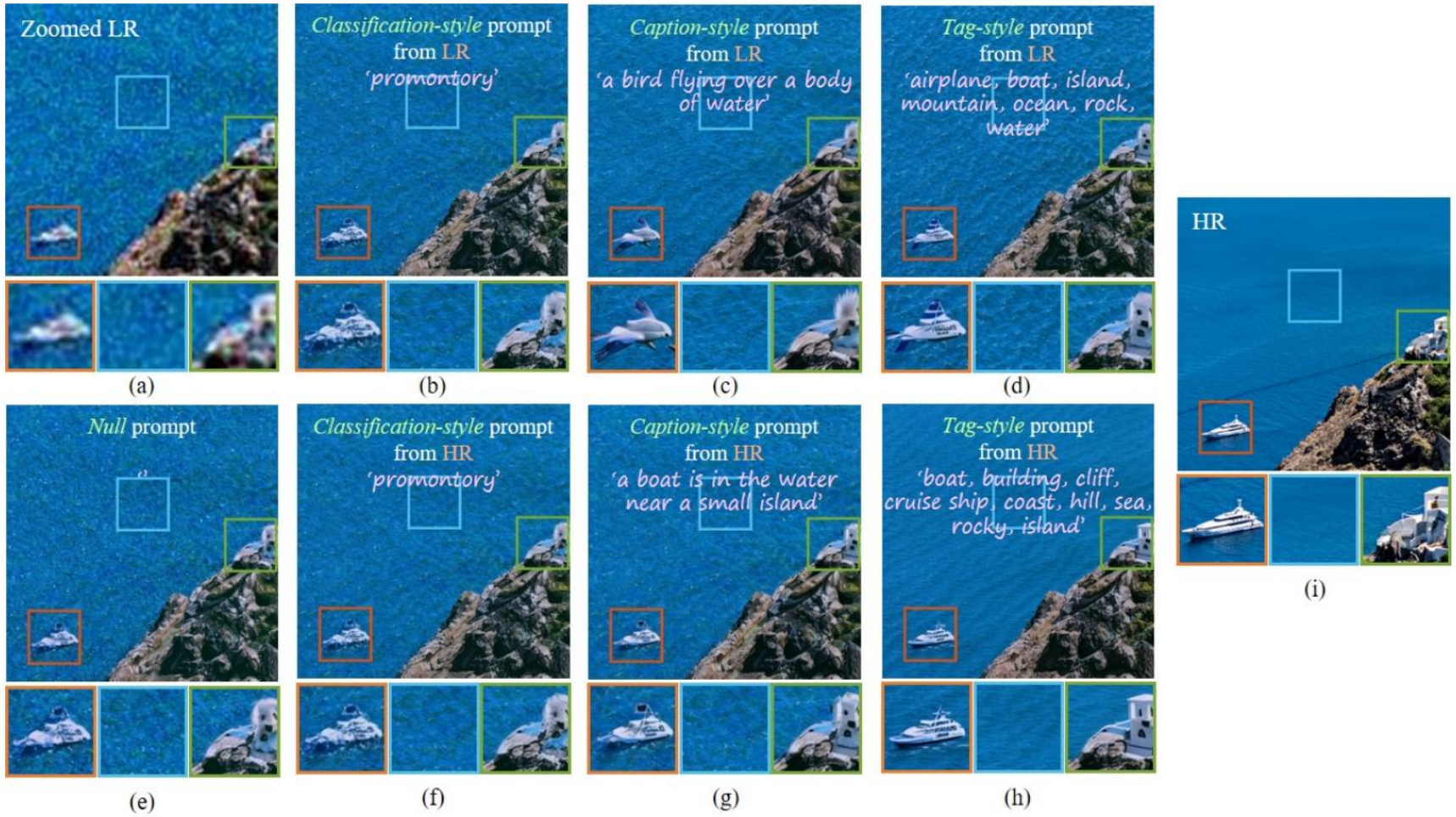
Background

- DiT4SR
  - ICCV 2025



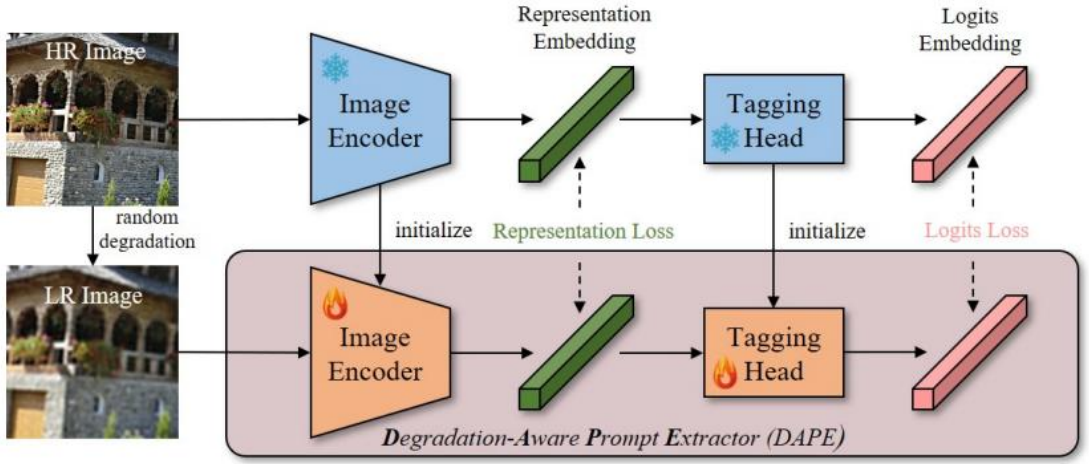
Background

- SeeSR
  - CVPR 2024

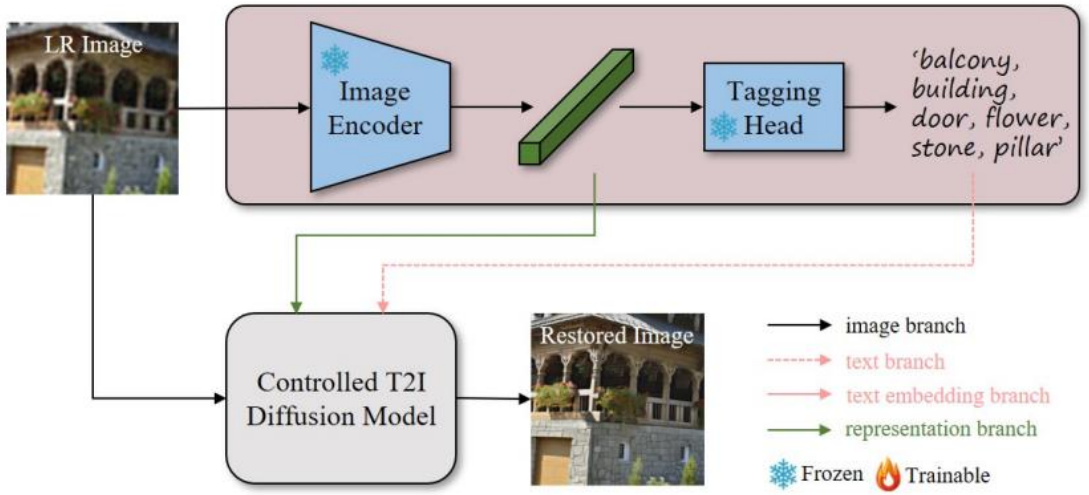


Background

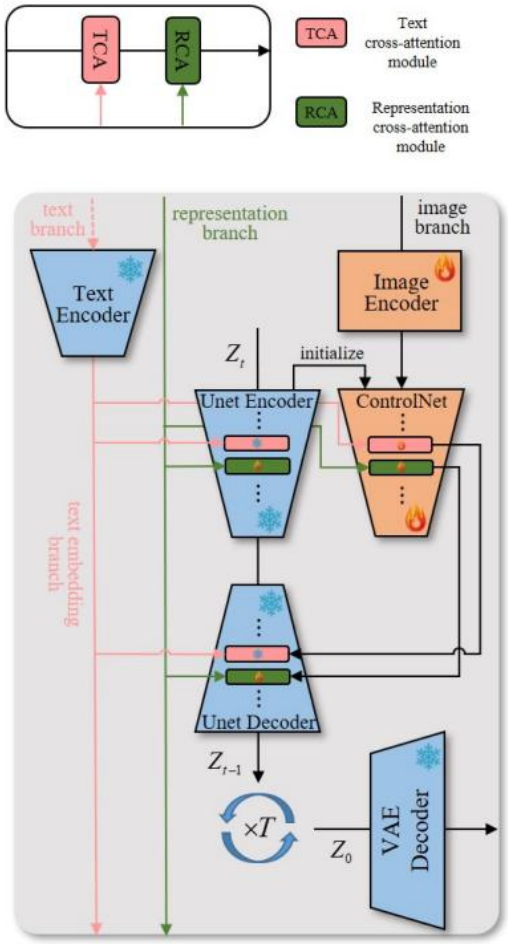
- SeeSR
  - CVPR 2024



(a) Degradation-aware prompt extractor



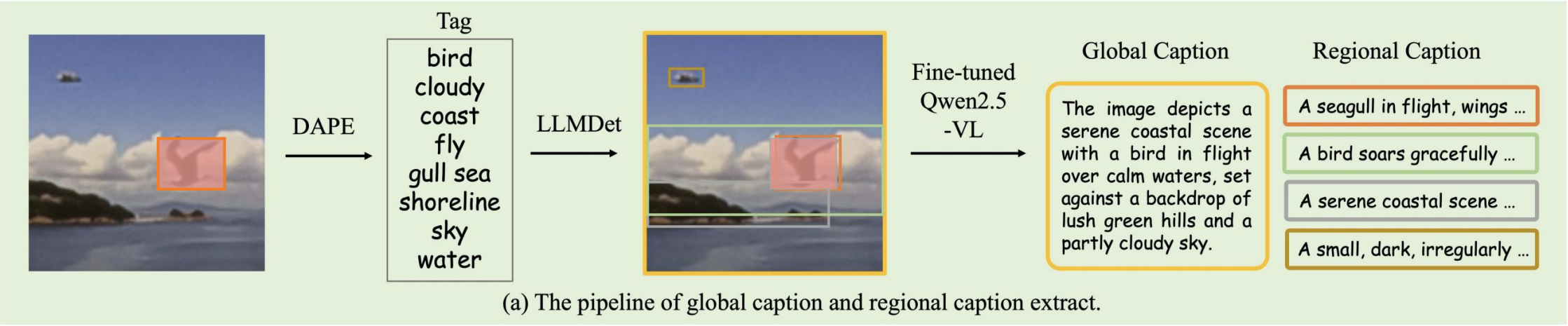
(b) Real-ISR with DAPE



(c) Controlled T2I diffusion model

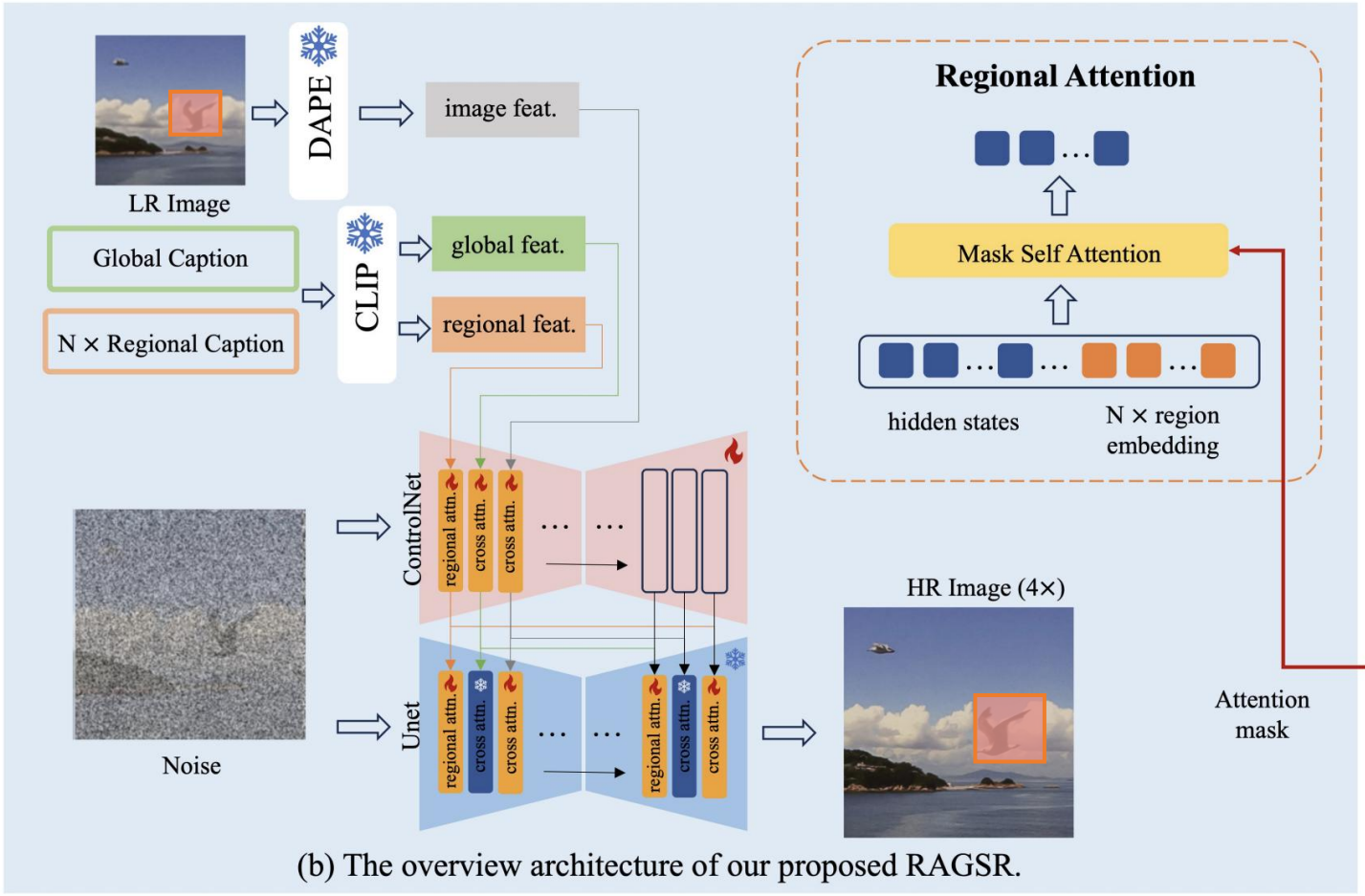
Background

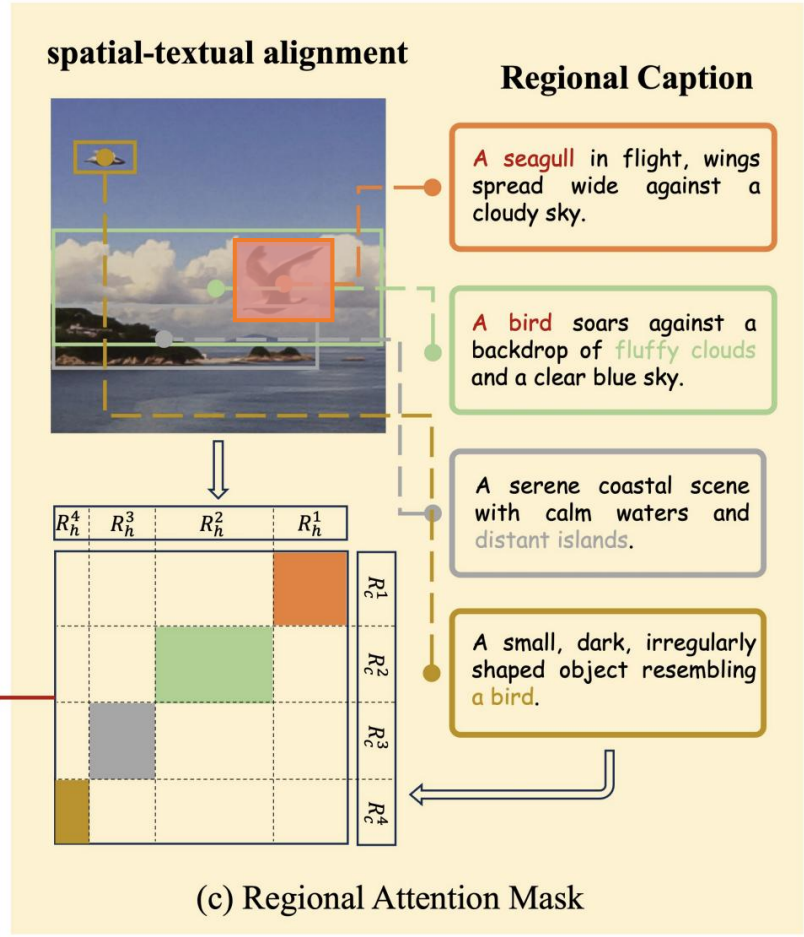
- RAGSR
  - arXiv 2025



(a) The pipeline of global caption and regional caption extract.

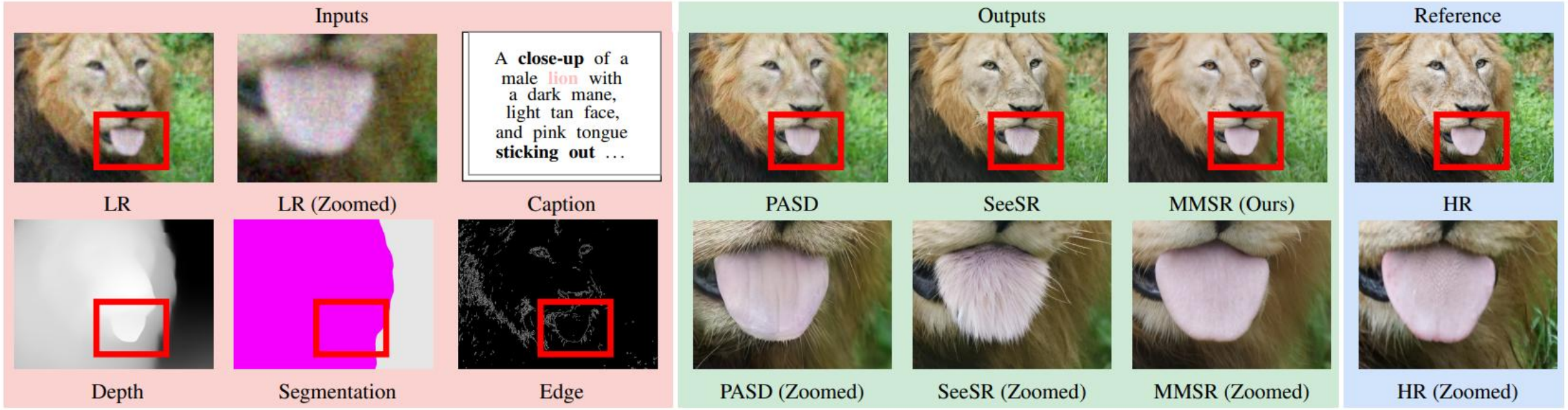
Background


LR Image
DAPE
image feat.
Global Caption
N × Regional Caption
CLIP
global feat.
regional feat.
Noise
ControlNet
Unet
regional attn.
cross attn.
HR Image (4×)
Regional Attention
Mask Self Attention
hidden states
N × region embedding
Attention mask
(b) The overview architecture of our proposed RAGSR.


spatial-textual alignment
Regional Caption
A seagull in flight, wings spread wide against a cloudy sky.
A bird soars against a backdrop of fluffy clouds and a clear blue sky.
A serene coastal scene with calm waters and distant islands.
A small, dark, irregularly shaped object resembling a bird.
(c) Regional Attention Mask

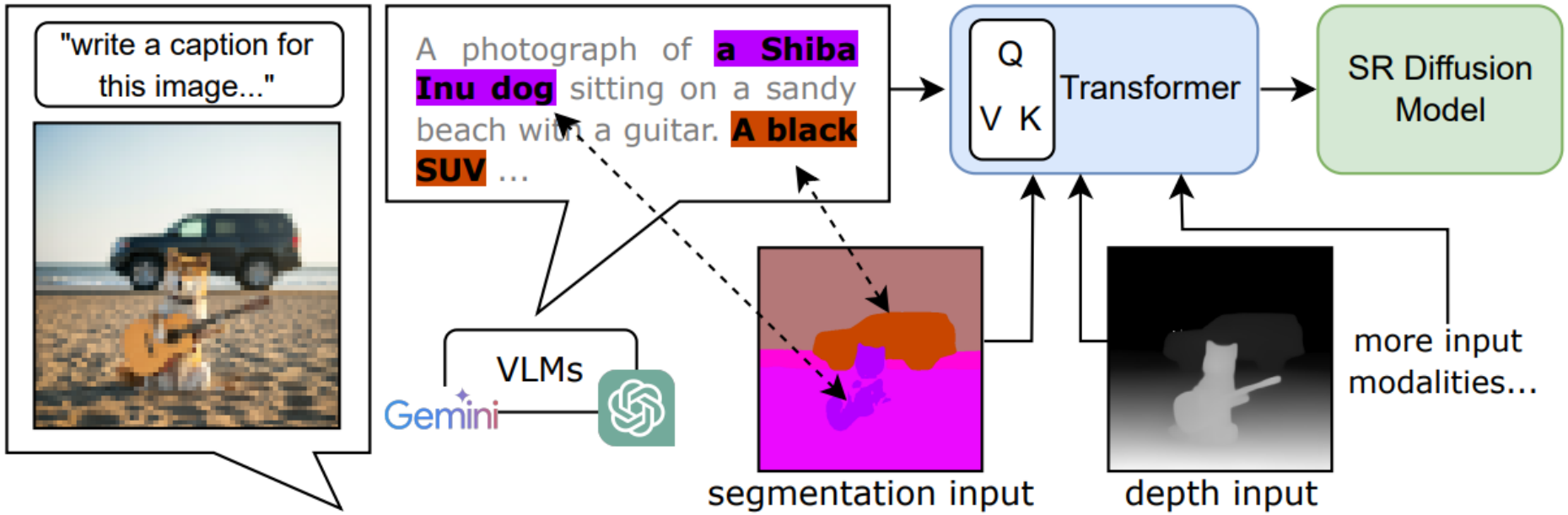
MMSR

- The Power of Context
  - CVPR 2025



MMSR

- Language models struggle to accurately represent spatial information



MMSR

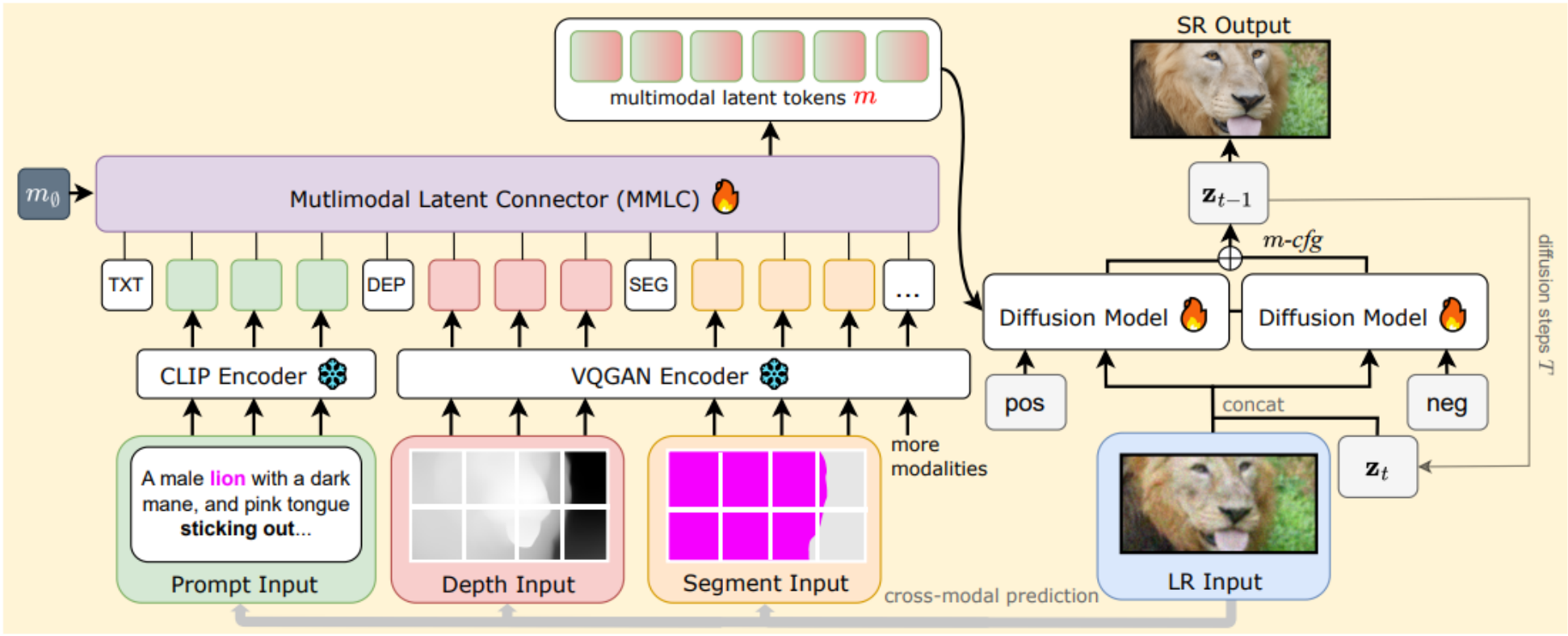
- **Problem**
  - **Spatial Ambiguity**: Text-based approaches struggle to accurately represent spatial relationships and details.
  - **Hallucinations**: Generative models can introduce false details that weren't present in the original image.
  - **Limited Controllability**: Users have minimal control over the super-resolution process.

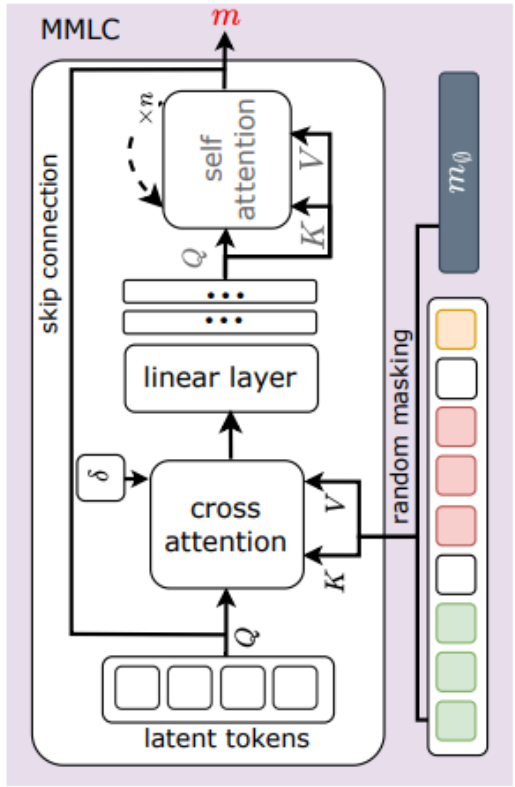
MMSR

- **Guidance types**
  - **Text captions**: Provide semantic understanding of the image content
  - **Depth maps**: Offer spatial information about object positioning and scene geometry
  - **Semantic segmentation**: Supply object boundaries and class information
  - **Edge maps**: Provide structural information about object contours

MMSR

- Framework





MMSR

- Discrete tokens for encoding each modality rather than continuous features

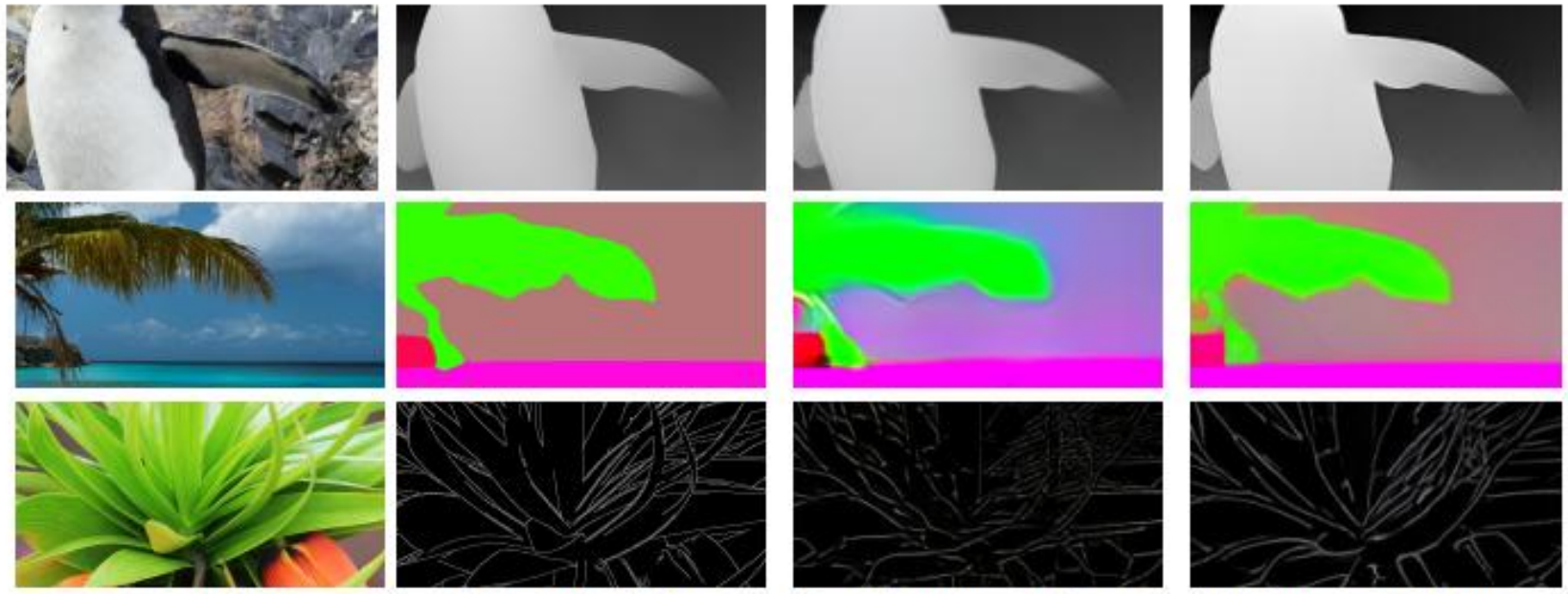


Figure 4. Using discrete multimodal tokens leads to superior reconstruction of modalities compared to continuous tokens.

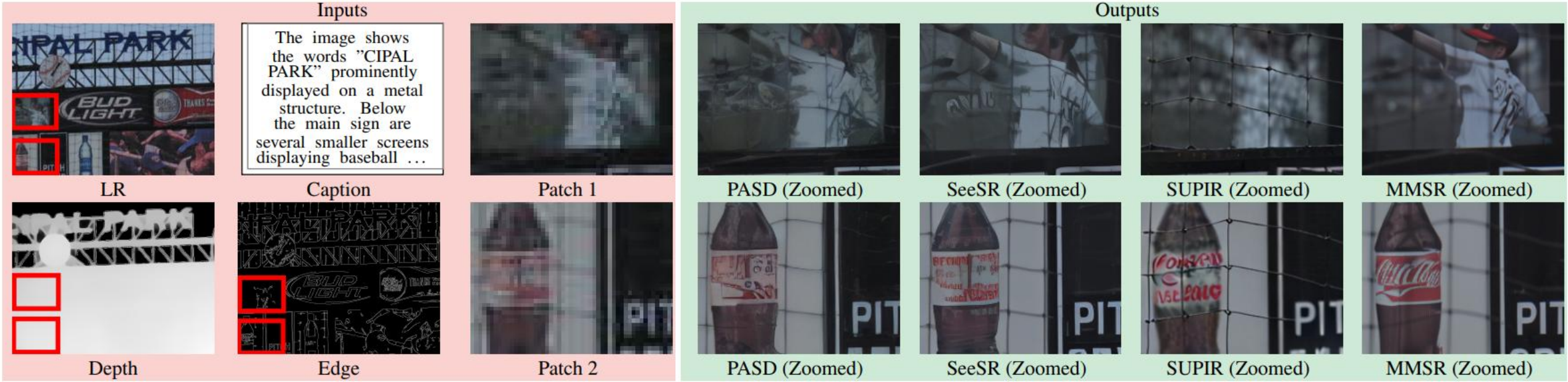
MMSR

- By scaling the temperature during sampling, achieve precise control over the super-resolution process

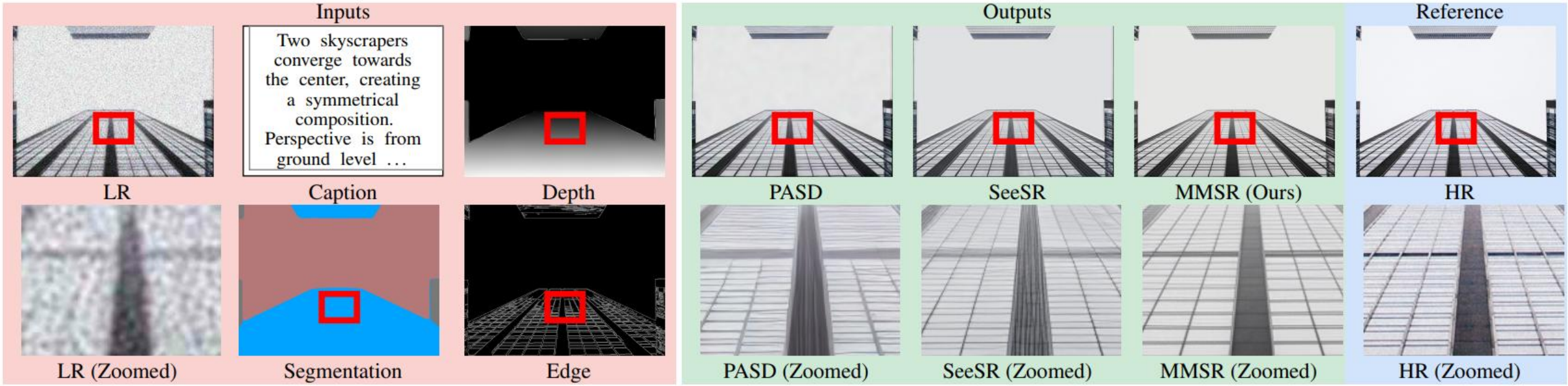
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\delta}\right)V.$$

MMSR

- Comparison Result



MMSR


Inputs
Outputs
Reference
Two skyscrapers converge towards the center, creating a symmetrical composition. Perspective is from ground level ...
LR
Caption
Depth
PASD
SeeSR
MMSR (Ours)
HR
LR (Zoomed)
Segmentation
Edge
PASD (Zoomed)
SeeSR (Zoomed)
MMSR (Zoomed)
HR (Zoomed)
A person wearing a blue hooded jacket and white pants carries a white surfboard down a grassy hill towards a dark sandy beach ...
LR
Caption
Depth
PASD
SeeSR
MMSR (Ours)
HR

MMSR

- Comparison Result

Methods	*DIV2K-Val-3k* 512 × 512 PSNR	SSIM	LPIPS ↓	DISTS ↓	NIQE ↓	FID ↓	MUSIQ	CLIPIQA
BSRGAN	21.87	0.5539	0.4136	0.2737	4.7615	64.28	59.11	0.5183
R-ESRGAN	**21.94**	**0.5736**	0.3868	0.2601	4.9209	53.46	58.64	0.5424
LDL	21.52	0.5690	0.3995	0.2688	5.0249	58.94	57.90	0.5313
DASR	21.72	0.5536	0.4266	0.2688	4.8596	67.22	54.22	0.5241
FeMASR	20.85	0.5163	0.3973	0.2428	4.5726	53.70	58.10	0.5597
LDM	21.26	0.5239	0.4154	0.2500	6.4667	41.93	56.52	0.5695
StableSR	20.84	0.4887	0.4055	0.2542	4.6551	36.57	62.95	0.6486
ResShift	21.75	0.5422	0.4284	0.2606	6.9731	55.77	58.23	0.5948
PASD	20.77	0.4958	0.4410	0.2538	4.8328	40.77	66.85	0.6799
DiffBIR	20.94	0.4938	0.4270	0.2471	4.7211	40.42	65.23	0.6664
SeeSR	21.19	0.5386	0.3843	0.2257	4.9275	31.93	68.33	0.6946
MMSR	21.74	0.5693	**0.3707**	**0.2071**	**4.2532**	**29.35**	**70.06**	**0.7164**

Methods	*DIV2K-Val-100* 1024 × 1024 PSNR	SSIM	LPIPS ↓	DISTS ↓	NIQE ↓	MANIQA	MUSIQ	CLIPIQA
R-ESRGAN	21.77	**0.5813**	0.3624	0.1990	3.6573	0.4046	47.54	0.5358
StableSR	20.07	0.3947	0.5097	0.2427	3.6260	0.4113	65.39	0.6938
PASD	21.27	0.5369	0.3473	0.1753	3.6321	0.4708	69.69	0.6914
SUPIR	20.65	0.5350	0.3849	0.1814	3.6458	0.4051	65.88	0.5697
SeeSR	21.31	0.5578	0.3273	0.1620	4.0215	**0.5439**	69.79	0.6941
MMSR	**21.87**	0.5565	**0.2810**	**0.1492**	**3.4243**	0.4885	**72.31**	**0.7294**

MMSR

- Comparison Result

Methods	*RealSR* 512 × 512 PSNR	SSIM	LPIPS ↓	LIQE	NIMA	MANIQA	MUSIQ	CLIPIQA
R-ESRGAN	**25.69**	**0.7616**	**0.2727**	3.3574	4.6548	0.5487	60.18	0.4449
StableSR	24.70	0.7085	0.3018	3.6106	4.8150	0.6221	65.78	0.6178
PASD	24.29	0.6630	0.3435	3.5749	4.8554	0.6493	68.69	0.6590
SUPIR	22.97	0.6298	0.3750	3.5682	4.5757	0.5745	61.49	0.6434
SeeSR	25.18	0.7216	0.3009	4.1360	4.9193	0.6442	69.77	0.6612
MMSR	24.83	0.7003	0.2952	**4.3468**	**5.1094**	**0.6578**	**71.33**	**0.6717**

Methods	*DrealSR* 512 × 512 PSNR	SSIM	LPIPS ↓	LIQE	NIMA	MANIQA	MUSIQ	CLIPIQA
R-ESRGAN	**28.64**	**0.8053**	**0.2847**	2.9255	4.3258	0.4907	54.18	0.4422
StableSR	26.71	0.7224	0.3284	3.2425	4.4861	0.5594	58.51	0.6357
PASD	27.00	0.7084	0.3931	3.5908	4.6618	0.5850	64.81	0.6773
SUPIR	24.61	0.6123	0.4294	3.4710	4.3815	0.5381	57.32	0.6758
SeeSR	26.75	0.7405	0.3174	4.1270	4.6942	0.6052	65.09	0.6908
MMSR	27.28	0.7456	0.3249	**4.5023**	**5.0558**	**0.6301**	**68.93**	**0.6999**

MMSR

- Ablation study

	MUSIQ	NIQE ↓	DISTS ↓	LPIPS ↓	Throughput
w/o. MMLC	69.69	3.4845	0.1781	0.3929	3.48 img/s
w. MMLC	**72.31**	**3.4243**	**0.1492**	**0.2810**	3.32 img/s

MMSR

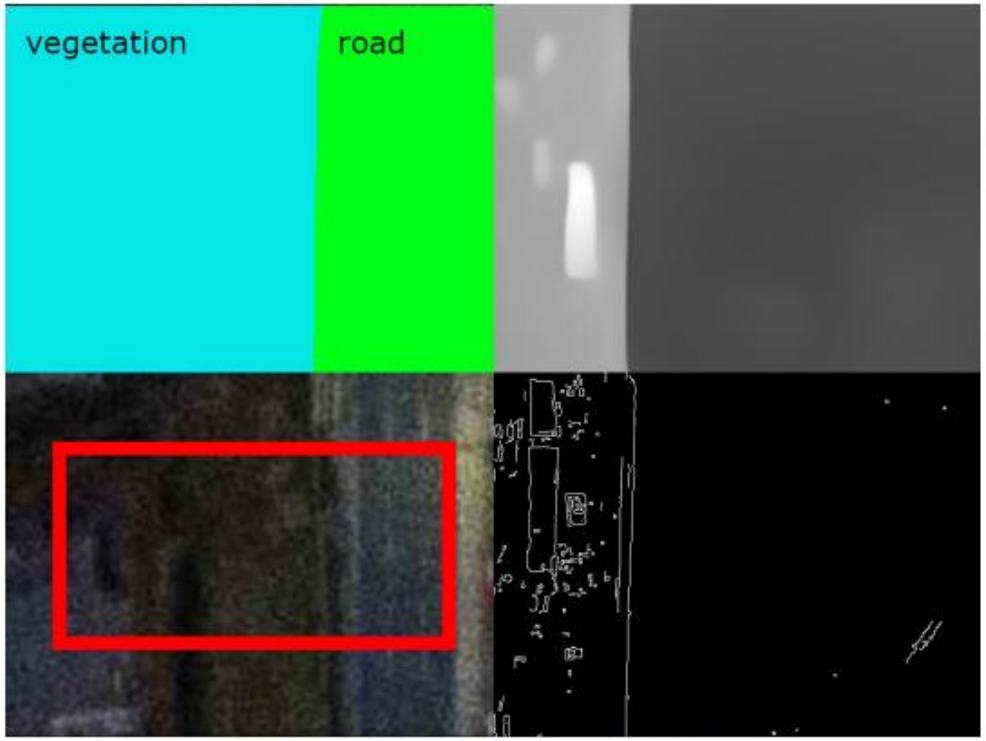
- Ablation study

guidance	2	10	14
cfg	0.3239	0.4491	0.5064
$m_{\emptyset}$-*cfg*	0.2815	0.4803	0.5493
m-cfg	**0.2810**	**0.3471**	**0.3772**

(a) LPIPS score

guidance	2	10	14
cfg	3.577	4.6179	5.1886
$m_{\emptyset}$-*cfg*	3.6261	5.1081	5.9175
m-cfg	**3.4679**	**3.7419**	**3.9815**

(b) NIQE score



Segment, Depth, LR, and Edge Input

cfg=2.0 CLIPIQA=0.4481

cfg=14.0 CLIPIQA=0.7091

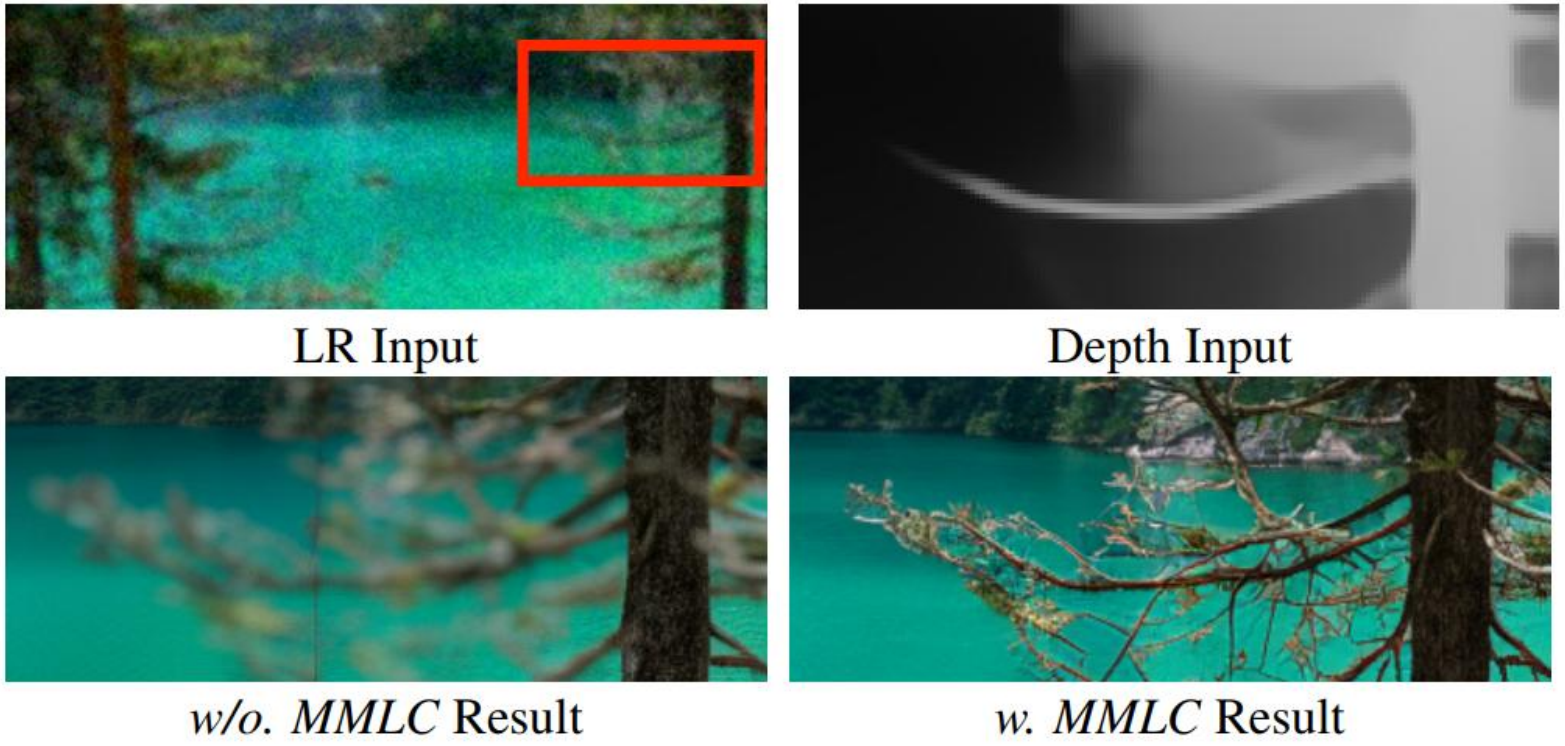
m-cfg=2.0 CLIPIQA=0.4122

m-cfg=14.0 CLIPIQA=0.7352

(c) Visual comaprisons

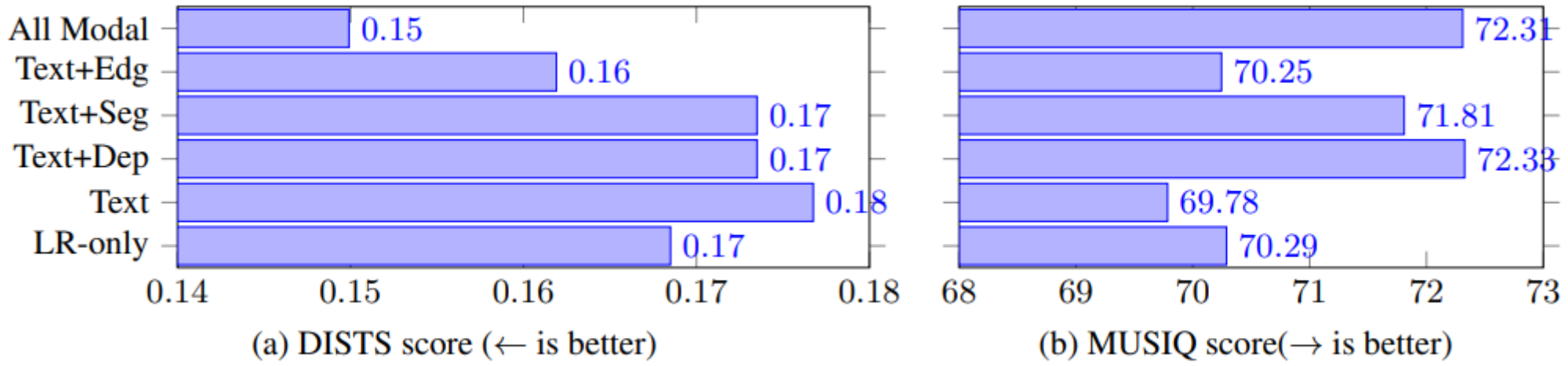
MMSR

- Ablation study



MMSR

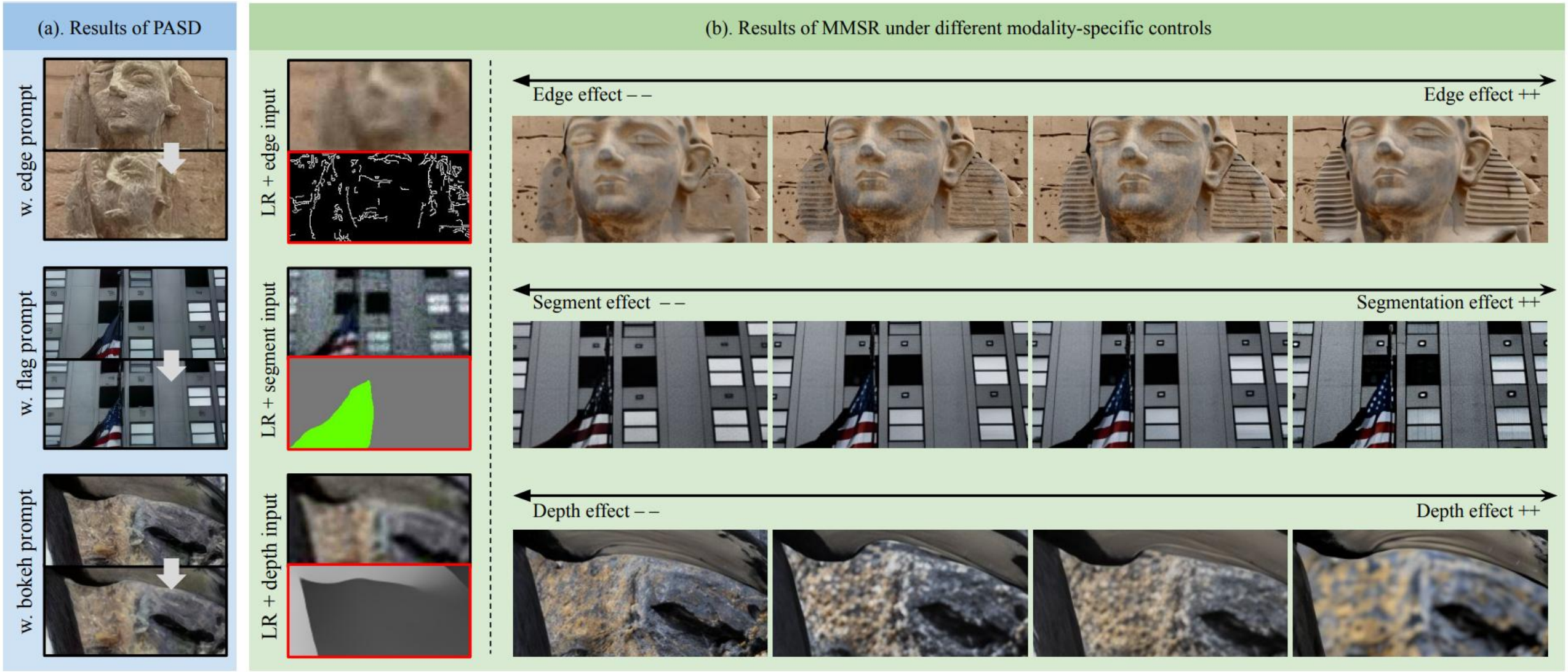
- Ablation study



(a) DISTS score (← is better)

(b) MUSIQ score(→ is better)

MMSR


(a). Results of PASD
w. edge prompt
w. flag prompt
w. bokeh prompt
(b). Results of MMSR under different modality-specific controls
LR + edge input
LR + segment input
LR + depth input
Edge effect – –
Edge effect ++
Segment effect – –
Segmentation effect ++
Depth effect – –
Depth effect ++

Conclusion

- A novel diffusion-based framework seamlessly integrates diverse modalities
- a learnable multimodal token and modality-specific contribution controls provide fine-grained control over the SR process

Thanks for your listening!