NEURAL INFORMATION PROCESSING SYSTEMS

Learning and Transferring Visual Relation with Diffusion Transformers

Yan Gong[1], Yiren Song[2], Yicheng Li[1], Chenglin Li[1], Yin Zhang[1*]

[1]Zhejiang University, [2]National University of Singapore

Presenter: Yan Gong

Learning and Transferring Visual Relation with Diffusion Transformers

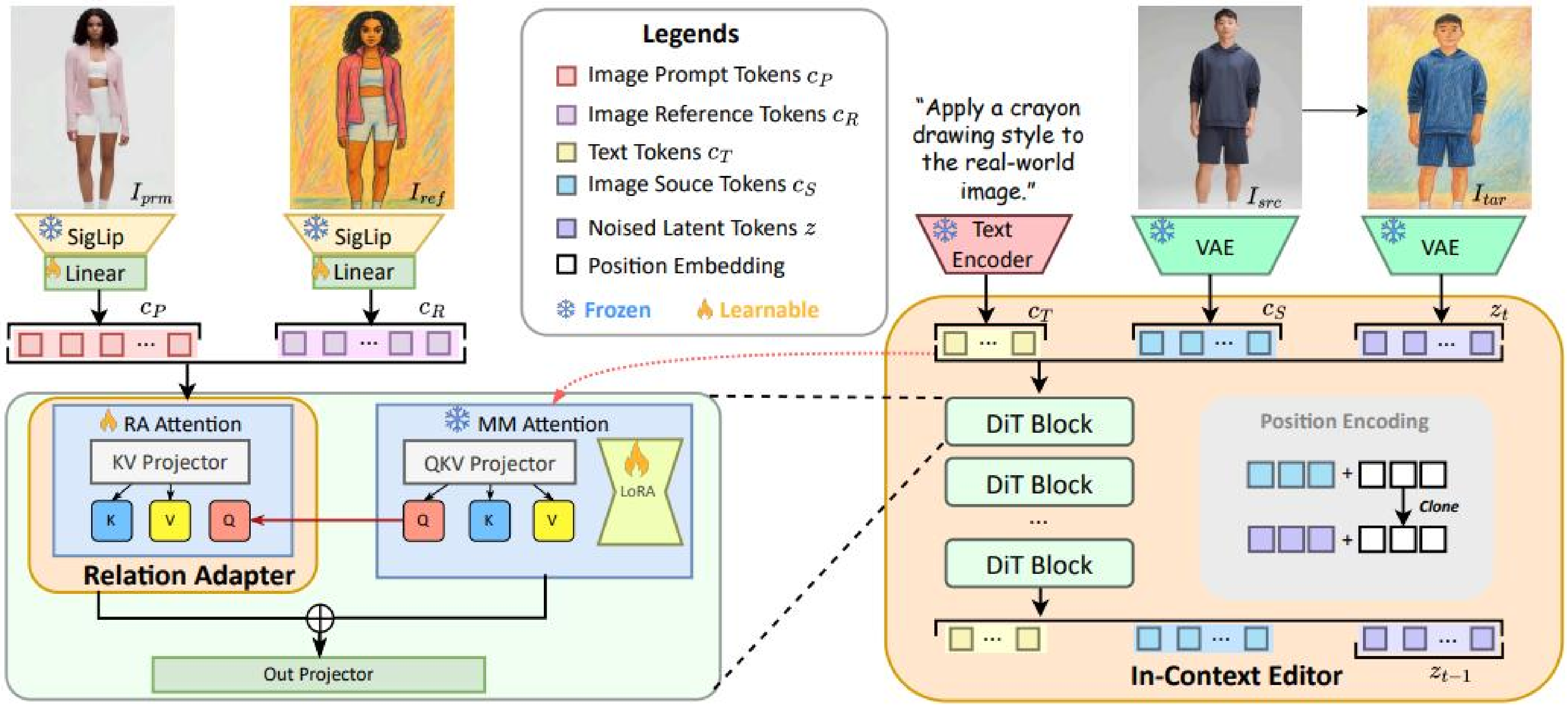
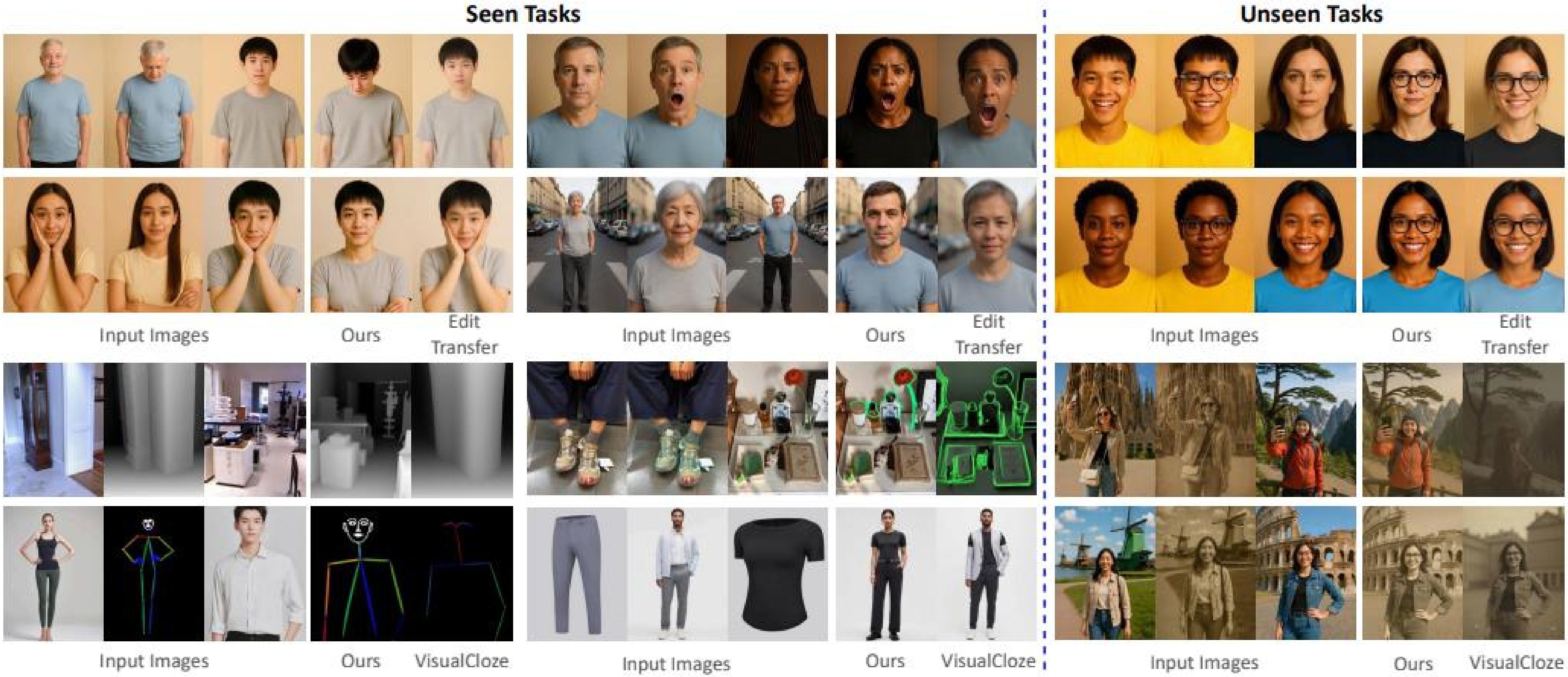


Figure 2: **The overall architecture and training paradigm of RelationAdapter.** We employ the RelationAdapter to decouple inputs by injecting visual prompt features into the MMAttention module to control the generation process. Meanwhile, a high-rank LoRA is used to train the In-Context Editor on a large-scale dataset. During inference, the In-Context Editor encodes the source image into conditional tokens, concatenates them with noise-added latent tokens, and directs the generation via the MMAttention module.

Learning and Transferring Visual Relation with Diffusion Transformers



Compared with in-context based methods, our architecture demonstrates image consistency, and editing effectiveness on both seen and unseen tasks.

Learning and Transferring Visual Relation with Diffusion Transformers

❖ Evaluation

✦ Outperforms baseline methods on common tasks, achieving lower MSE and higher CLIP-I, GPT-C, and GPT-A scores.

✦ The results are improved on both seen and unseen tasks, demonstrating its effectiveness and generalization ability in different editing scenarios.

Table 1: Quantitative Comparison of Baseline Methods Trained on a Common Task (ET: Edit Transfer, VC: VisualCloze). The best results are denoted as Bold.

Method	*MSE* ↓	*CLIP-I* ↑	*GPT-C* ↑	*GPT-A* ↑
EditTransfer	0.043	0.827	4.234	3.508
Ours ∩ ET	**0.020**	**0.905**	**4.437**	**4.429**
VisualCloze	0.049	0.802	3.594	3.411
Ours ∩ VC	**0.025**	**0.894**	**4.084**	**3.918**

Table 2: Ablation Study on the Effectiveness of the RelationAdapter(RA) in Seen and Unseen Tasks (-S for Seen, -U for Unseen). The best results are denoted as Bold.

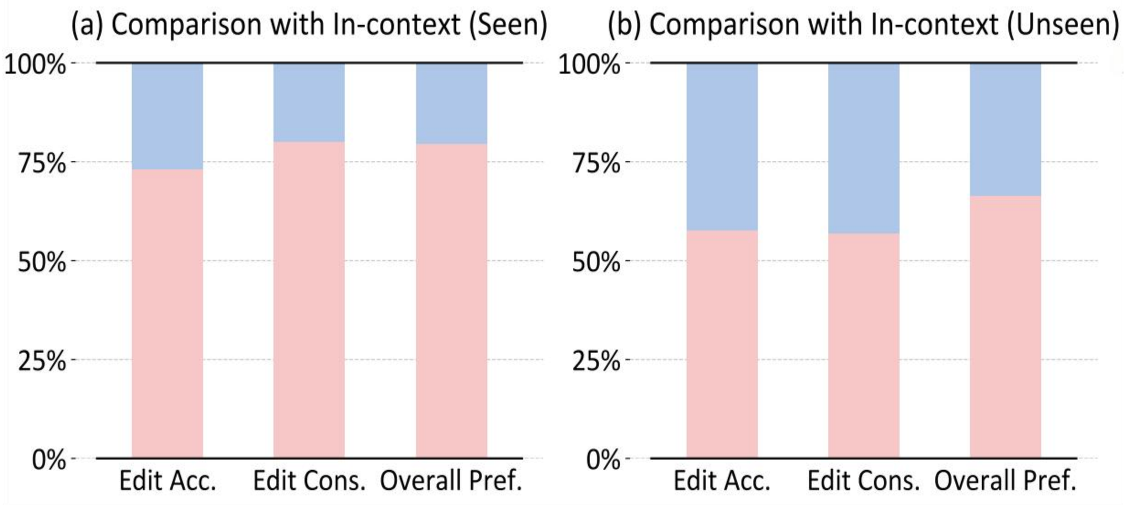
Method	*MSE* ↓	*CLIP-I* ↑	*GPT-C* ↑	*GPT-A* ↑
w/o RA -S	0.055	0.787	3.909	3.597
Ours -S	**0.044**	**0.852**	**4.079**	**4.106**
w/o RA -U	0.061	0.778	3.840	3.566
Ours -U	**0.053**	**0.812**	**4.187**	**4.173**

❖ Overall performance

✦ Efficiency improvement of our method over the In-Context baseline.

✦ The inference time is shortened (<9.0 seconds vs. 13.0+ seconds), the inference speed is increased by 30.8%, and the training speed is increased by 6.8%, with significant improvements in both training and inference efficiency.

❖ User Study



Thanks for listening!

NEURAL INFORMATION PROCESSING SYSTEMS

Learning and Transferring Visual Relation with Diffusion Transformers

Yan Gong[1], Yiren Song[2], Yicheng Li[1], Chenglin Li[1], Yin Zhang[1*]

[1]Zhejiang University, [2]National University of Singapore

Resources

Arxiv

Wechat

Snapchat

Feel free to connect!