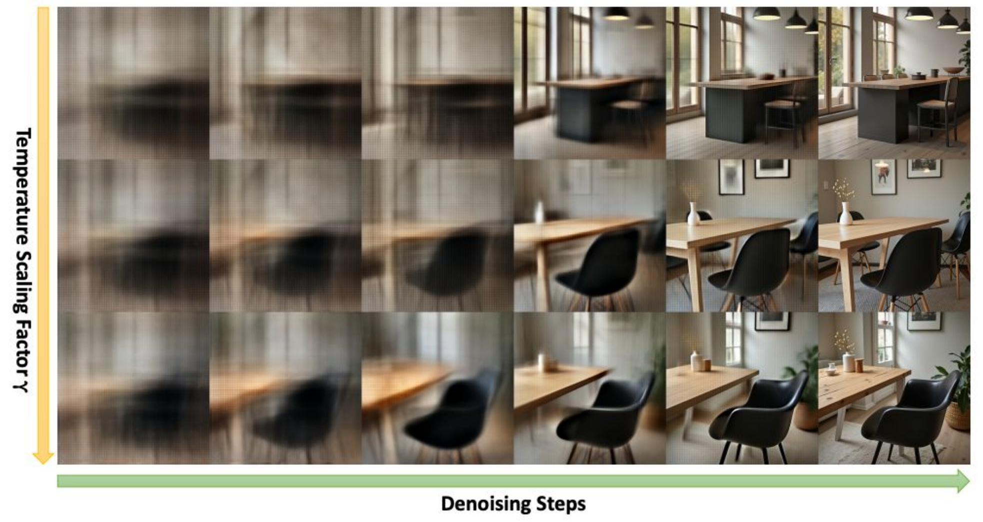
TACA: Rethinking Cross-Modal Interaction in Multimodal Diffusion Transformer

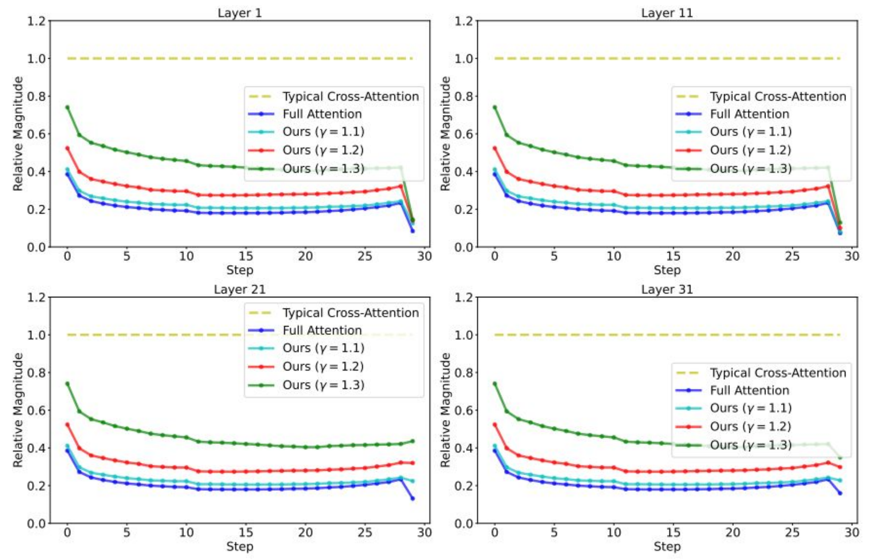
Zhengyao Lv*1, Tianlin Pan*2,3, Chenyang Si*2, Zhaoxi Chen4, Wangmeng Zuo5, Ziwei Liu4†, Kwan-Yee K. Wong1†
HKU1 NJU2 UCAS3 NTU4 HIT5

ICCV HONOLULU HAWAII OCT 19-23, 2025

Motivation

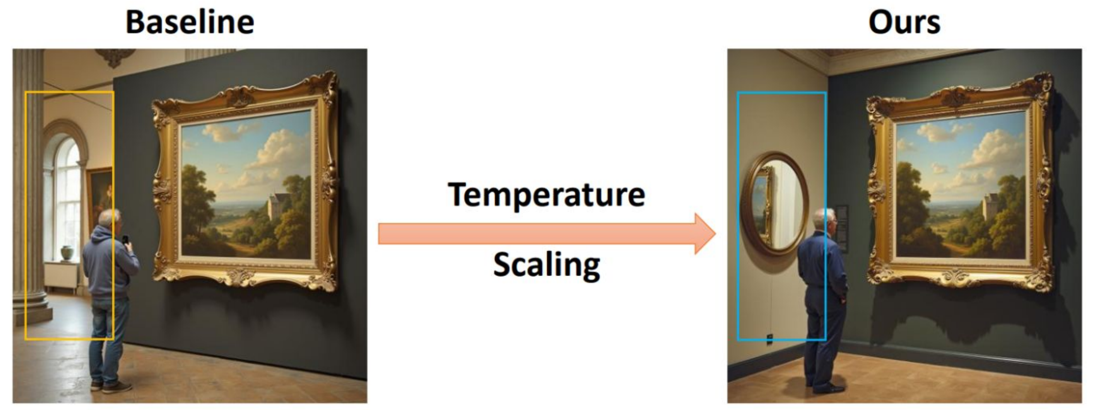
- **Task:** We propose **TACA**, a parameter-efficient method that dynamically rebalances cross-modal attention in multimodal diffusion transformers to **improve text-image alignment**.
- **Motivation:**
  - In **MM-DiT,** the cross-modal attention between visual and text tokens is suppressed **due to the significant imbalance in their numbers**
  - The attention weighting in DiT **does not adapt to the varying needs** of the denoising process across different timesteps.



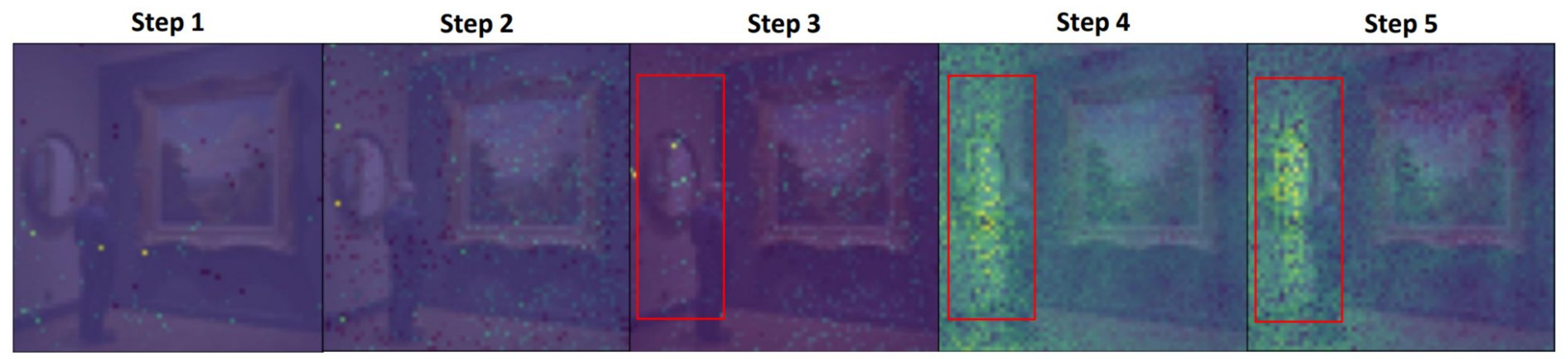


Investigation

- **Temperature scaling helps visual-text alignment.**
- To mitigate the suppression of cross-attention caused by the dominance of visual tokens (N_vis ≫ N_txt), **we amplify the logits of visual-text interactions through a temperature coefficient gamma**



Prompt: The square painting was next to the *round mirror*



Visualization of attention map difference (Ours – Baseline, First 5/30 steps)

Method

- **Temperature-Adjusted Cross-modal Attention (TACA)**
- In MM-DiT: $\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{D}}\right)\boldsymbol{V},$

where the $\boldsymbol{Q}\boldsymbol{K}^T$ term can be expanded to

$$\boldsymbol{Q}\boldsymbol{K}^T = \begin{pmatrix} \boldsymbol{W}_c^Q\boldsymbol{c}(\boldsymbol{W}_c^K\boldsymbol{c})^T & \boldsymbol{W}_c^Q\boldsymbol{c}(\boldsymbol{W}_x^K\boldsymbol{x})^T \\ \boldsymbol{W}_x^Q\boldsymbol{x}(\boldsymbol{W}_c^K\boldsymbol{c})^T & \boldsymbol{W}_x^Q\boldsymbol{x}(\boldsymbol{W}_x^K\boldsymbol{x})^T \end{pmatrix} = \begin{pmatrix} \boldsymbol{Q}_{\mathrm{txt}}\boldsymbol{K}_{\mathrm{txt}}^T & \boldsymbol{Q}_{\mathrm{txt}}\boldsymbol{K}_{\mathrm{vis}}^T \\ \boxed{\boldsymbol{Q}_{\mathrm{vis}}\boldsymbol{K}_{\mathrm{txt}}^T} & \boldsymbol{Q}_{\mathrm{vis}}\boldsymbol{K}_{\mathrm{vis}}^T \end{pmatrix}.$$

- After Softmax:

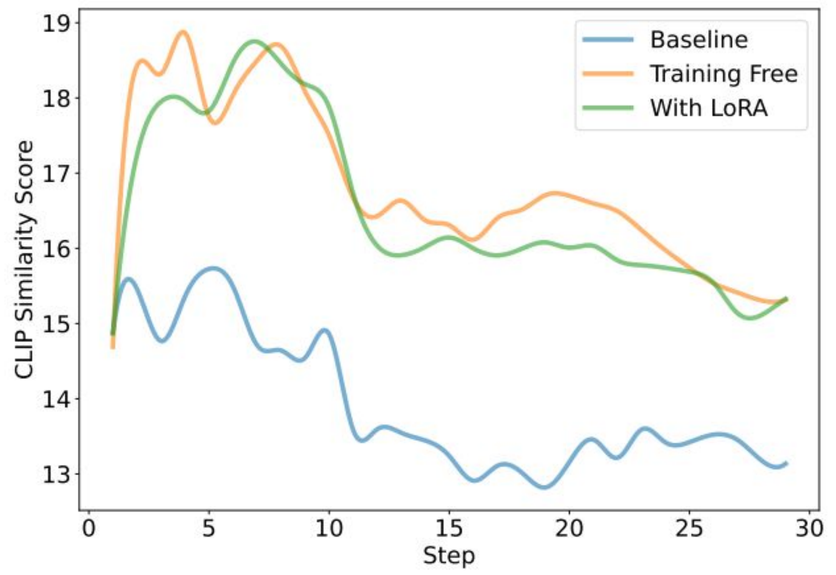
$$P_{\mathrm{vis-txt}}^{(i,j)} = \frac{e^{\gamma s_{ij}^{\mathrm{vt}}/\tau}}{\sum_{k=1}^{N_{\mathrm{txt}}} e^{\gamma s_{ik}^{\mathrm{vt}}/\tau} + \sum_{k=1}^{N_{\mathrm{vis}}} e^{s_{ik}^{\mathrm{vv}}/\tau}},$$

where $s_{ik}^{\mathrm{vt}} = \boldsymbol{Q}_{\mathrm{vis}}^{(i)}\boldsymbol{K}_{\mathrm{txt}}^{T\,(k)}/\sqrt{D}$ and $s_{ik}^{\mathrm{vv}} = \boldsymbol{Q}_{\mathrm{vis}}^{(i)}\boldsymbol{K}_{\mathrm{vis}}^{T\,(k)}/\sqrt{D}$

- Only enhance the early timesteps:

$$\gamma(t) = \begin{cases} \gamma_0 & t \geq t_{\mathrm{thresh}} \\ 1 & t < t_{\mathrm{thresh}} \end{cases}$$

- We trained a LoRA to mitigate artifacts caused by the shift in the output distribution of each attention layer



Temperature Scaling

a brown backpack and a blue horse

a fabric hat and ***a glass mirror***

A white cat with ***black stomach*** takes a pose

Results

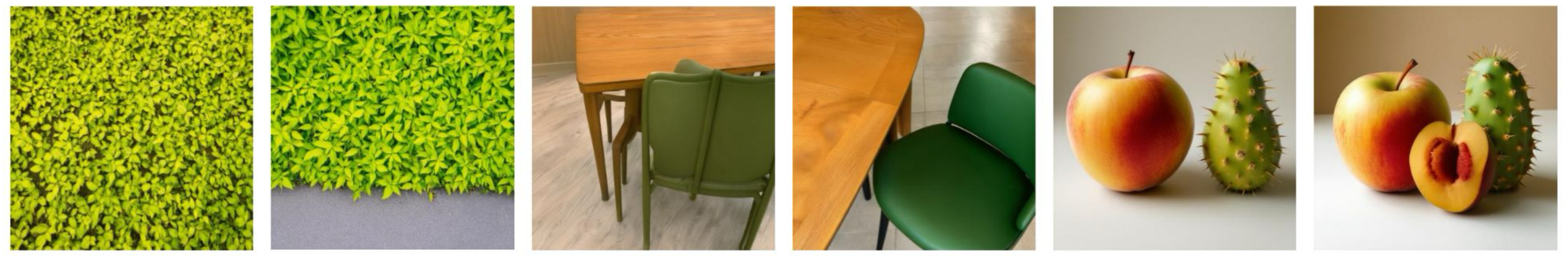
Model	Attribute Binding			Object Relationship		Complex ↑
	Color ↑	Shape ↑	Texture ↑	Spatial ↑	Non-Spatial ↑	
FLUX.1-Dev	0.7678	0.5064	0.6756	0.2066	0.3035	0.4359
FLUX.1-Dev + TACA ($r = 64$)	**0.7843**	**0.5362**	**0.6872**	**0.2405**	0.3041	**0.4494**
FLUX.1-Dev + TACA ($r = 16$)	0.7842	0.5347	0.6814	0.2321	**0.3046**	0.4479
SD3.5-Medium	0.7890	0.5770	0.7328	0.2087	0.3104	0.4441
SD3.5-Medium + TACA ($r = 64$)	**0.8074**	**0.5938**	**0.7522**	**0.2678**	0.3106	0.4470
SD3.5-Medium + TACA ($r = 16$)	0.7984	0.5834	0.7467	0.2374	**0.3111**	**0.4505**

FLUX.1 Dev | Ours | FLUX.1 Dev | Ours | FLUX.1 Dev | Ours

The sweet ***banana*** split topped with the crunchy ***nuts*** and the soft whipped cream — The red vase was ***on top of*** the wooden shelf — The ***juicy*** watermelon quenched the ***thirsty mouth*** and the dry throat

The fluffy white clouds floated ***above the tall skyscrapers*** — The crisp green lettuce lay next to the juicy red ***tomato*** — The crisp apple lay beside the ***rough stone*** and the silky fabric

SD3.5-Medium | Ours | SD3.5-Medium | Ours | SD3.5-Medium | Ours



The bright green leaves ***contrasted with the dull grey pavement*** — The green chair is ***on the right of*** the wooden table — The glossy apple sat next to the ***fuzzy peach*** and the prickly pear

The flickering ***fireflies*** danced in the dark forest and the ***moonlit*** meadow — The green plant was ***on the left of the white lamp*** — The ***blue pencil*** was next to the white paper

Project Page — GitHub Repo — Paper