THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Snap Research

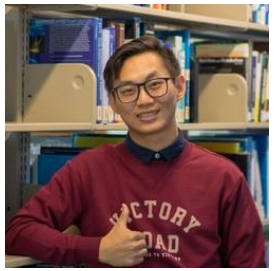
ICML International Conference On Machine Learning

I Think, Therefore I Diffuse: Enabling Multimodal In-Context Reasoning in Diffusion Models

ICML2025


Zhenxing Mi


Kuan-Chieh Wang

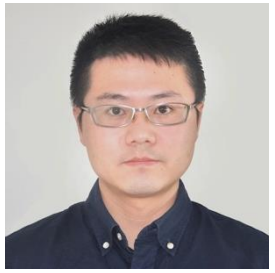

Guocheng Qian


Hanrong Ye

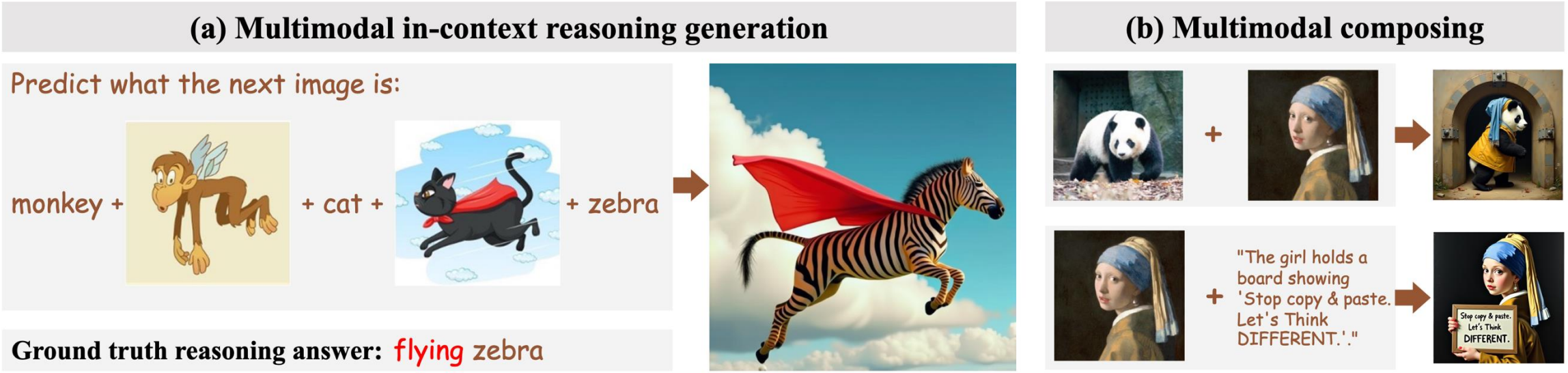

Runtao Liu


Sergey Tulyakov


Kfir Aberman


Dan Xu

Let’s make diffusion models think before generating

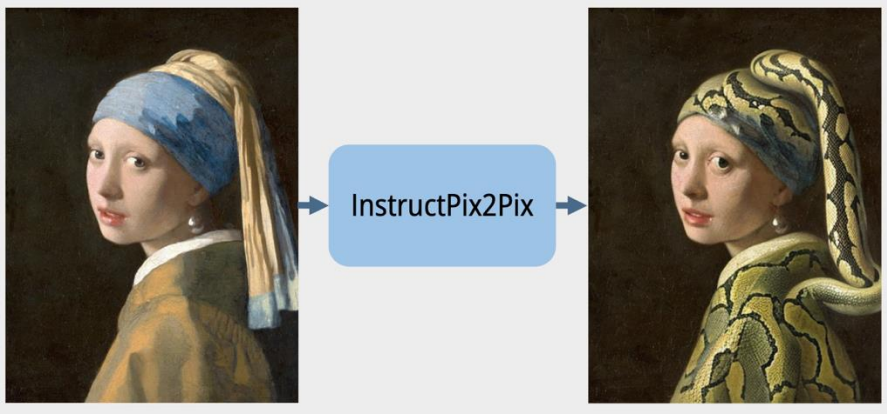


With large vision-language model (LVLM) + Diffusion decoder,
we get a model with strong multimodal understanding and generation capabilities.

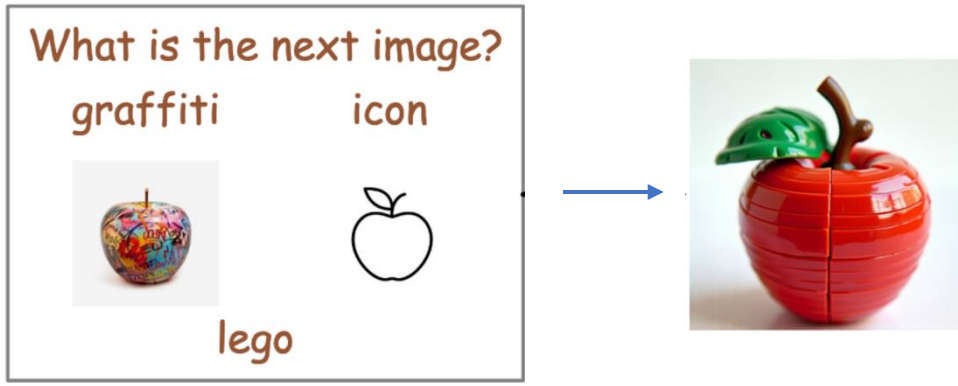
We align a VLM (Qwen2-VL) and a diffusion decoder (Flux).

Problem 1: Dataset

Direct diffusion training needs complex multimodal reasoning datasets.



Editing image pairs

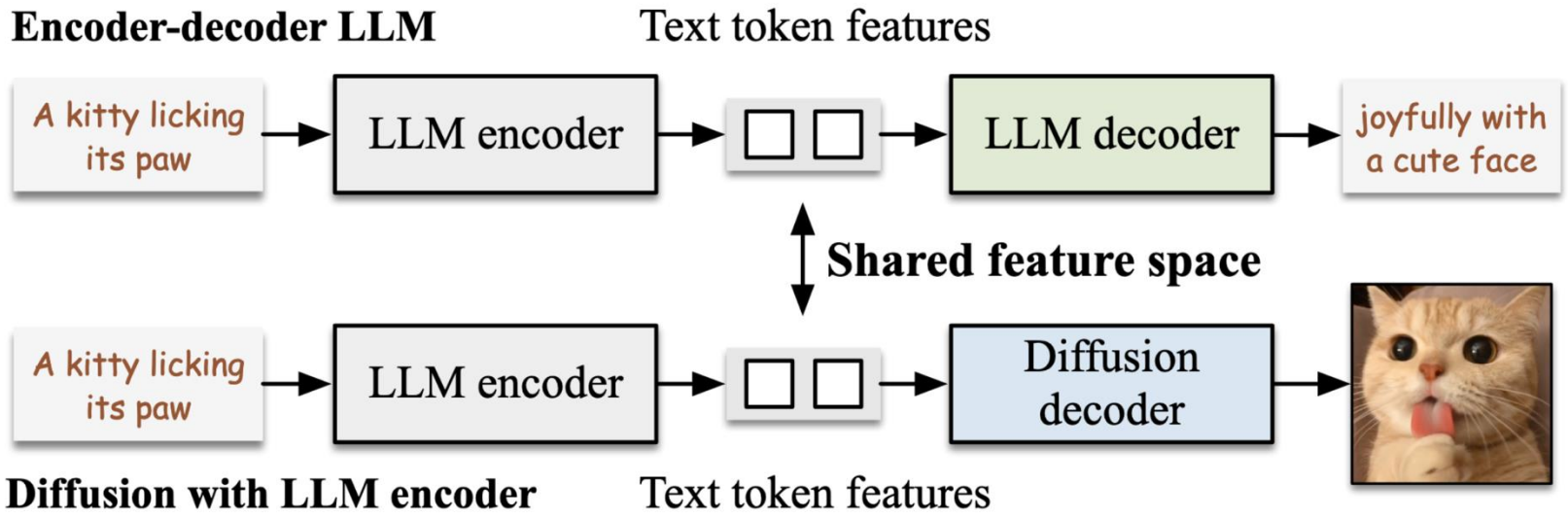


Multimodal reasoning samples

The reasoning capabilities will mainly come from the diffusion training, instead of inheriting from the VLM's existing capabilities.

Insight 1: Shared feature space

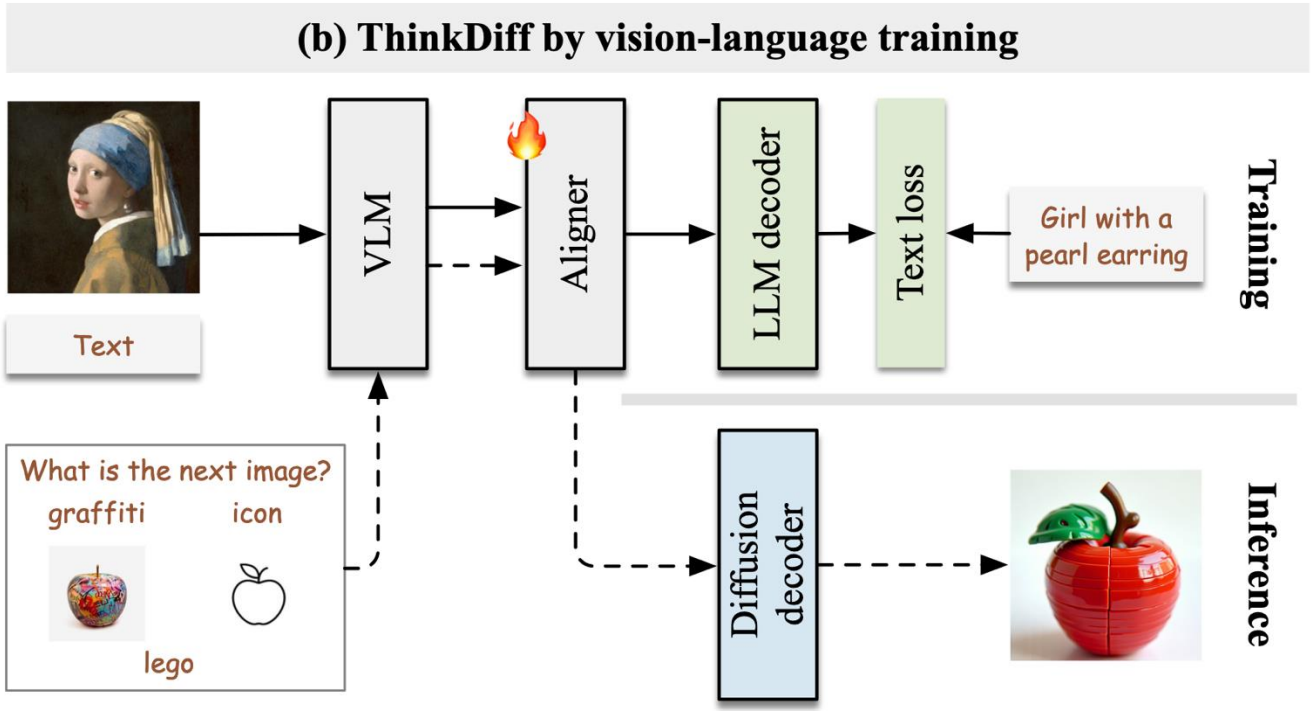
A text-to-image diffusion model uses a LLM **encoder** (T5) for text encoding.
It's diffusion **decoder** shares the same input space with the LLM **decoder**.



We can align the VLM with the LLM decoder (T5) by vision-language training.
Then the VLM is aligned with the diffusion decoder.

Training and inference

Training: align the VLM to the LLM (T5) decoder by vision-language training.
Inference: replace the LLM (T5) decoder as the diffusion decoder.

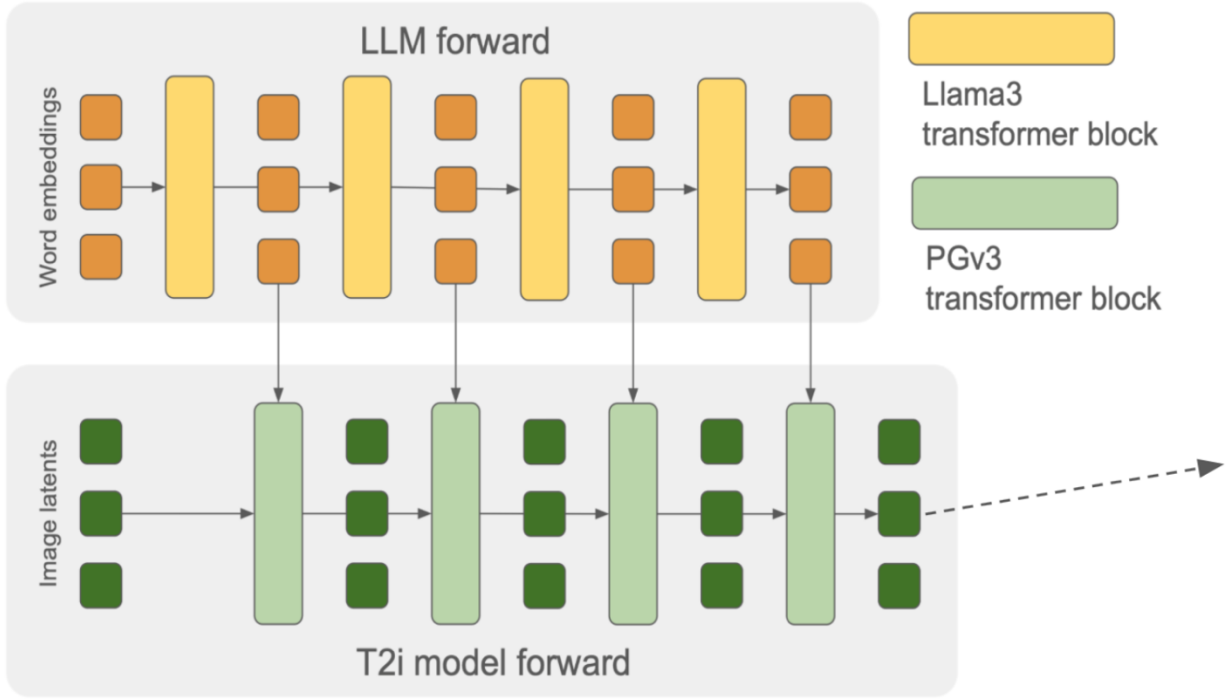


We only need image-text pairs for training.
But we support interleaved image and text as input in inference.

Problem 2: Transferring LVLM's reasoning capabilities.

How to make the diffusion decoder fully inherent VLM's capabilities?

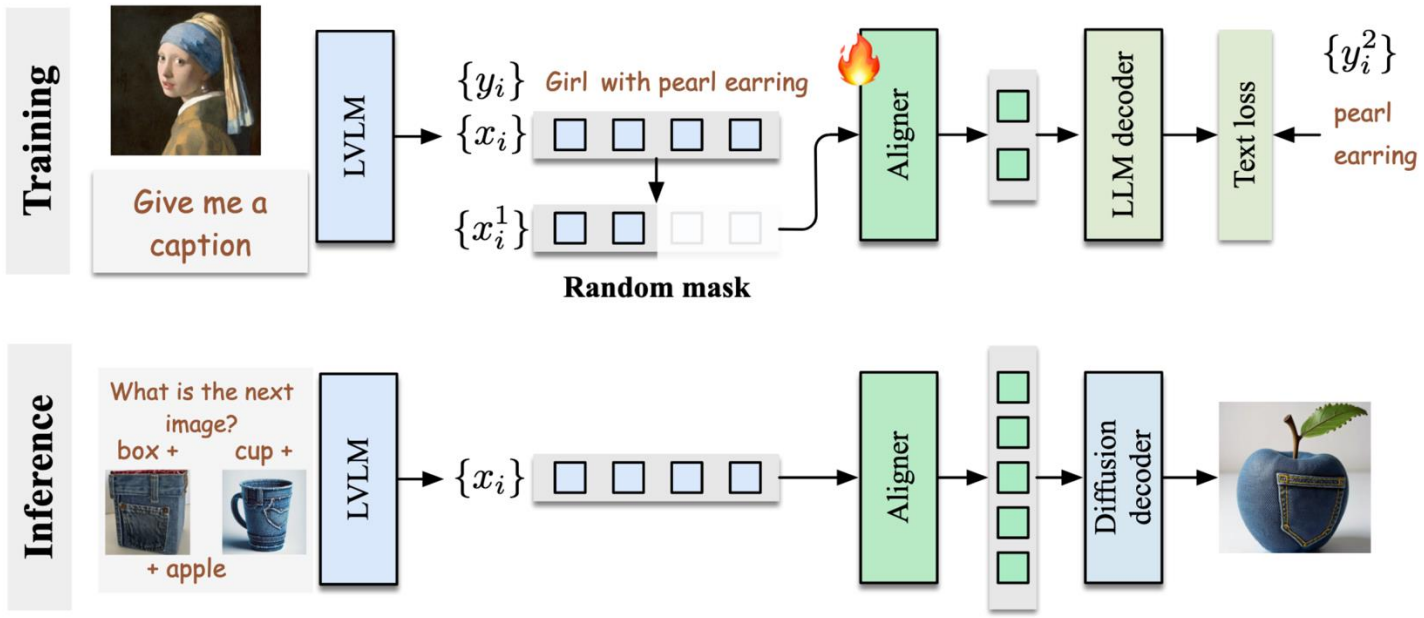
Playground V3



Only using the deep features of **input** tokens does not fully transfer the reasoning information.

Insight 2: generating is reasoning

The deep features of **generated** tokens fully capture the reasoning process and results.

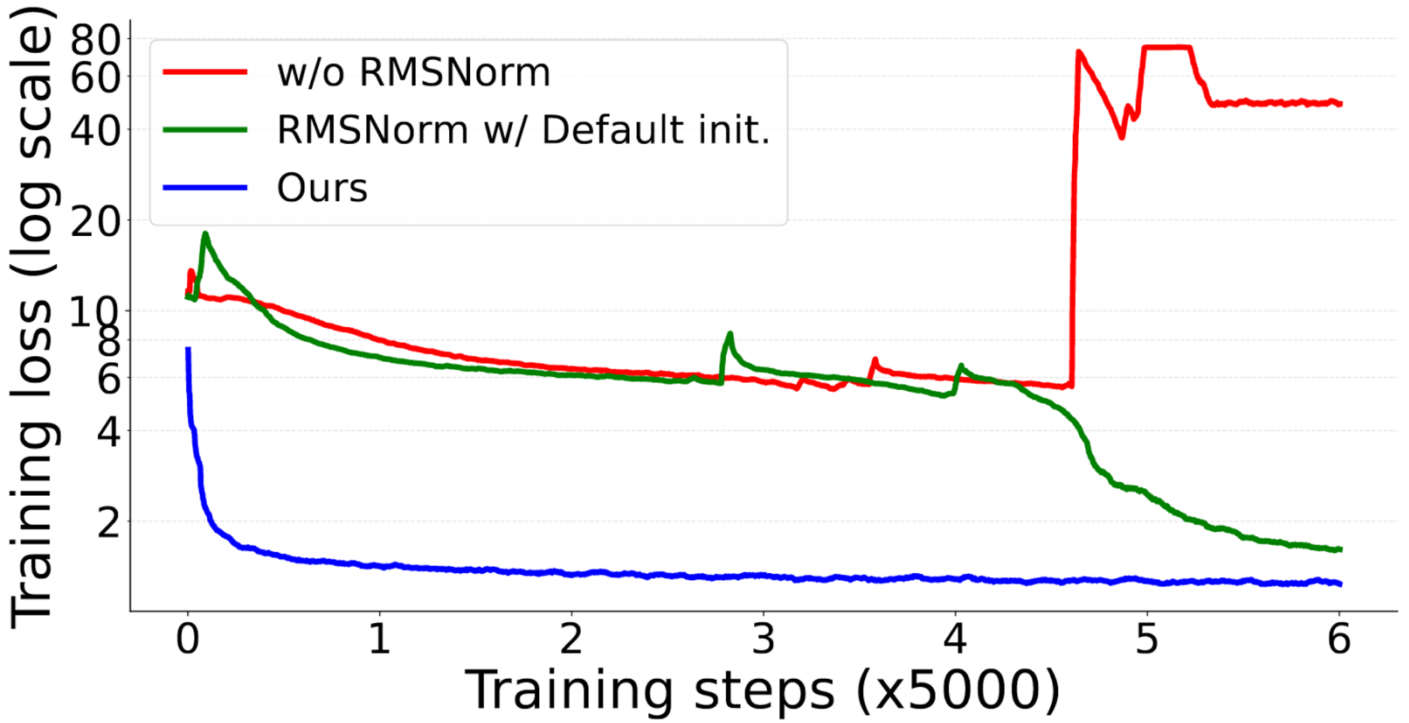


(a) ThinkDiff-LVLM

By aligning the deep features of VLM's generated tokens to the diffusion decoder, we successfully transfers the reasoning capabilities to the diffusion models.

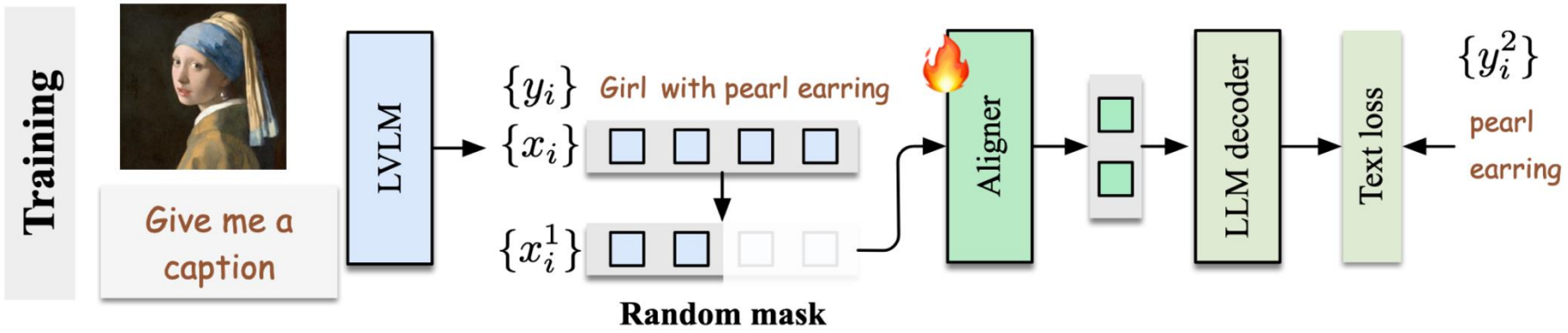
Problem 3: Stable training.

By adding a layer norm in the Aligner and initialize it by the parameters of the LLM (T5) encoder, we stable the training loss.



Problem 4: shortcut mapping issue.

When aligning the deep features of the VLM's generated tokens, the input of the Aligner has a one-to-one correspondence with the text supervision, causing the aligner to learn a trivial mapping.



We use a random mask strategy for training.

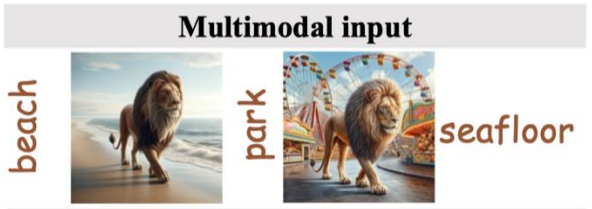
Evaluation

We CoBSAT for the evaluation of the multimodal in-context reasoning generation of ThinkDiff-LVLM.

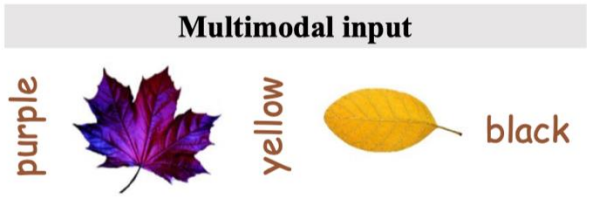
Table 2. 4-shot CoBSAT accuracy of ThinkDiff-LVLM shows a 27% average improvement over other methods and a 4.7% increase over its 2-shot results, highlighting its ability to handle complex in-context reasoning. In contrast, SEED-LLaMA (Ge et al., 2024), Emu (Sun et al., 2023), and GILL (Koh et al., 2024) exhibit reduced performance in 4-shot evaluations, indicating their struggle with increased input complexity. Improvement ratios over SoTA are also provided.

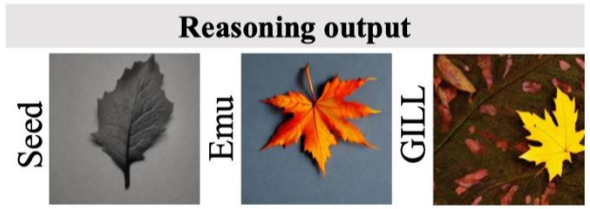
	Color-I	Background-I	Style-I	Action-I	Texture-I	Color-II	Background-II	Style-II	Action-II	Texture-II
SEED-LLaMA	0.482	0.211	0.141	0.053	0.122	0.252	0.076	0.268	0.207	0.105
Emu	0.063	0.018	0.045	0.048	0.097	0.037	0.122	0.109	0.077	0.088
GILL	0.106	0.044	0.041	0.073	0.087	0.022	0.059	0.044	0.032	0.067
Ours	**0.638**	**0.362**	**0.254**	**0.434**	**0.317**	**0.610**	**0.590**	**0.432**	**0.664**	**0.332**
Improvement (Δ%)	**32.4%**	**71.6%**	**80.1%**	**718.9%**	**159.8%**	**142.1%**	**676.3%**	**61.2%**	**220.8%**	**216.2%**

Quality results



GT answer: lion on seafloor

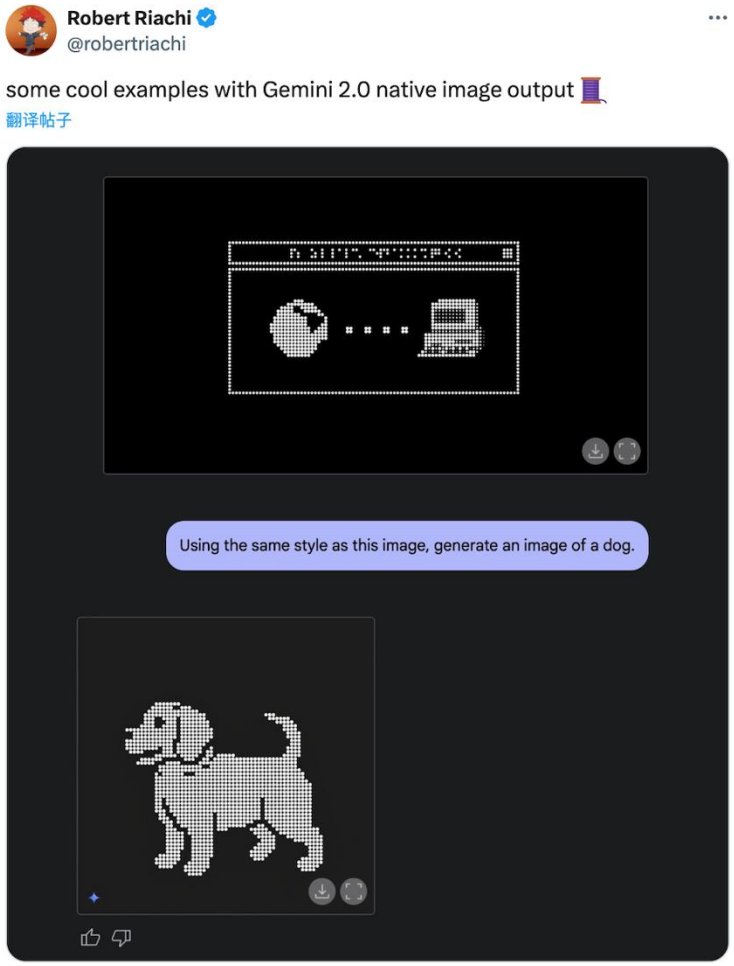




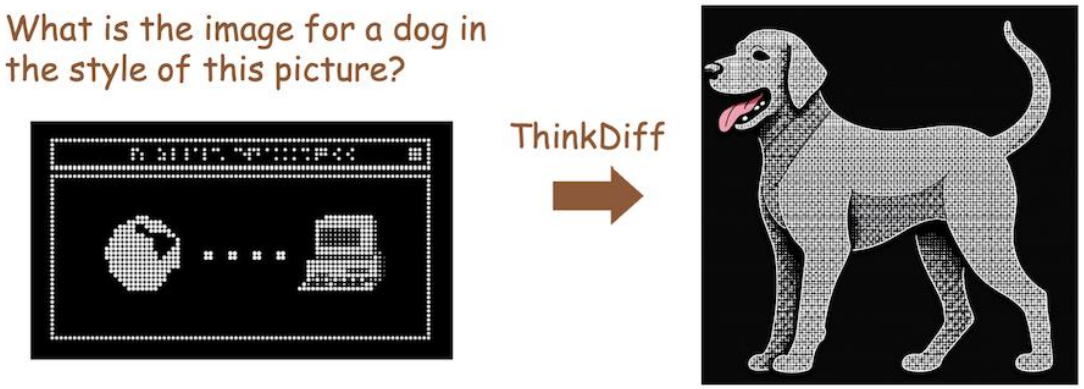
GT answer: black leaf

GT answer: wicker apple

A case comparing with Gemini.



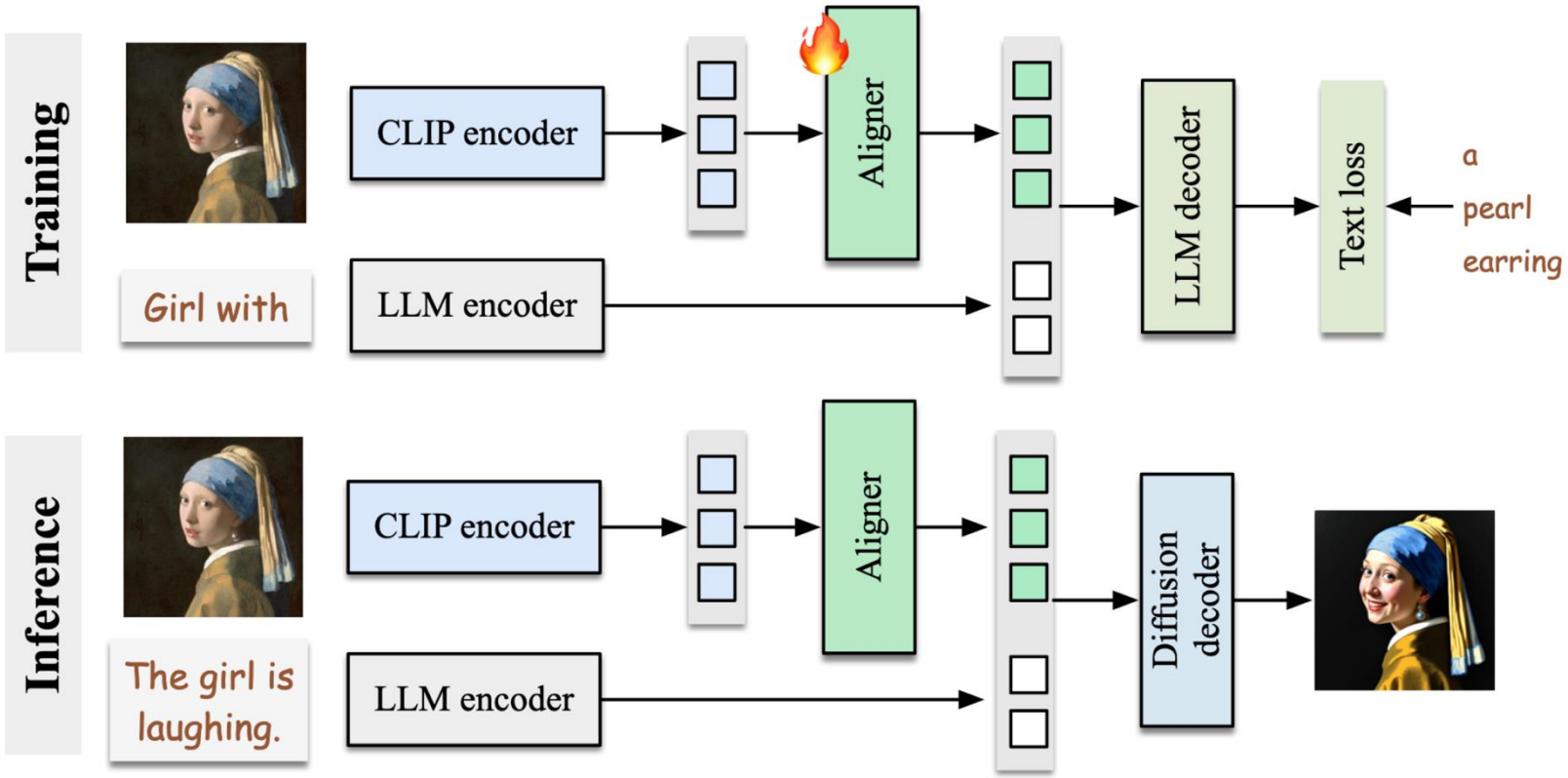
Google Gemini



ThinkDiff-LVLM

ThinkDiff-CLIP

What if we use CLIP as the VLM.



(b) ThinkDiff-CLIP

ThinkDiff-CLIP

Excellent multimodal composition.

ThinkDiff-CLIP with CogVideo

Excellent multimodal composition.

A panda, dressed in a small, red jacket and a tiny hat, sits on a wooden stool in a serene bamboo forest. The panda's fluffy paws strum a miniature acoustic guitar, producing soft, melodic tunes. Nearby, a few other pandas gather, watching curiously and some clapping in rhythm. Sunlight filters through the tall bamboo, casting a gentle glow on the scene. The panda's face is expressive, showing concentration and joy as it plays. The background includes a small, flowing stream and vibrant green foliage, enhancing the peaceful and magical atmosphere of this unique musical performance.

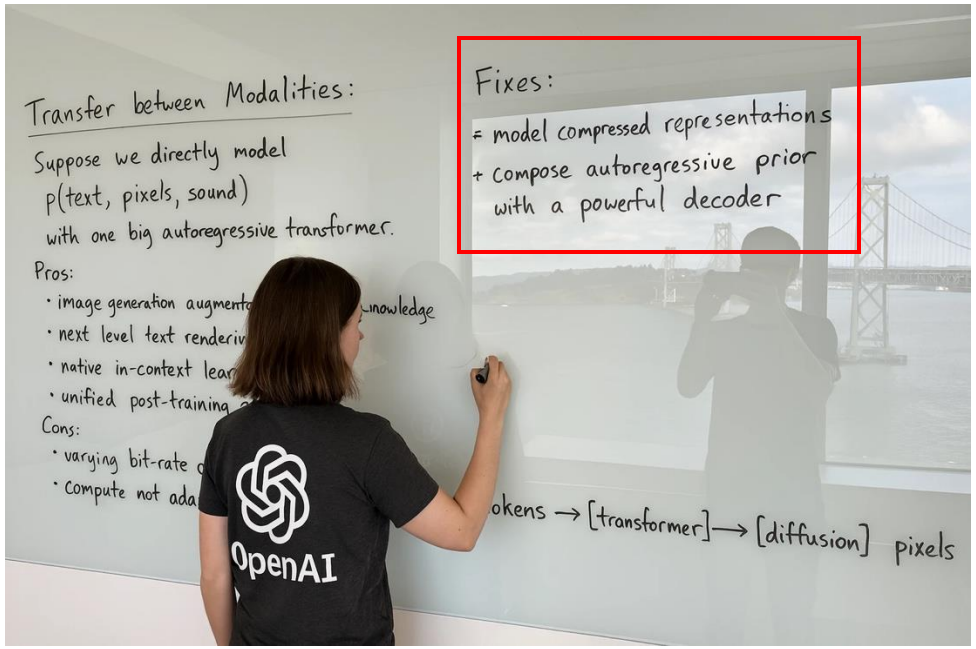
Future work

Our method is more semantic.
However, fidelity is important for design and editing.

Input

ThinkDiff

"There is a speech balloon showing 'Meow'"



We may use intermediate VLM features to transfer more vision information, and end-to-end diffusion training to improve the fidelity of the generated images.

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Snap Research

ICML International Conference On Machine Learning

Thanks for listening

Code

github.com/MiZhenxing/ThinkDiff