ICML
International Conference
On Machine Learning

GMAIL: Generative Modality Alignment for generated Image Learning

Shentong Mo, Sukmin Yun (presenter)

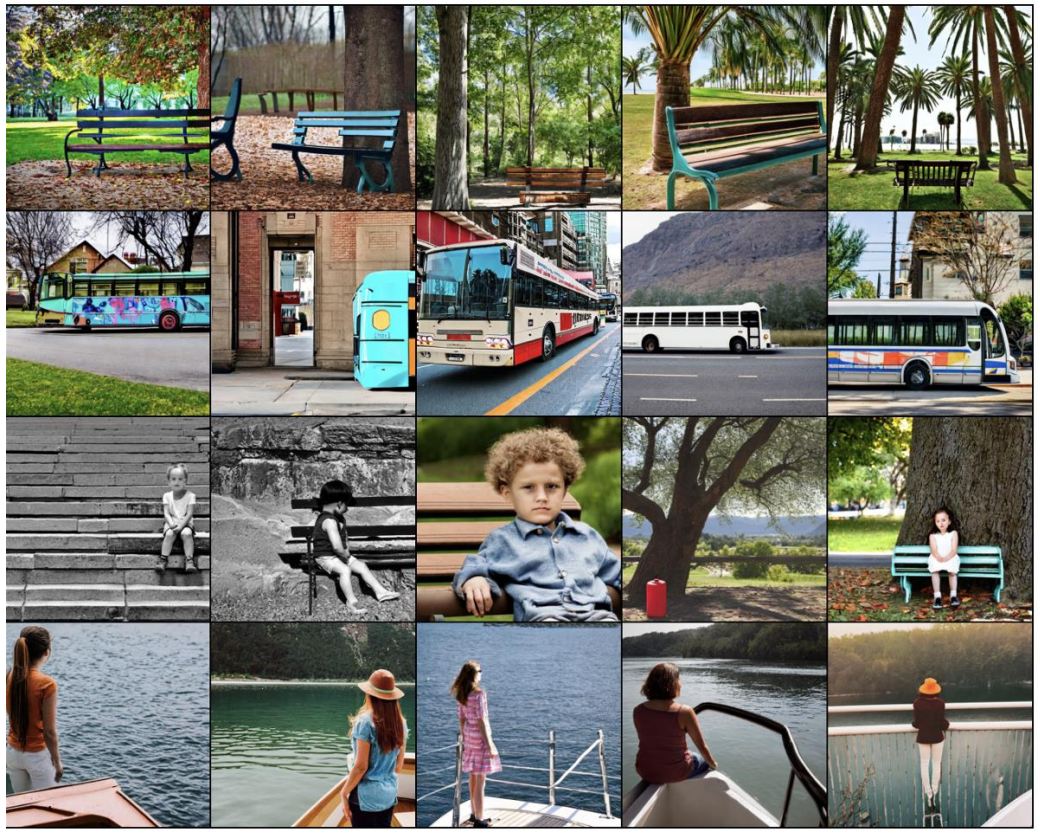
VAIL@HYU
VISION
ARTIFICIAL
INTELLIGENCE
LAB.

Introduction

Recent generative models have made it possible to synthesize highly realistic images

- Potentially providing an abundant data source for training machine learning models
- Generated images could generally capture high-level semantics in real images

Real Images

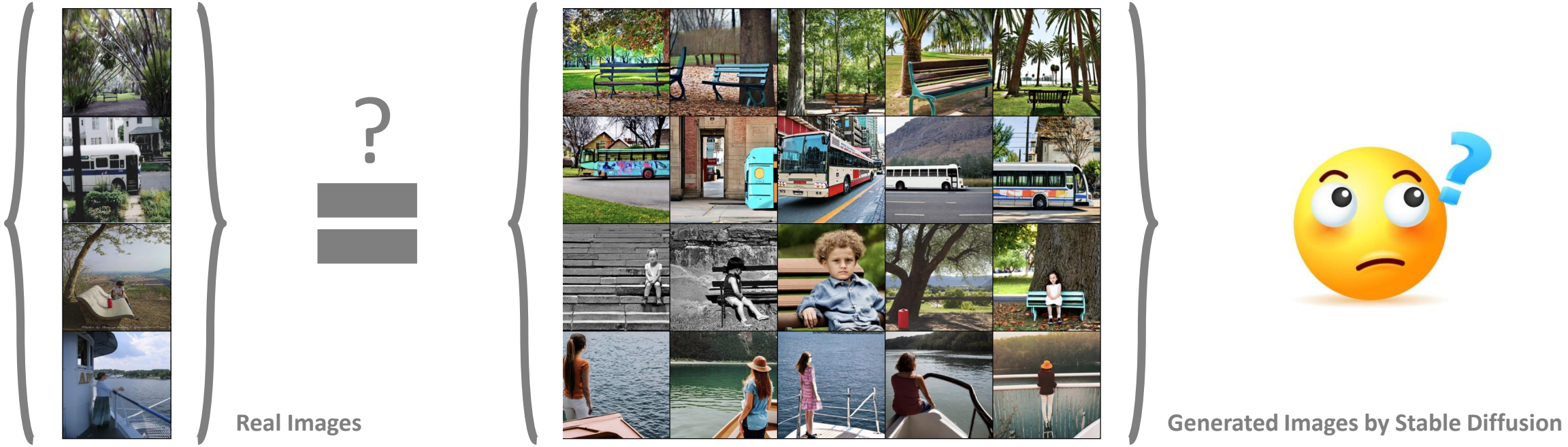


Generated Images by Stable Diffusion
Same caption used

Motivation

However, they may still exhibit subtle artifacts or variations that could contribute to the modality gap

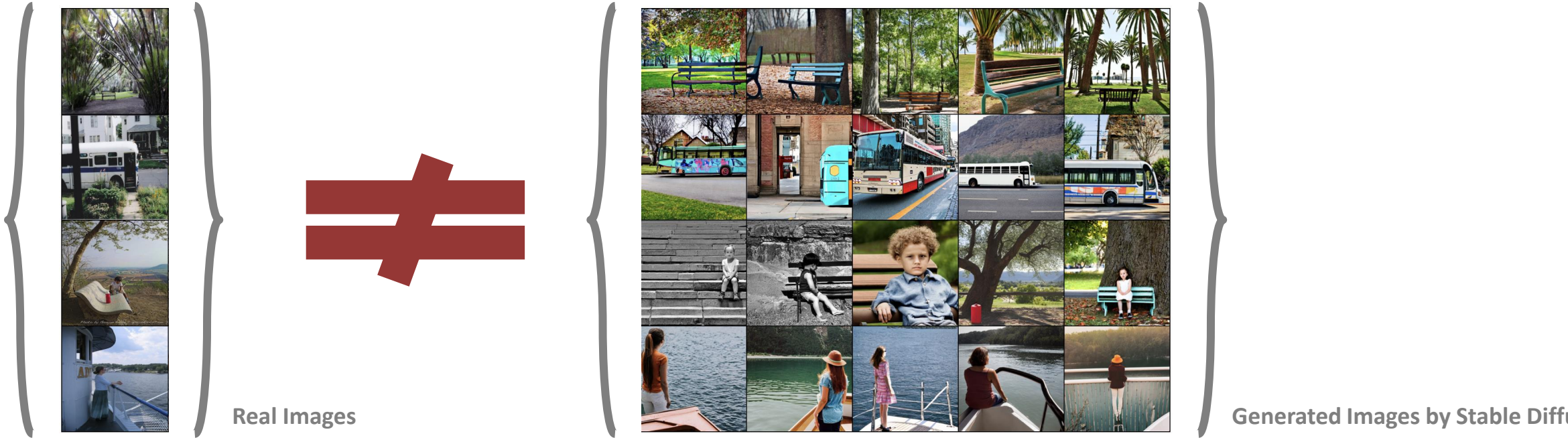
- Over-reliance on synthetic data could be harmful when models are applied to real-world task
- **Questions:** Do they really have the same modality?
  - Previously, we indiscriminately use generated images as real images for training



Motivation

Idea: We assume that they have different modalities

- We treat generated images as a separate modality from real images
- We bridge the two distinct modalities in the same latent space through a multi-modal learning approach
  - E.g., Contrastive Language-Image Pretraining (CLIP)



Real Images

Generated Images by Stable Diffusion

GMAIL: Generative Modality Alignment for generated Image Learning

- We aim to train Vision-Language models on generated images made by diffusion models

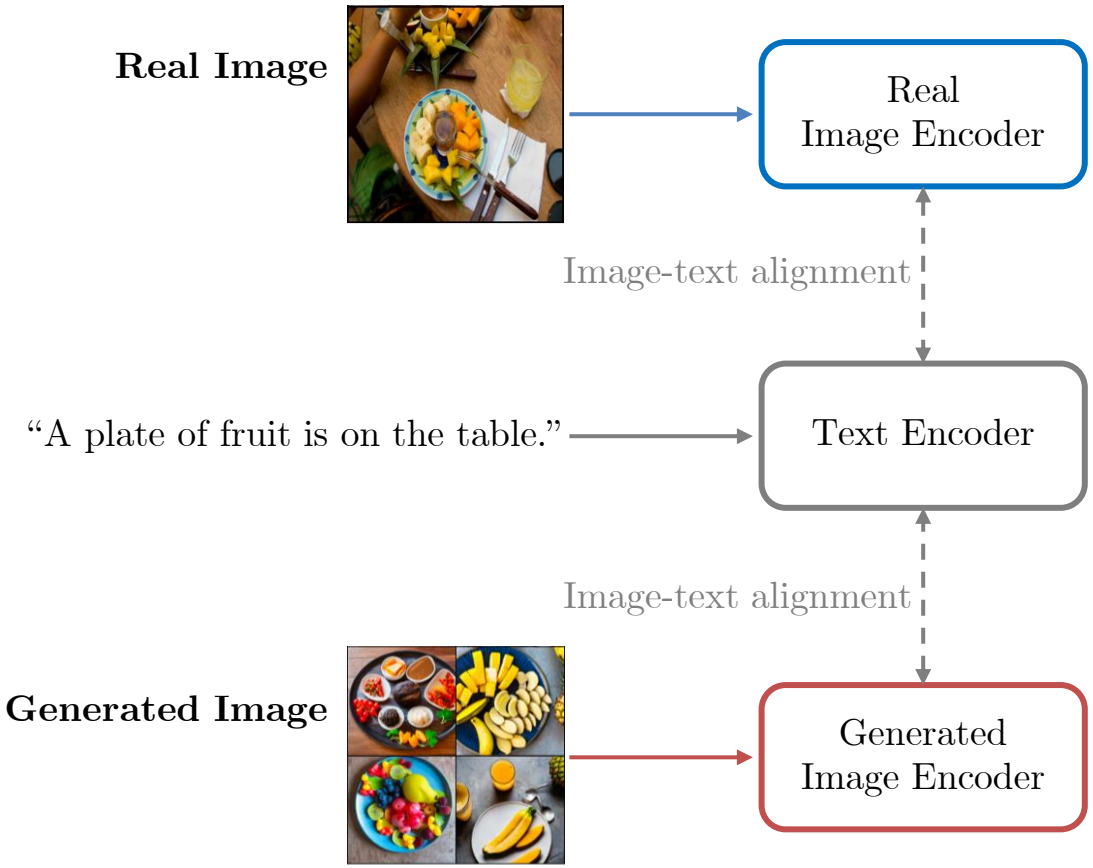
Real Image

"A plate of fruit is on the table."

Generated Image

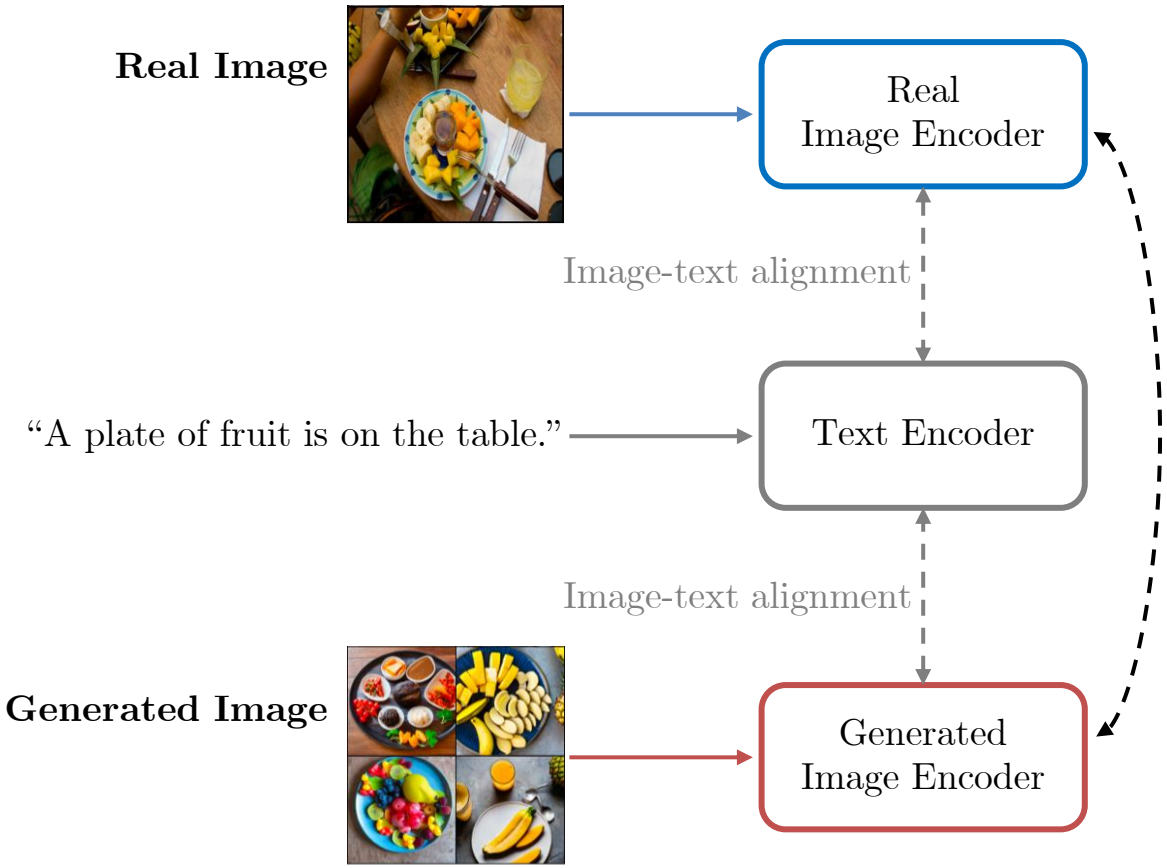
GMAIL: Generative Modality Alignment for generated Image Learning

- We aim to train Vision-Language models on generated images made by diffusion models



GMAIL: Generative Modality Alignment for generated Image Learning

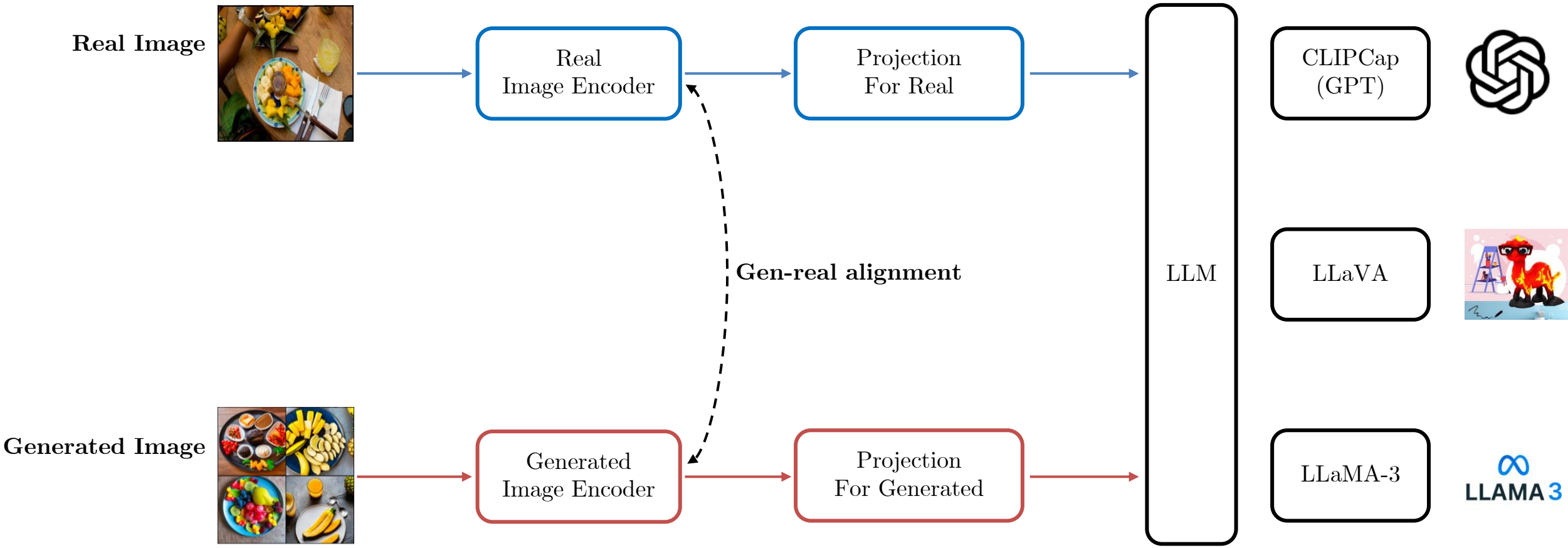
- We aim to train Vision-Language models on generated images made by diffusion models



$$\mathcal{L}_{align} = -\frac{1}{|\mathcal{B}|} \sum_{(x_g, x_r) \in \mathcal{B}} \log \frac{\exp(\text{sim}(f_g(x_g), f_r(x_r))/\tau)}{\sum_{x'_r \in \mathcal{B}} \exp(\text{sim}(f_g(x_g), f_r(x'_r))/\tau)},$$

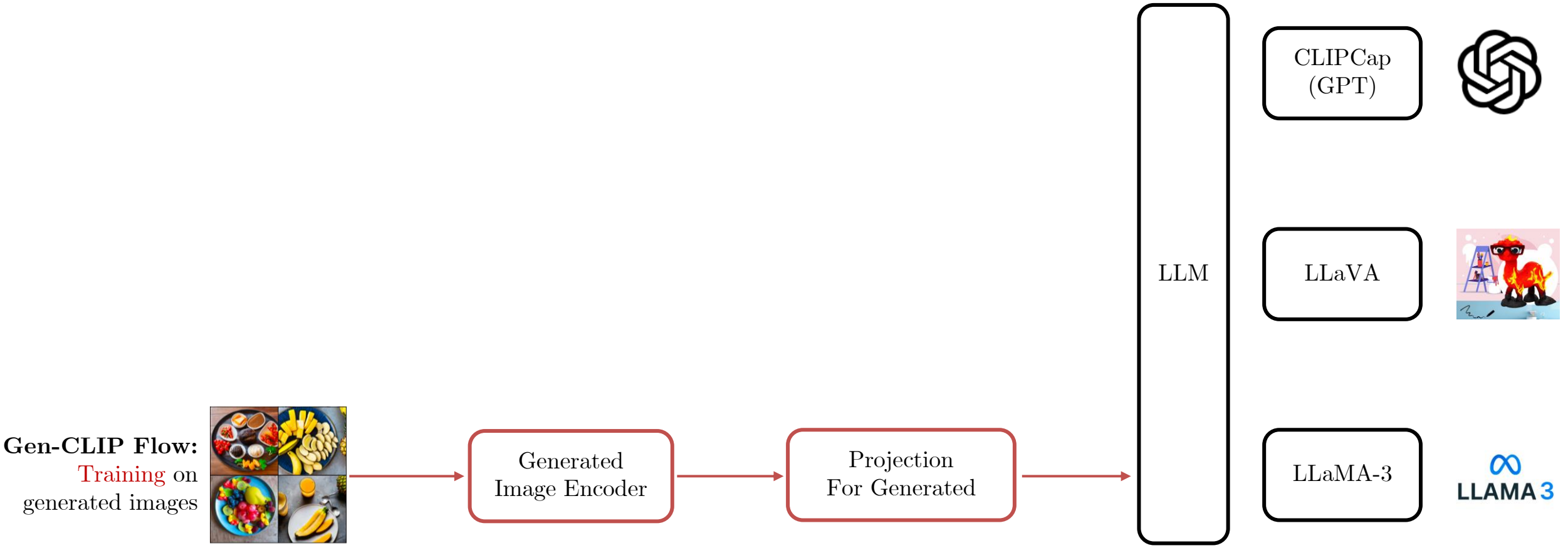
GMAIL: Generative Modality Alignment for generated Image Learning

- We aim to train Vision-Language models on generated images made by diffusion models



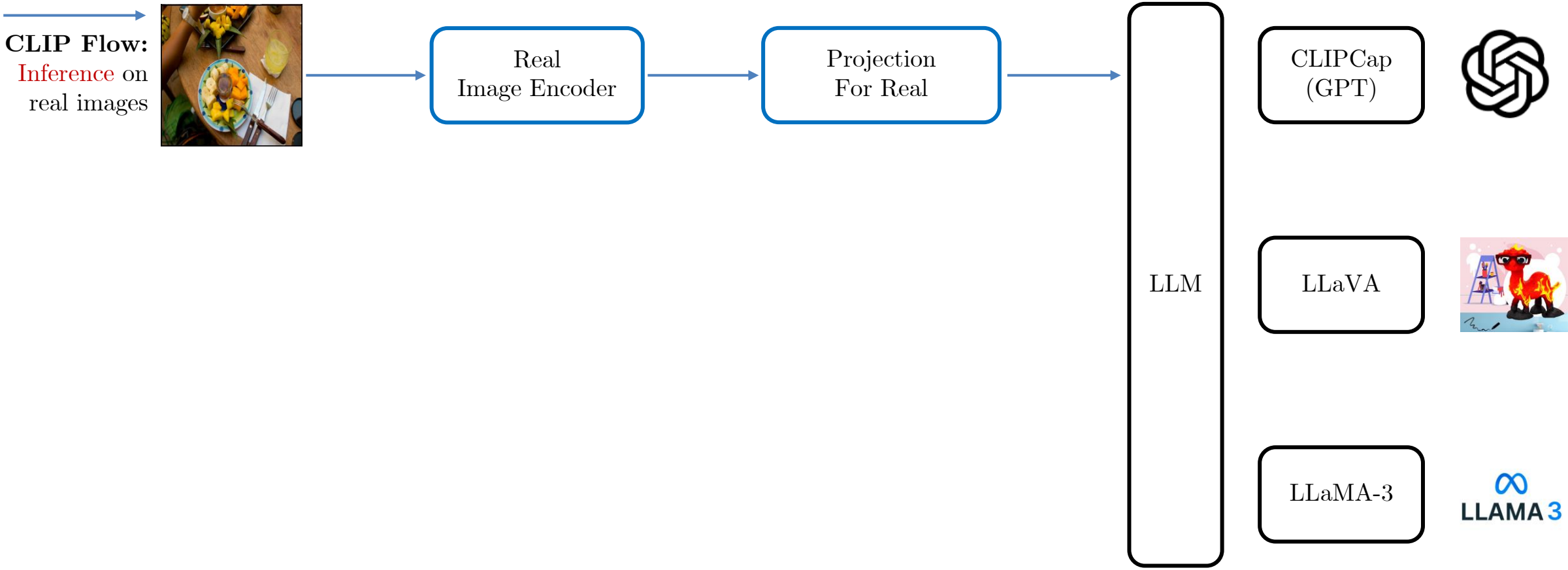
GMAIL: Generative Modality Alignment for generated Image Learning

- We aim to train Vision-Language models on generated images made by diffusion models



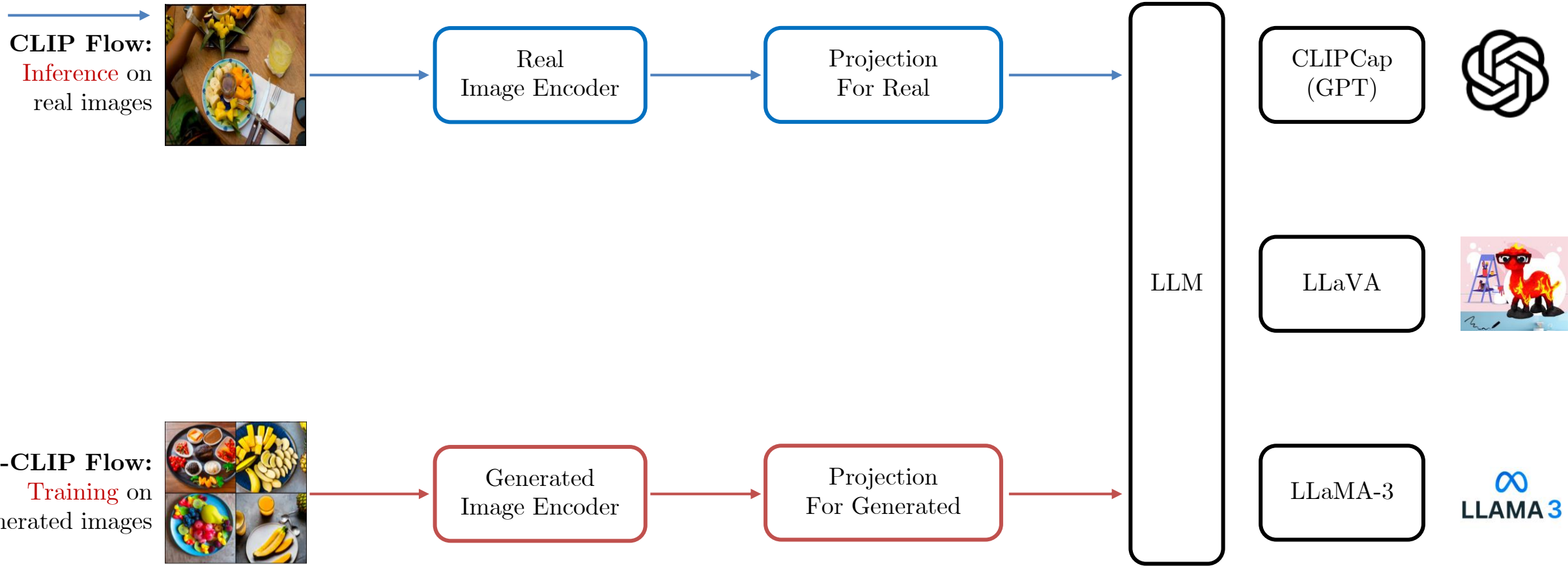
GMAIL: Generative Modality Alignment for generated Image Learning

- We aim to train Vision-Language models on generated images made by diffusion models



GMAIL: Generative Modality Alignment for generated Image Learning

- We aim to train Vision-Language models on generated images made by diffusion models



Experiments

Performances on various vision-language models and tasks

- We generate images using Stable Diffusion v2
- **Image captioning**

Method	B@4 (↑)	METEOR(↑)	CIDEr (↑)	SPICE (↑)	ROUGE-L (↑)	WMD (↑)
ClipCap (Mokady et al., 2021)	32.15	27.10	108.35	20.12	–	–
ClipCap + GMAIL (ours)	**38.12**	**31.67**	**119.53**	**23.75**	**56.27**	**62.16**
IFCap (Lee et al., 2024)	33.25	28.60	115.27	21.58	51.35	56.72
IFCap + GAMIL (ours)	**39.32**	**32.07**	**127.86**	**23.98**	**59.83**	**63.51**
LLaVA (Liu et al., 2023)	39.67	32.38	134.29	24.17	61.36	65.78
LLaVA + GMAIL (ours)	**43.26**	**34.89**	**146.38**	**27.23**	**65.25**	**71.39**
Llama3 (Meta, 2024)	47.36	35.21	158.13	28.35	68.32	75.13
Llama3 + GMAIL (ours)	**50.21**	**38.59**	**168.53**	**32.58**	**73.29**	**80.25**

- **VQA on ScienceQA and MMMU**

Method	Accuracy (%)
LLaVA	85.2
LLaVA + GMAIL (ours)	**87.6**
LLaMA-3	88.5
LLaMA-3 + GMAIL (ours)	**91.2**

ScienceQA

Method	Accuracy (%)
LLaVA	44.7
LLaVA + GMAIL (ours)	**48.3**

MMMU

Experiments

Comparisons with CLIP-based approaches and tasks

- **Zero-shot image retrieval on Flickr30k**

Method	Image-to-Text			Text-to-Image		
	R@1 (↑)	R@5 (↑)	R@10 (↑)	R@1 (↑)	R@5 (↑)	R@10 (↑)
CLIP (Radford et al., 2021)	44.1	68.2	77.0	24.7	45.1	54.6
CLIP + GMAIL (ours)	**47.1**	**71.2**	**79.6**	**30.2**	**50.3**	**60.5**
Long-CLIP (Zhang et al., 2024)	47.2	71.5	80.0	33.1	55.6	64.9
Long-CLIP + GMAIL (ours)	**51.6**	**75.3**	**83.6**	**39.3**	**61.5**	**71.8**

  - This demonstrates that GMAIL consistently enhances model performance across different backbone architectures by facilitating better alignment of generated images with real-world data

- **Scaling trend of GMAIL on Flickr30k zero-shot image retrieval**

Train Data	Image-to-Text			Text-to-Image		
	R@1 (↑)	R@5 (↑)	R@10 (↑)	R@1 (↑)	R@5 (↑)	R@10 (↑)
COCO	47.1	71.2	79.6	30.2	50.3	60.5
CC3M	48.6	73.6	82.2	32.6	52.6	62.3
CC12M	**50.9**	**75.3**	**84.6**	**34.9**	**54.7**	**64.8**

  - The results reveal a clear scaling trend, where increasing the volume of training data from COCO to CC3M and then to CC12M consistently enhances the model's performance on both image-to-text and text-to-image retrieval tasks

Experiments

- **Effects of Gen-Real Alignment**

Alignment	B@4 (↑)	METEOR(↑)	CIDEr (↑)	SPICE (↑)	ROUGE-L (↑)	WMD (↑)
✗	36.15	30.32	115.35	22.95	55.12	61.08
✓	**38.12**	**31.67**	**119.53**	**23.75**	**56.27**	**62.16**

  - These improvements highlight the critical role of alignment fine-tuning in bridging the modality gap between generated and real images, which enables the model to better capture and replicate the semantic richness found in real-world data

- **Image generation with FLUX**

Method	B@4 (↑)	CIDEr (↑)	SPICE (↑)
FLUX (without alignment)	37.20	117.82	23.40
FLUX + GMAIL (ours)	**39.54**	**122.36**	**24.15**

  - We have conducted experiments using FLUX, which introduces a more powerful and differently parameterized generation pipeline compared to Stable Diffusion v2. The performance improvements remain consistent with FLUX, indicating robust alignment across varying artifact styles and photorealism levels.

Thank You for Your Attention