Towards Safe Concept Transfer of Multi-Modal Diffusion via Causal Representation Editing

Peiran Dong[1]* Bingjie Wang[1]* Song Guo[2] Junxiao Wang[3,4]
Jie Zhang[2] Zicong Hong[1]

[1]Hong Kong Polytechnic University [2]Hong Kong University of Science and Technology
[3]Guangzhou University [4]King Abdullah University of Science and Technology

Presenter: Bingjie Wang

Background

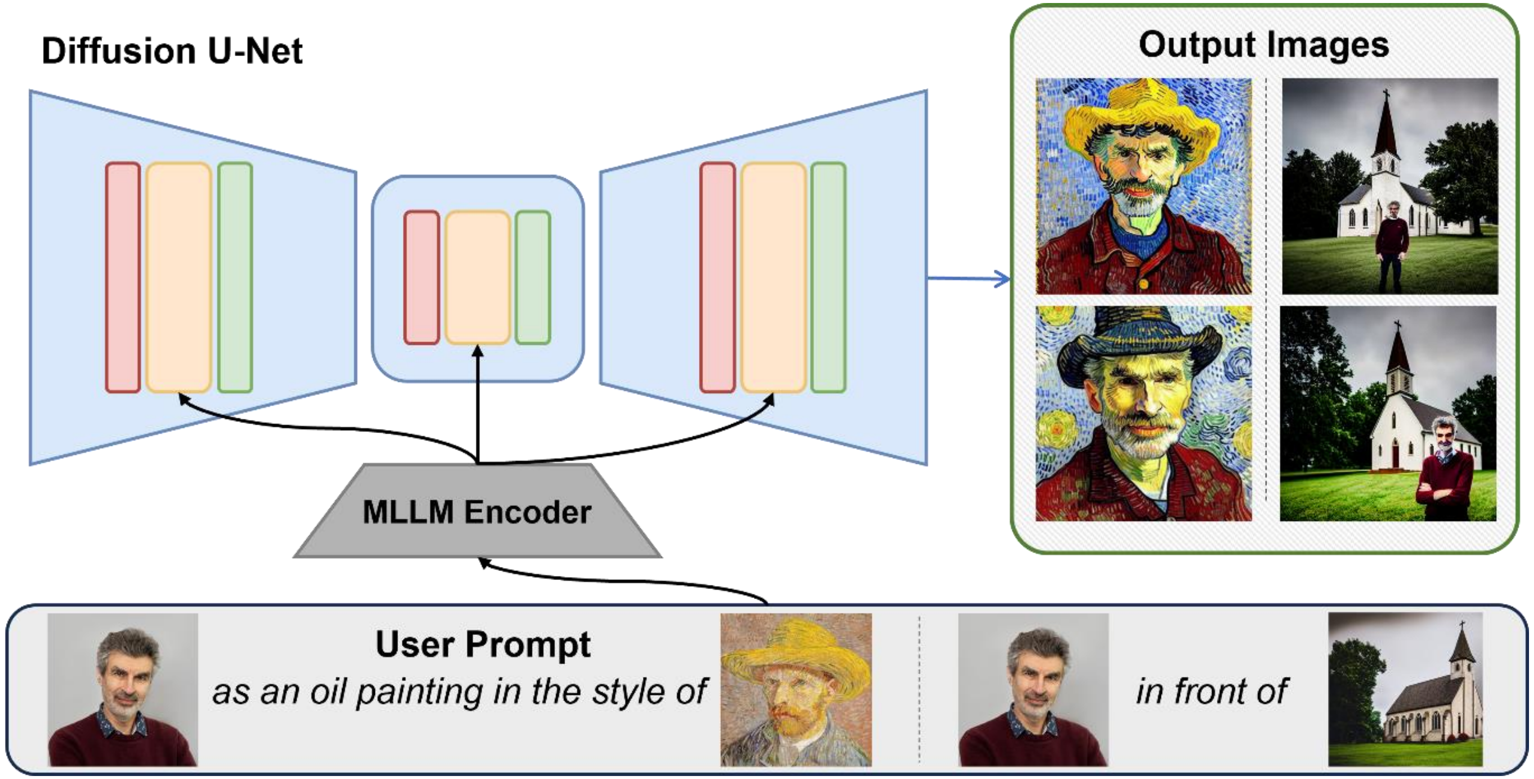


Figure 1. Multimodal Diffusion

Misuse of Multimodal Diffusion (like Kosmos-g[1]), like copying objects/styles in other images, leads to concerns about intellectual property rights.

[1] Pan, Xichen, et al. "Kosmos-g: Generating images in context with multimodal large language models." arXiv preprint arXiv:2310.02992 (2023).

Motivation

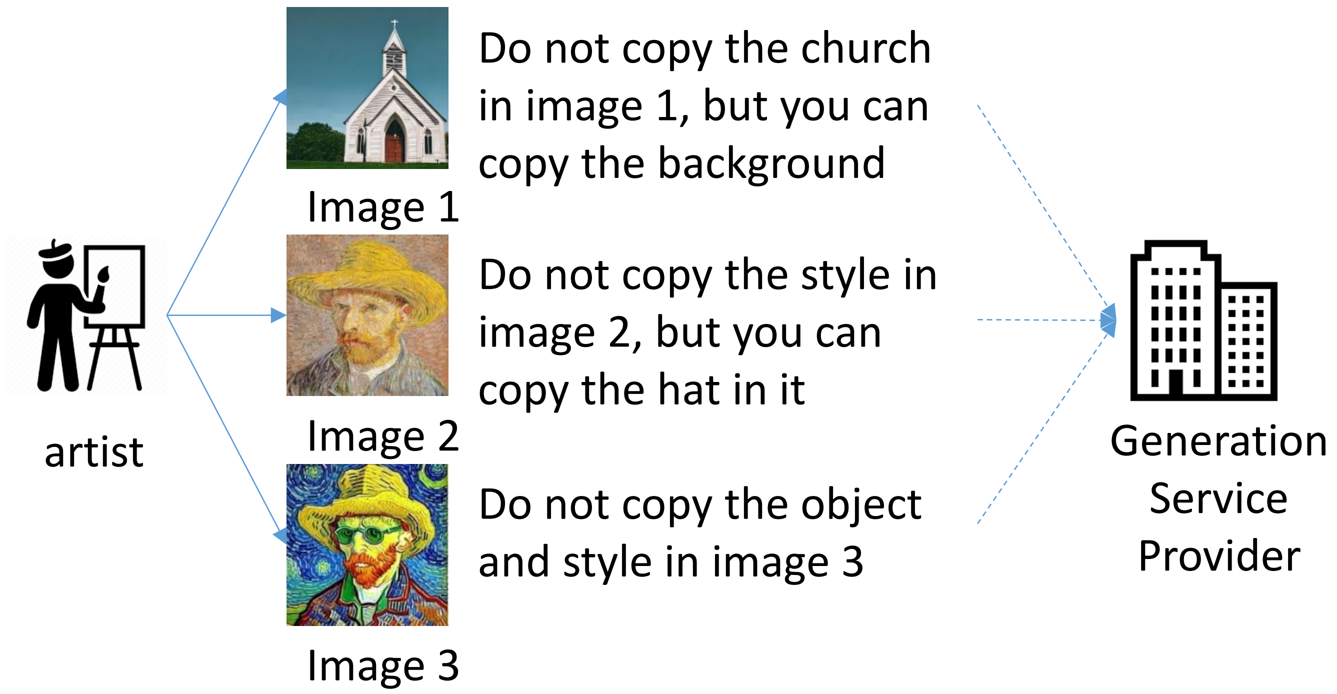


Figure 2. A possible scenario -- Artist demands

When Generation Service meets precise demands, for example, an artist tells the service providers which part they can use and what they can not use.

Motivation

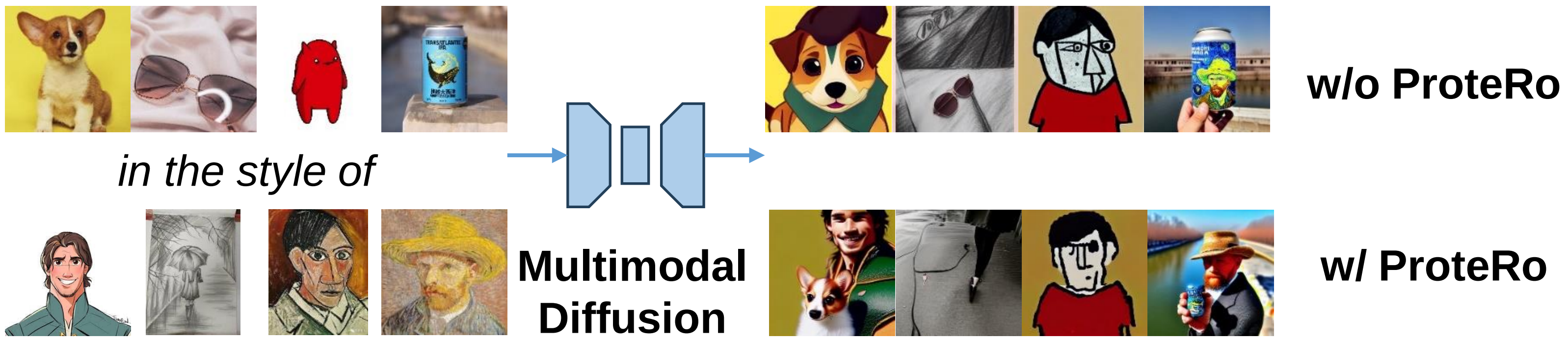


Figure 3. ProteRo [2] under multimodal setting

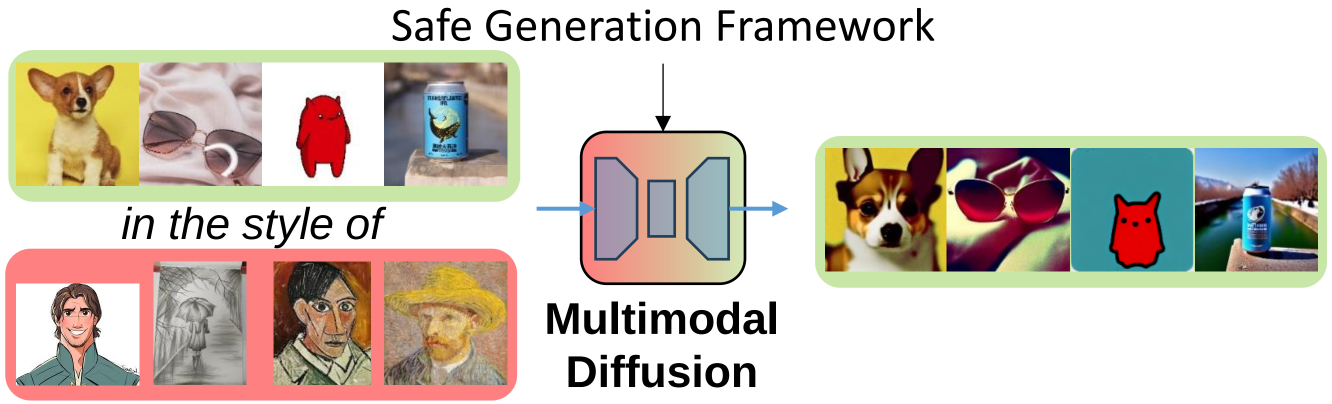


Figure 4. Wanted results with unsafe style

[2] Dong, Peiran, et al. "Towards Test-Time Refusals via Concept Negation." Advances in Neural Information Processing Systems 36 (2023).

Method

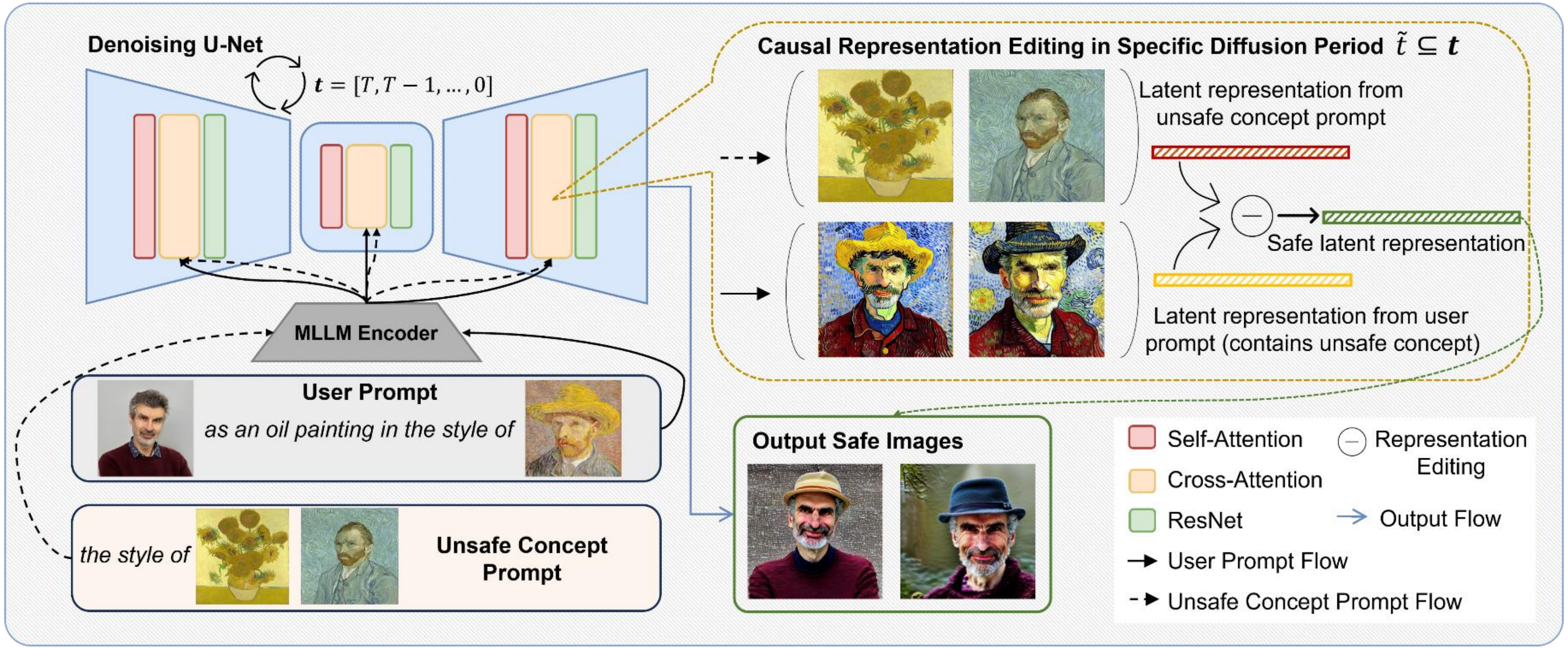


Figure 5. Pipeline for Safe Concept Transfer

Three Phases:

1. **Search: Searching for the unsafe input;**
2. **Prototype: Utilizing the MLLM encoder to get the unsafe embedding;**
3. **Refine: Remain safe parts of the embedding.**

Method

Phase 1 – Search: Searching for the unsafe input

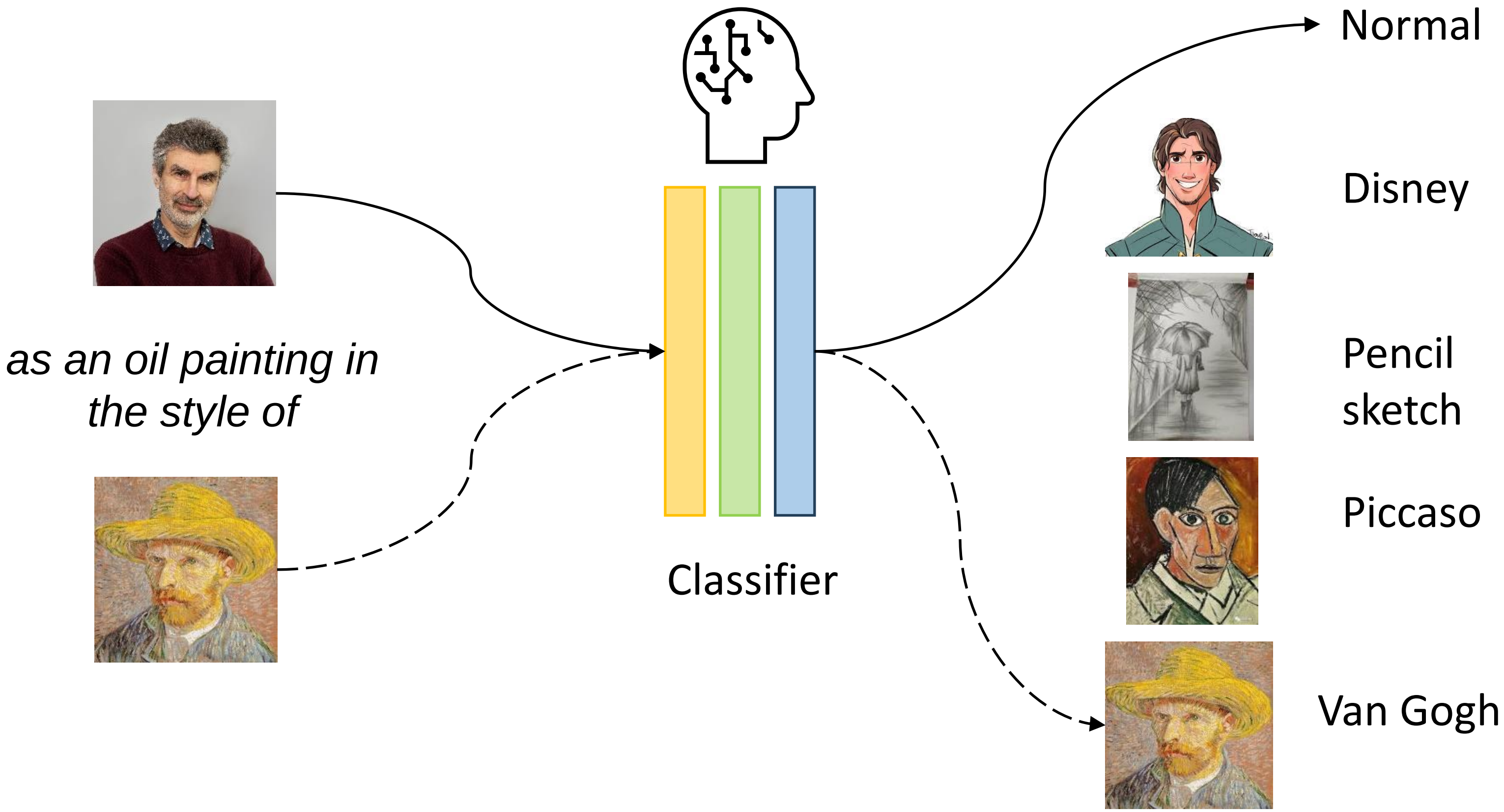


Figure 6. Searching for the unsafe input

Method

Phase 2 - Prototype: Utilizing the MLLM encoder to get the unsafe embedding

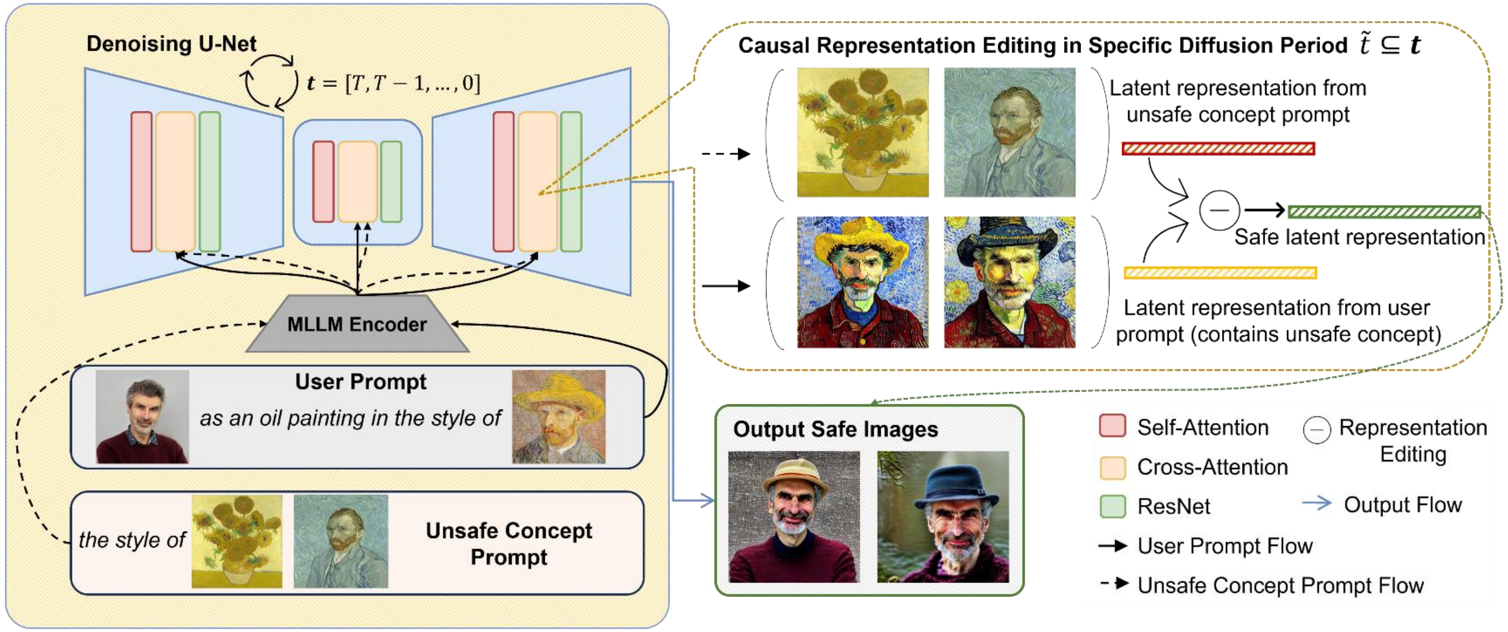
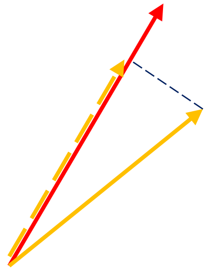


Figure 7. Utilizing the MLLM encoder to get the unsafe embedding

Method

Phase 3 - Refine: Using projection to get unsafe parts



User prompt embedding

Unsafe prompt embedding

Projected unsafe prompt embedding

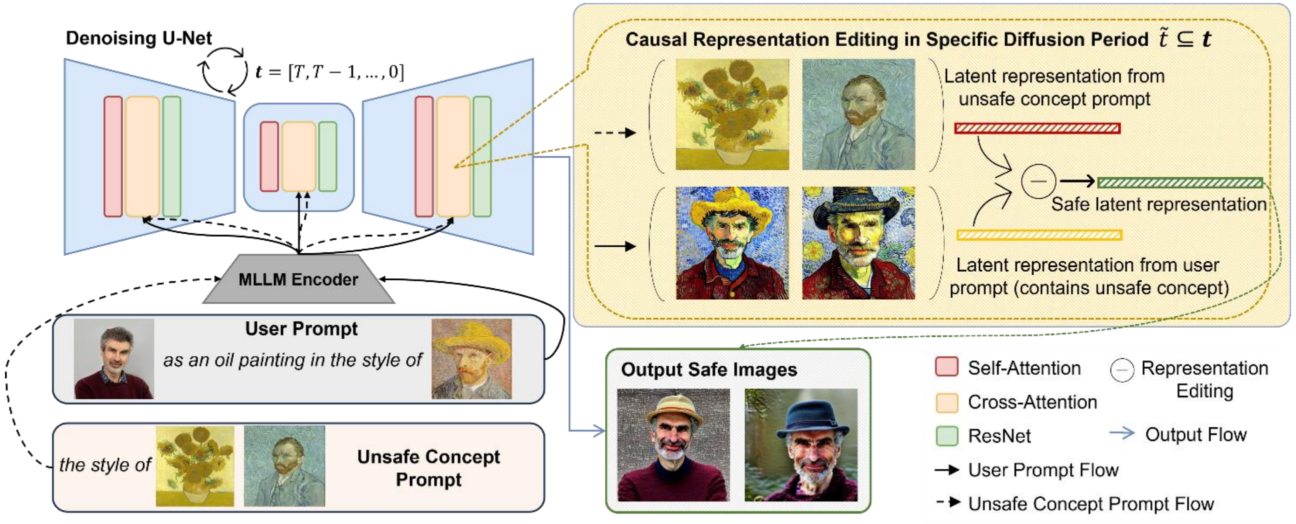


Figure 8. Using projection to get unsafe parts

Experiment - object

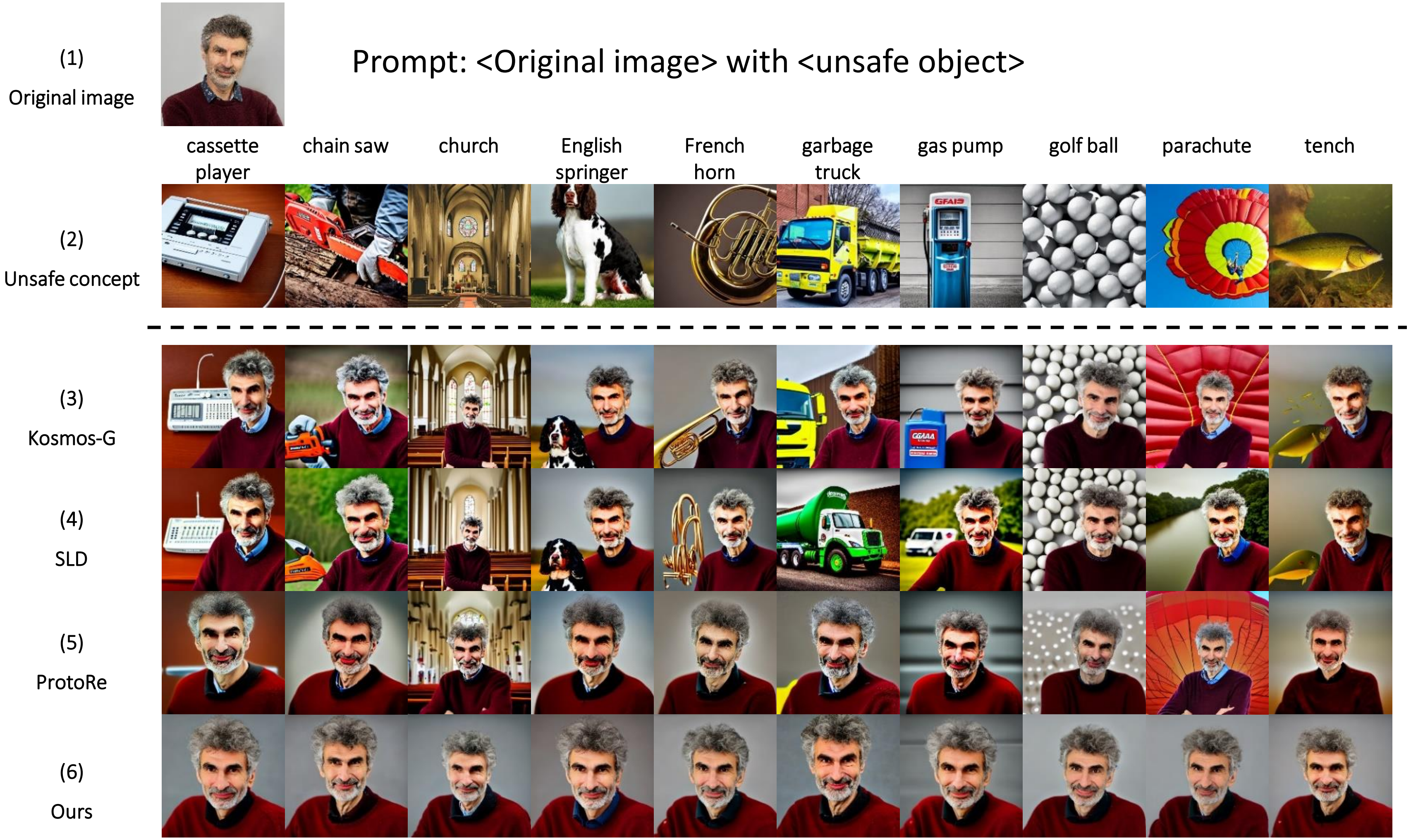


Figure 9. Qualitative results of safe object generation

Experiment - style

Prompt: <Original image> in the style of <unsafe style>

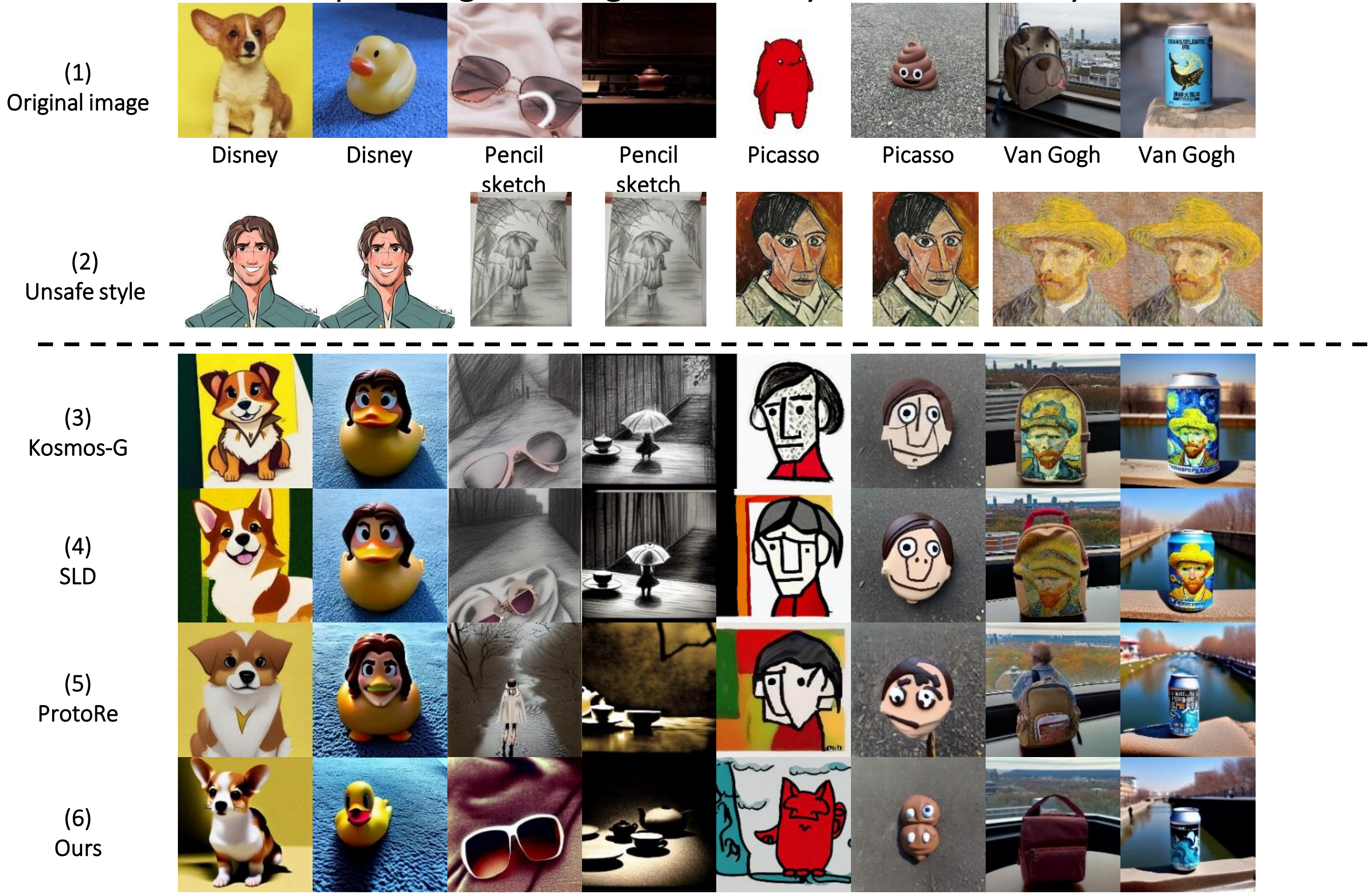


Figure 10. Qualitative results of safe object generation

NEURAL INFORMATION PROCESSING SYSTEMS

PERVASIVE INTELLIGENCE

THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

香港科技大學 THE HONG KONG CYBER UNIVERSITY OF SCIENCE AND TECHNOLOGY

Experiment – Quantitative Results

Object	Top-1 Accuracy of Object Transfer (%) ↓										
	cassette player	chain saw	church	English springer	French horn	garbage truck	gas pump	golf ball	parachute	tench	Average
Kosmos-G	5.2	50.6	96.6	27.2	12.0	52.6	34.4	24.2	43.2	16.6	36.26
Kosmos-G-Neg	9.4	51.6	95.6	31.8	6.6	59.6	32.4	28.6	39.4	11.4	36.76
SLD	0.8	18.4	95.6	15.4	11.4	30.6	16.2	7.0	27.6	1.8	22.48
ProtoRe	**0**	**0**	15.6	**0**	**0**	**0**	**0**	0.2	0.8	**0**	1.66
CRE	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

Table 1. Quantitative Results of safe object transfer

Discriminator	Style	Top-1 Accuracy of Style Transfer (%) ↓				
		Kosmos-G	Kosmos-G-Neg	SLD	ProtoRe	CRE
ResNet-50	Disney	53.9241	61.4557	56.7089	47.5949	**11.3924**
	Pencil Sketch	19.2405	44.3671	14.8101	12.9747	**0.6962**
	Picasso	21.8354	36.519	11.2658	3.6709	**0.3165**
	Van Gogh	44.4304	60.443	26.2658	2.7848	**0.5696**
ViT-base	Disney	39.557	44.2405	36.6456	29.557	**1.3291**
	Pencil Sketch	15.5063	35.8861	10.5063	6.7722	**0.6329**
	Picasso	22.1519	35.1266	15.3165	5.1899	**1.6456**
	Van Gogh	44.1139	60.443	27.9114	3.2278	**0.3797**
Average		32.5949	47.3101	24.9288	13.9715	**2.1202**

Table 2. Quantitative Results of safe style transfer

Experiment – Complex Senarios

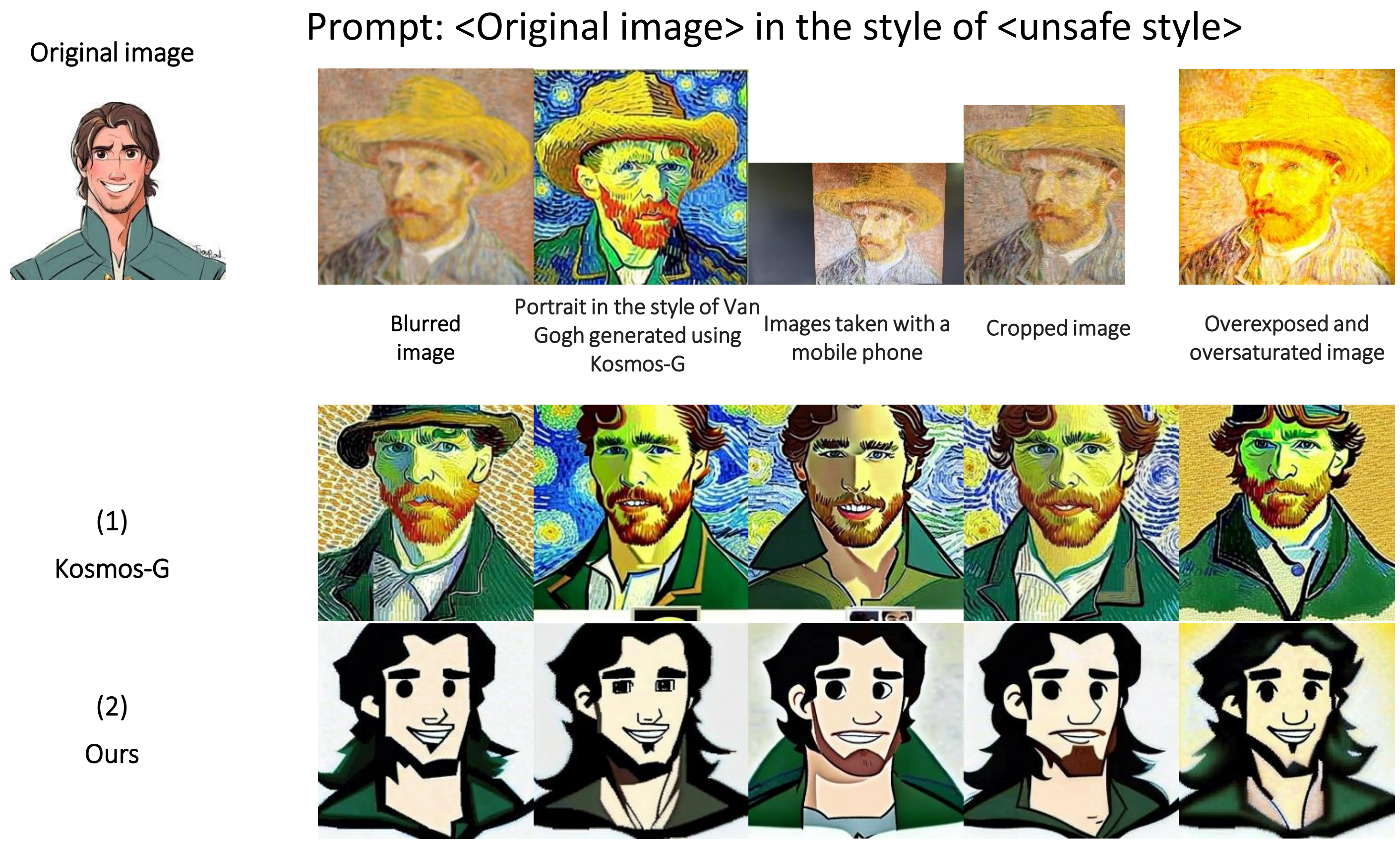
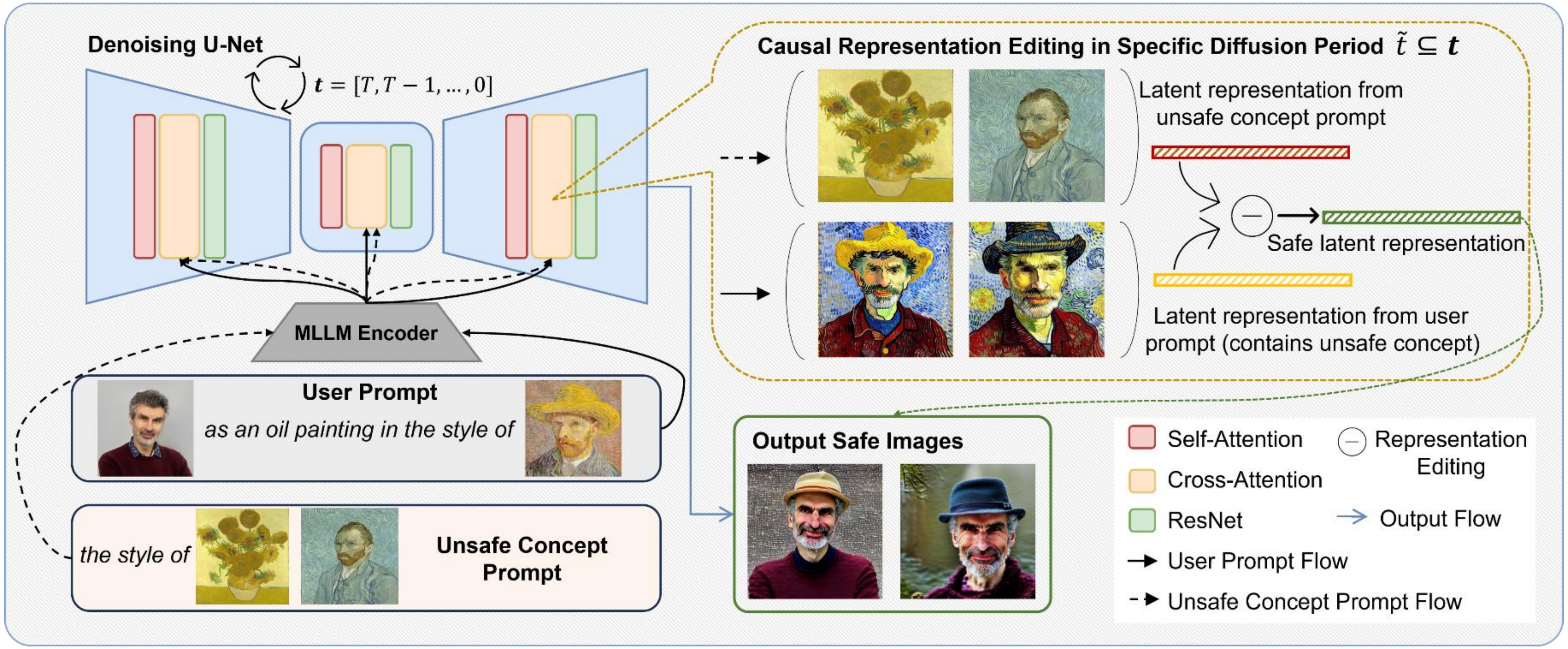


Figure 11. Qualitative results of complex Scenarios

Conclusion


Denoising U-Net
$t = [T, T-1, \ldots, 0]$
Causal Representation Editing in Specific Diffusion Period $\tilde{t} \subseteq t$
Latent representation from unsafe concept prompt
Safe latent representation
Latent representation from user prompt (contains unsafe concept)
MLLM Encoder
User Prompt
as an oil painting in the style of
the style of
Unsafe Concept Prompt
Output Safe Images
Self-Attention
Cross-Attention
ResNet
Representation Editing
Output Flow
User Prompt Flow
Unsafe Concept Prompt Flow