NEURAL INFORMATION PROCESSING SYSTEMS

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Magnet

We Never Know How Text-to-Image Diffusion Models Work, Until We Learn How Vision-Language Models Function

*Chenyi Zhuang[1], Ying Hu[1], Pan Gao[1,2]**

[1] Nanjing University of Aeronautics and Astronautics
[2] Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education

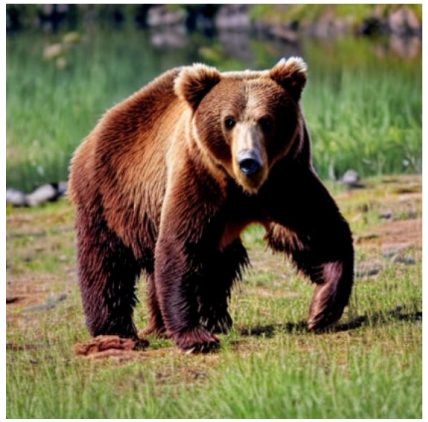
{chenyi.zhuang,ying.hu,pan.gao}@nuaa.edu.cn

Background

Stable Diffusion struggles to generate text-aligned images.

Prompt: “an orange dog wearing an gray bow tie laying on a sofa”

Attribute leakage



Prompt: “a brown bear wearing a pair of red glasses”

Missing Objects

Prior study: “the contextualization of **CLIP** embeddings is a potential cause.”

WHY CLIP’s fault?

Analysis – How does CLIP understand attribute?

Non-contextualized embeddings

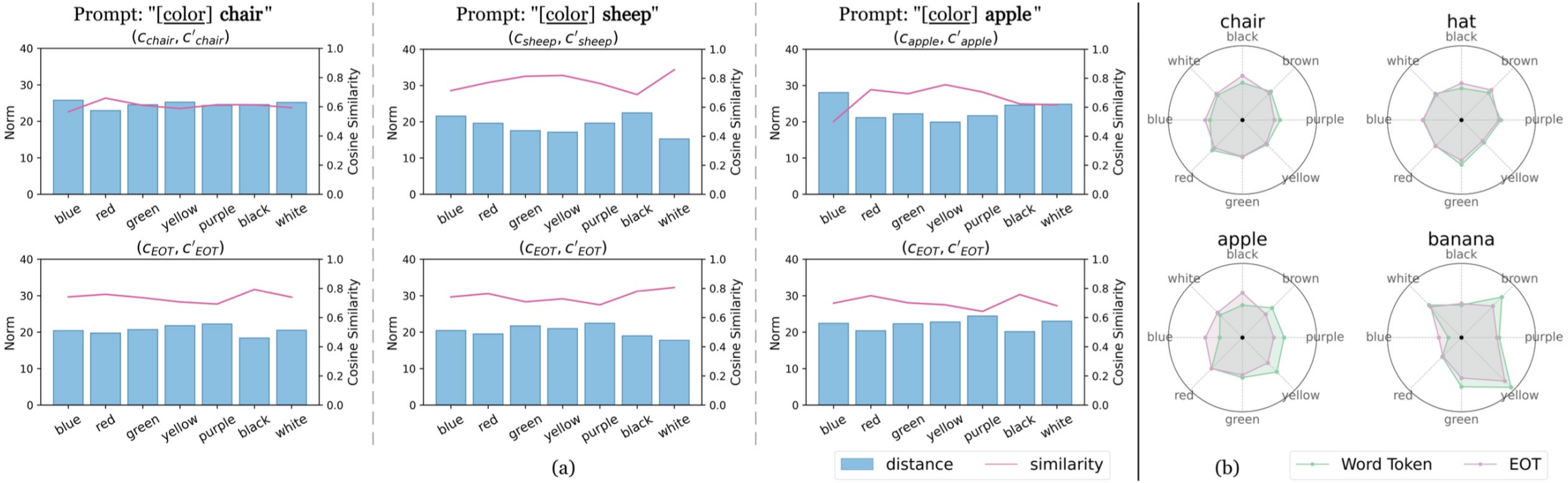
$c_{<SOT>}, c_{obj}, c_{<EOT>}$

Euclidean distance

Cosine similarity

Contextualized embeddings

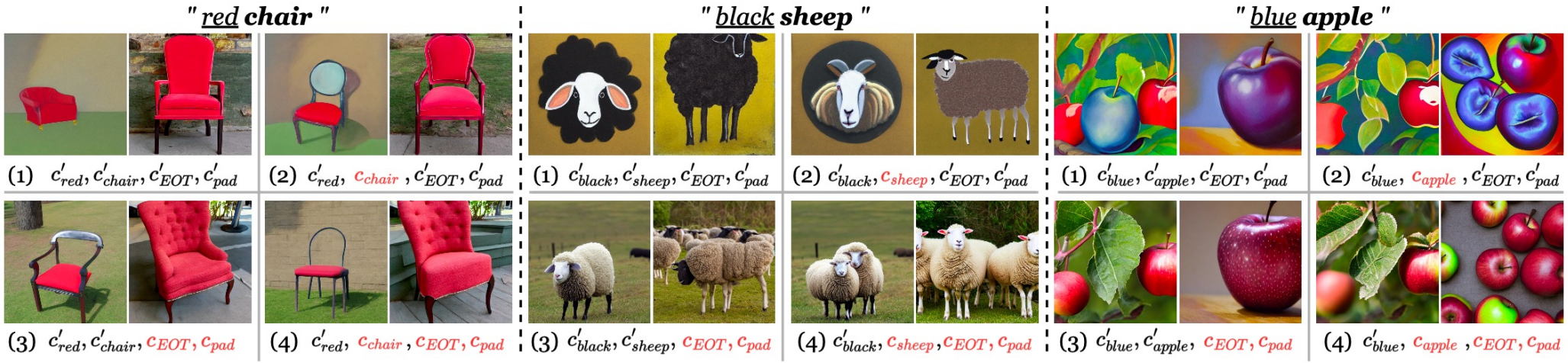
$c'_{<SOT>}, c'_{attr}, c'_{obj}, c'_{<EOT>}$



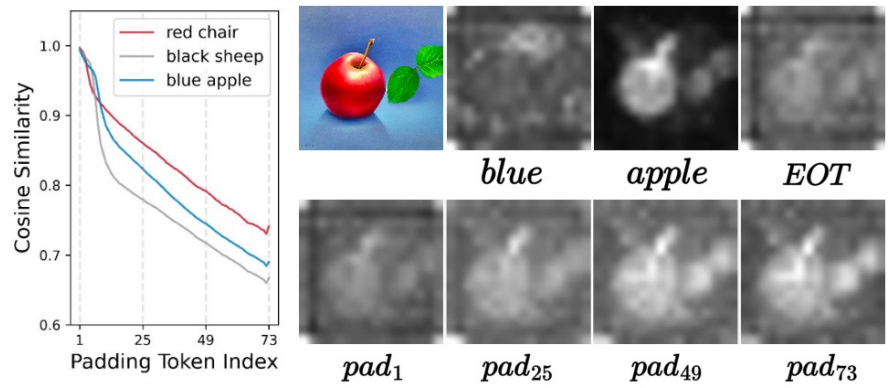
Attribute bias – the tendency of an object to *favor* certain attributes over others

Analysis – How does it affect T2I generation?

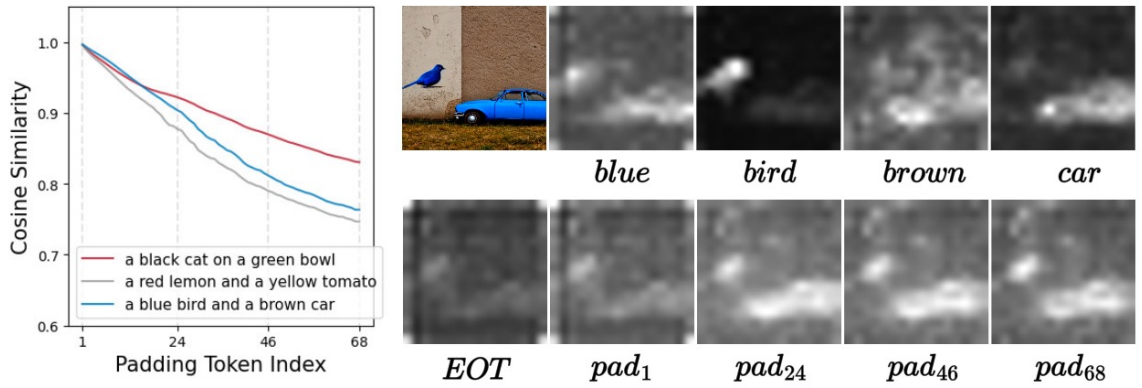
Replace specific parts of the original text embeddings to disturb the context information



(a) Fine-grained 4 cases



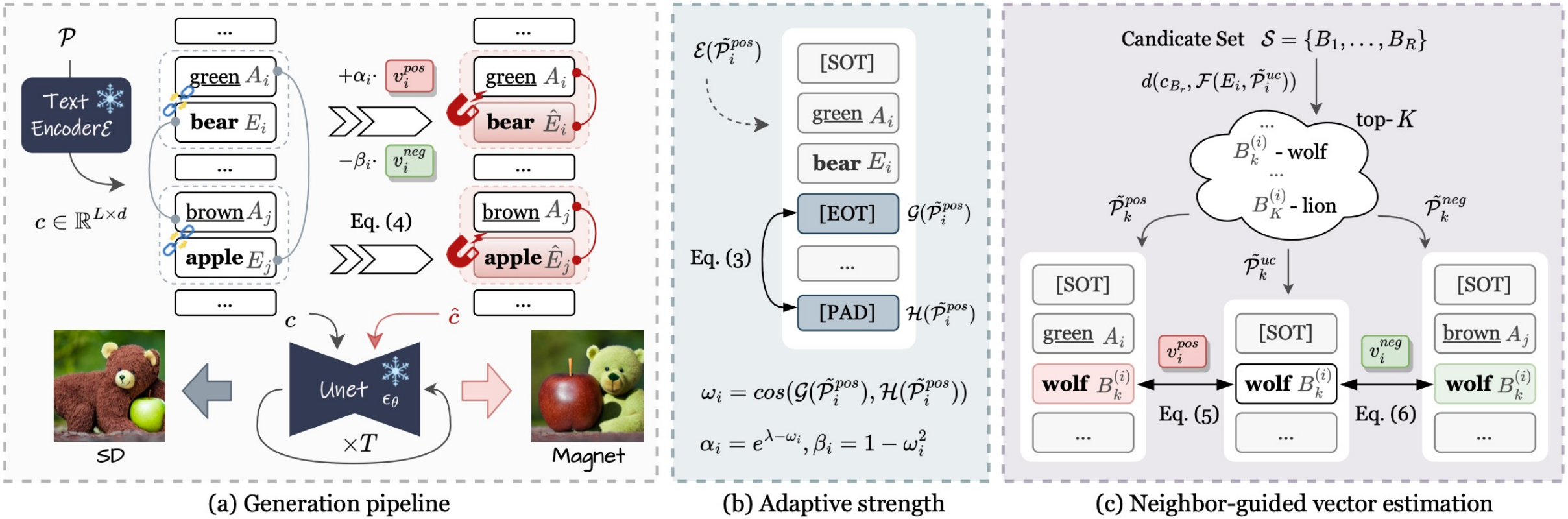
(b) Single-concept context issue



(c) Multi-concept context issue

Context issue – the padding embeddings contain inaccurate representations

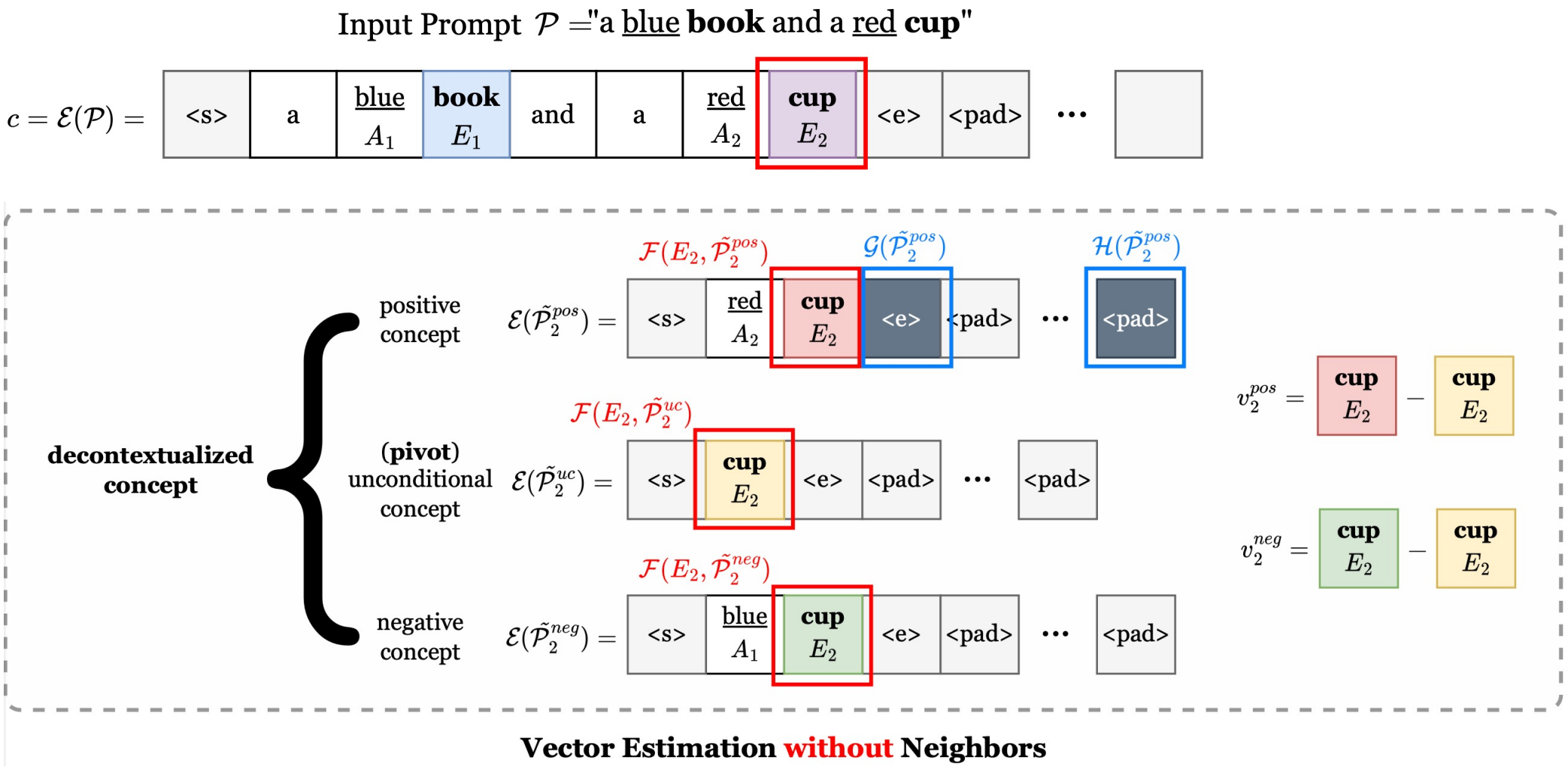
Method – Overview



(a) Generation pipeline (b) Adaptive strength (c) Neighbor-guided vector estimation

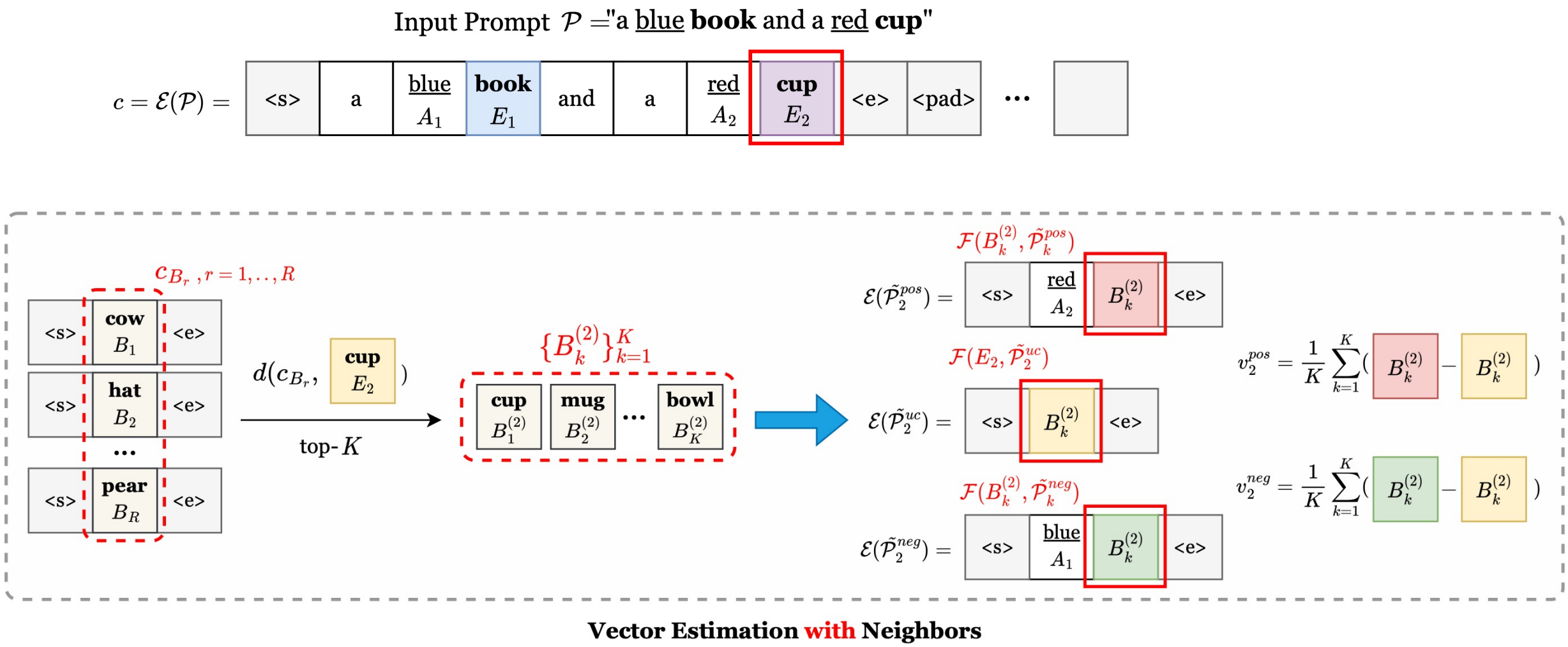
We introduce binding vectors, which can be applied on the object embedding to attract the target attribute and repulse other attributes, analogous to a **Magnet**, with adaptive strength, and neighbor-guided vector estimation strategies to improve robustness.

Method – Details



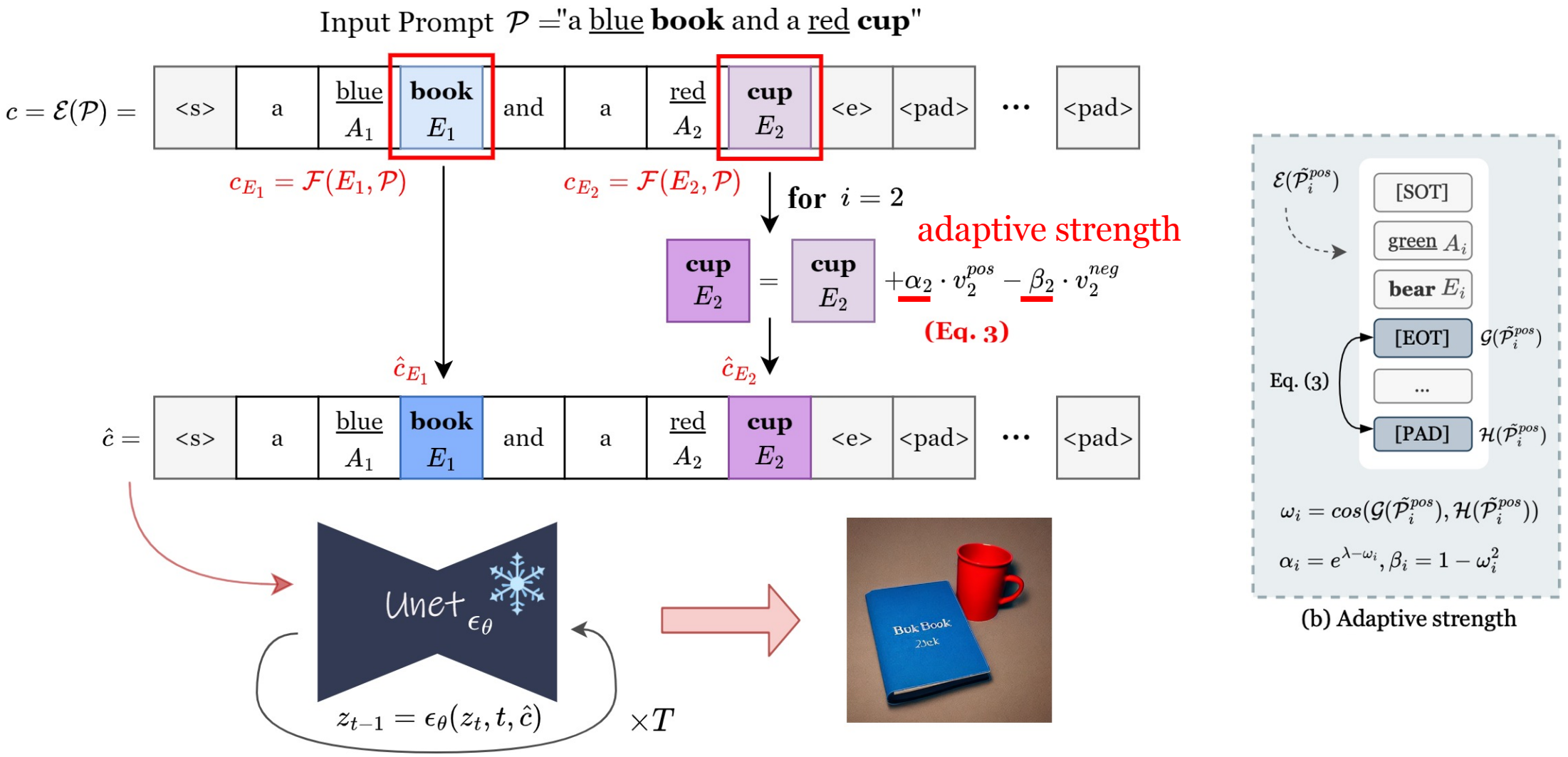
Vector Estimation without Neighbors

Method – Details



Vector Estimation with Neighbors

Method – Details



(b) Adaptive strength

Results – Magnet V.S. Baselines

Qualitative comparison (coarse-grained)

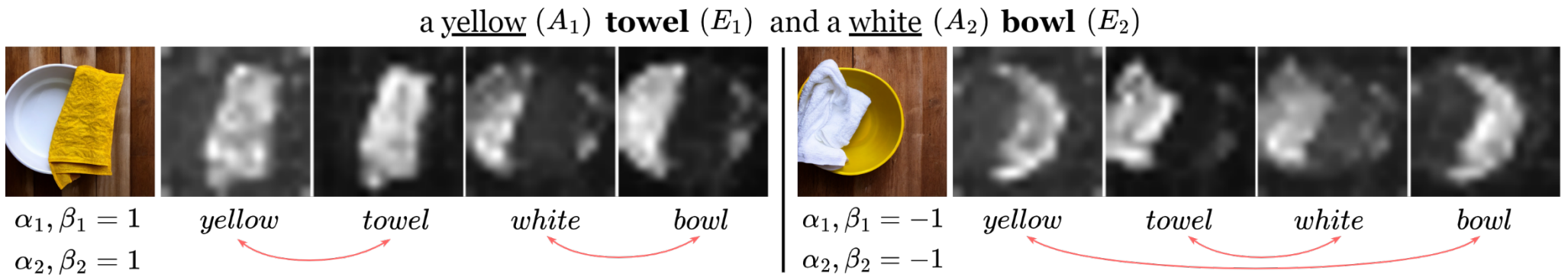
	ABC-6K			CC-500		
	Image Quality	Disentanglement Object	Disentanglement Attribute	Image Quality	Disentanglement Object	Disentanglement Attribute
Magnet (Ours)	26.57	25.71	27.14	25.43	24.86	29.43
Attend-and-Excite	15.43	21.43	19.71	22.86	26.29	18.57
Structure Diffusion	12.28	7.14	10.29	12.29	6.86	11.14
Stable Diffusion	10.29	6.57	8.57	11.14	7.71	13.42
No Winner	35.43	39.15	34.29	28.28	34.28	27.44

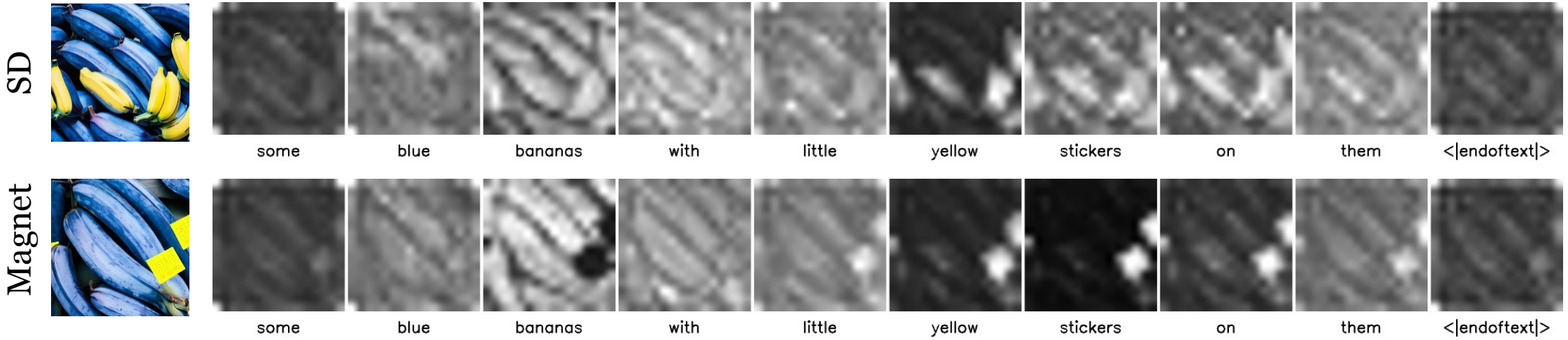
Qualitative comparison (fine-grained)

	Automatic		Manual		Runtime	Memory Usage
Method	Det.	Conf.	Obj.	Attr.	(s)	(GB)
Stable Diffusion	71.5	56.4	65.8	59.1	6.62	6.1
Structure Diffusion	72.1	56.0	64.0	63.9	7.94 (+20.0%)	7.0 (+83.5%)
Attend-and-Excite	84.3	62.6	84.6	66.2	13.4 (+102.4%)	15.6 (+155.7%)
Magnet (Ours)	76.5	59.8	68.6	74.0	6.81 (+2.9%)	6.5 (+1.0%)

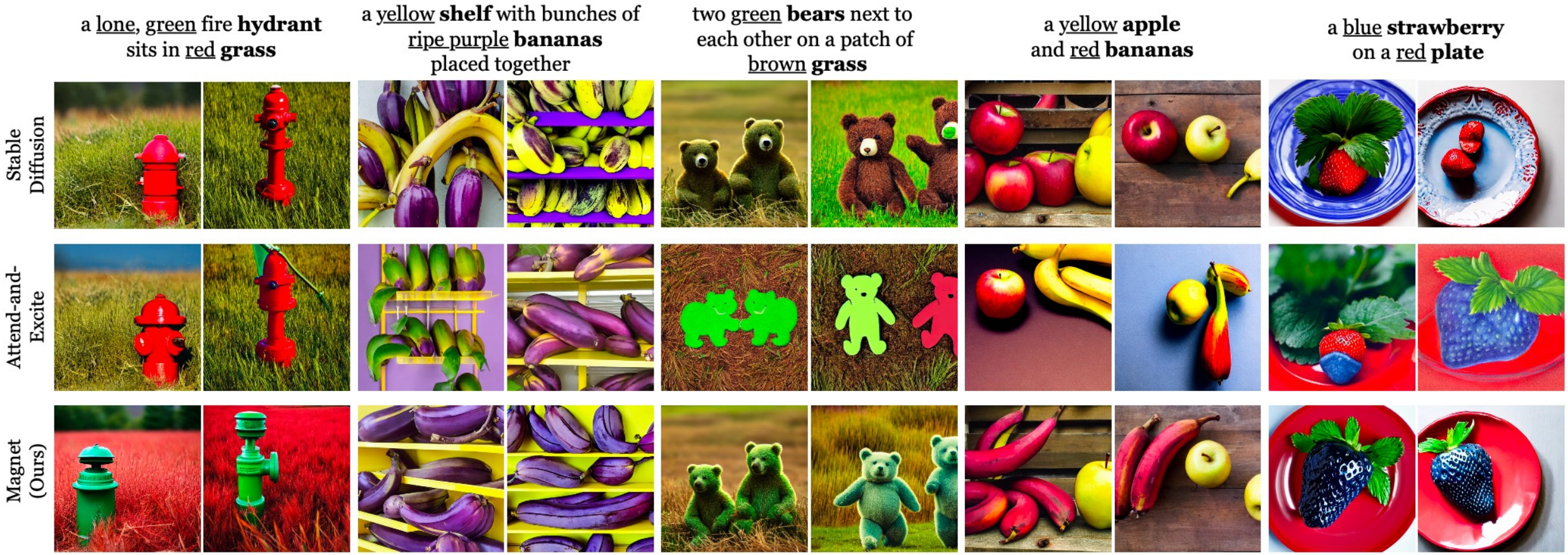
Qualitative comparison

Results – Cross-attention visualization



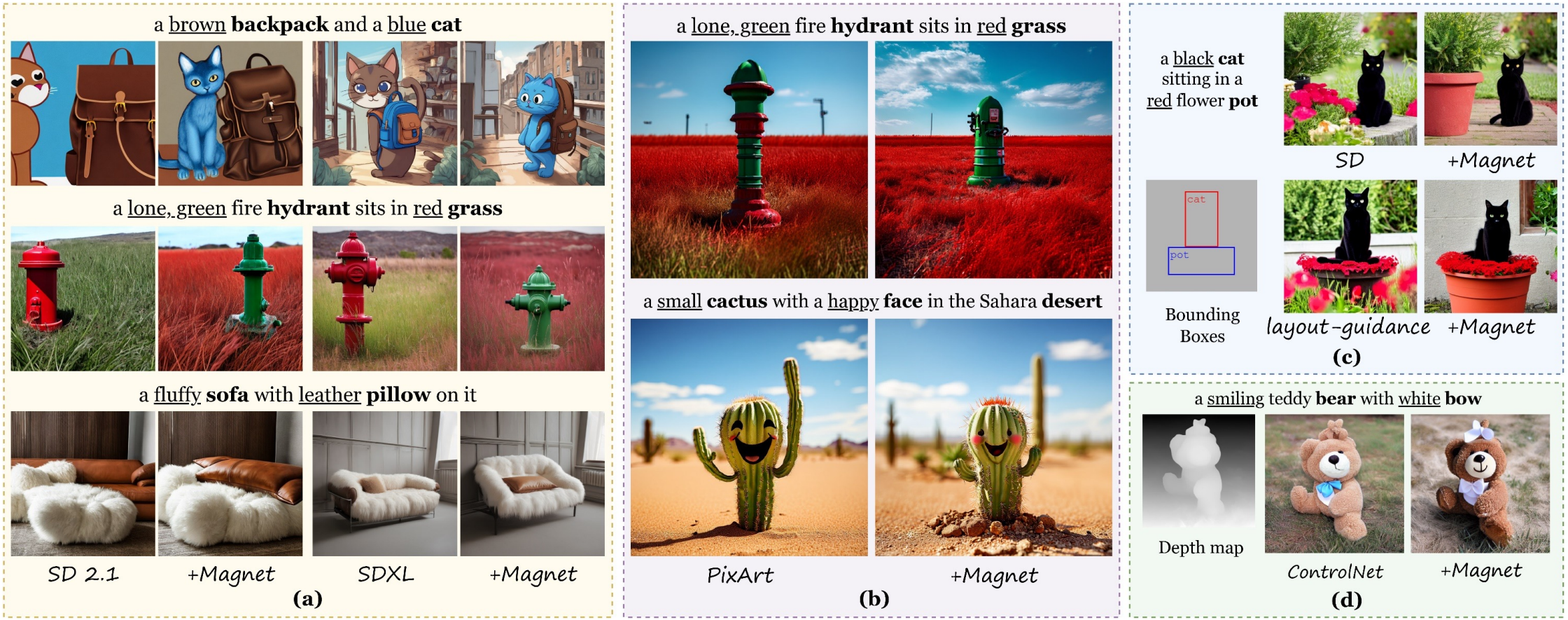


Results – Disentanglement



Magnet's *anti-prior* ability to generate unnatural concepts by enhancing disentanglement

Results – Compatibility



Magnet is ***plug-and-play*** to other T2I models and controlling modules

NEURAL INFORMATION PROCESSING SYSTEMS

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Thanks for your attention!

*Chenyi Zhuang[1], Ying Hu[1], Pan Gao[1,2]**

[1] Nanjing University of Aeronautics and Astronautics
[2] Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education

{chenyi.zhuang,ying.hu,pan.gao}@nuaa.edu.cn