Diagnosing and Correcting Concept Omission in Multimodal Diffusion Transformers

Kanghyun Baek[1], Jaihyun Lew[1], Chaehun Shin[1], Jungbeom Lee[2], Sungroh Yoon[1]
[1]Seoul National University, [2]Korea University

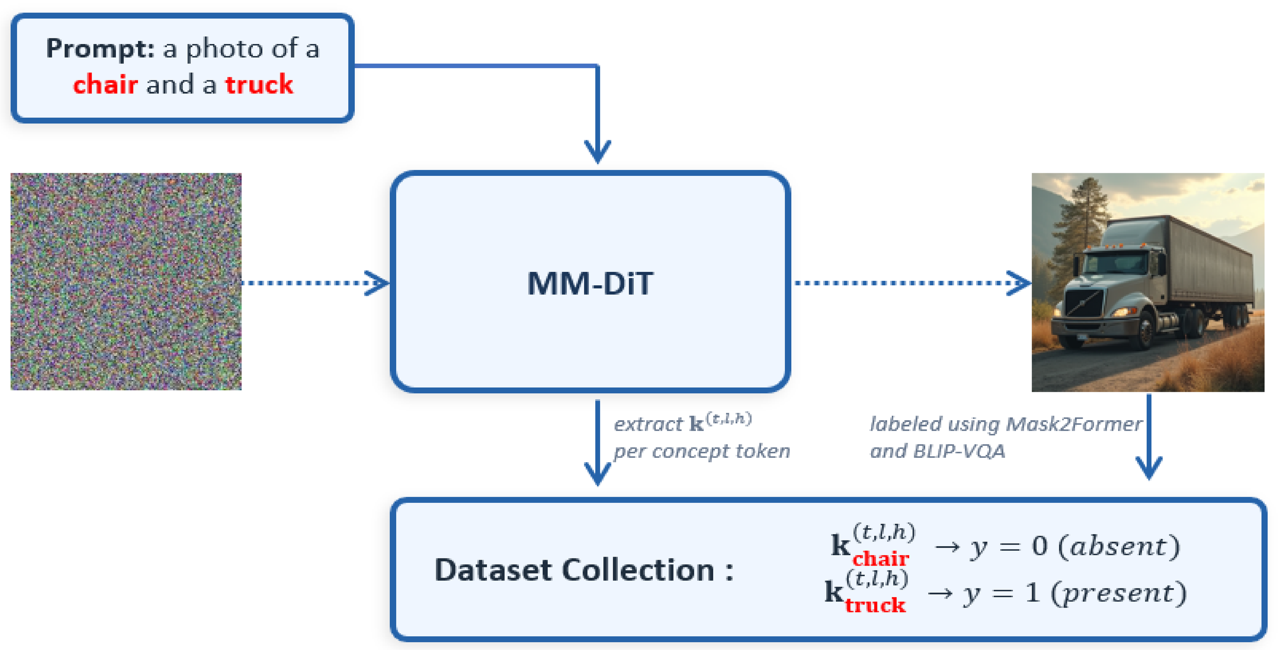
We reveal that text embeddings in MM-DiT internally encode concept omission status, and propose amplifying this signal to effectively recover omitted concepts.

Analysis on Concept Omission

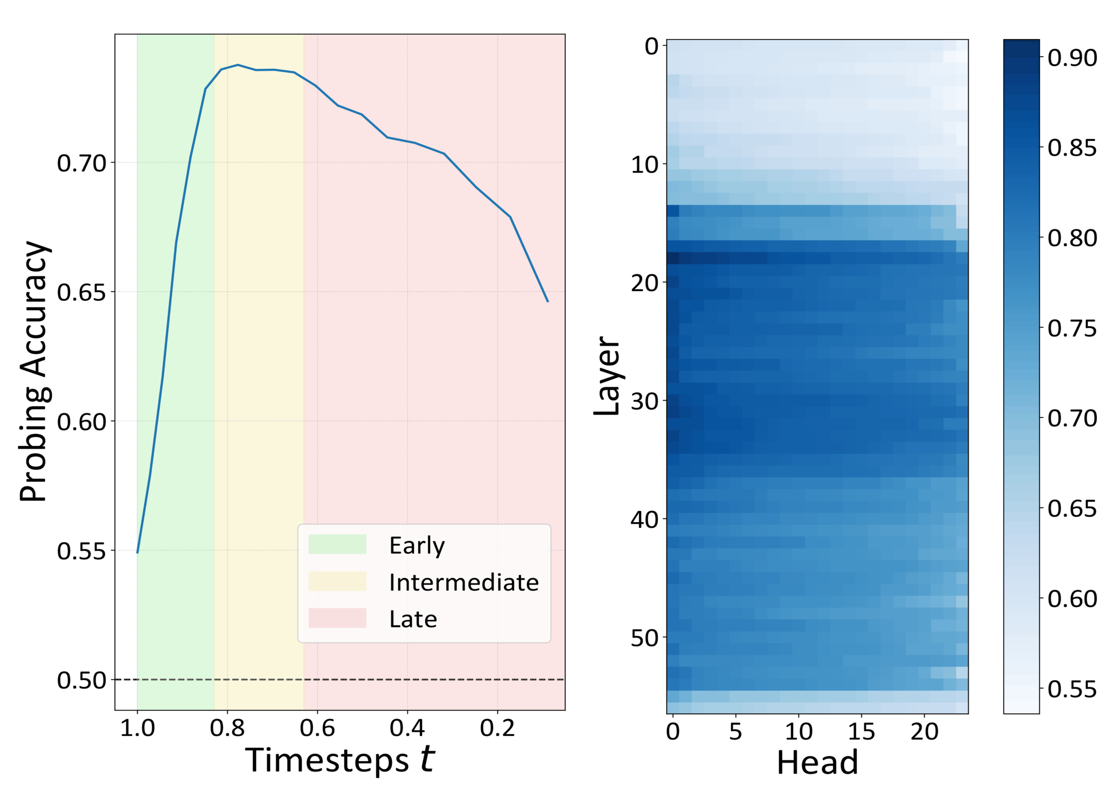
We investigate whether text token embeddings in MM-DiT encode information about concept omission status via linear probing.

1) Dataset Construction for Probing

- We train a linear probe on the key vector of each concept token at every attention head — predicting whether the concept appears in the final image.
- We generate images from two-object prompts, collecting the key vectors.
- Each object is labeled by an evaluation model on the generated images.

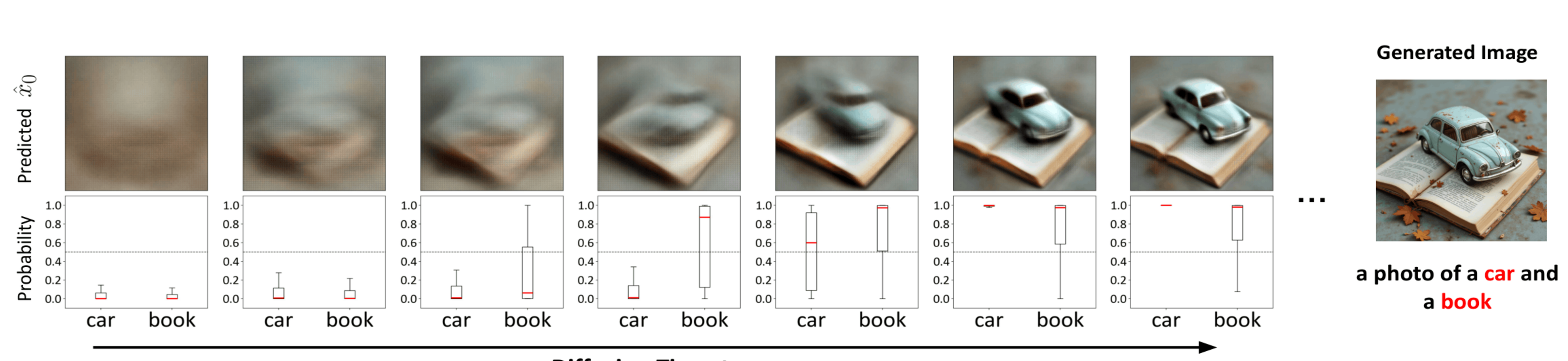


- We filter the collected data based on two criteria — (1) the alignment between labels and internal representations, and (2) the sufficiency of the signal encoded within those representations — and retain only the intermediate timesteps for training.



2) Analyzing Internal Representations of Omission

- Probes in the middle layers achieve up to 91.0% accuracy, confirming that concept tokens encode information about omission status.
- Probe predictions are tightly synchronized with the visual emergence of the corresponding concepts.



Problem & Motivation

- Text-to-image (T2I) generation has advanced rapidly with Multimodal Diffusion Transformers (MM-DiT). Despite remarkable improvements, these models still suffer from concept omission — including object omission and attribute neglect.
- Prior work has addressed this issue by focusing on visual embeddings, while text embeddings remain largely underexplored. We analyze concept omission from the perspective of text embeddings, leveraging the bidirectional text-image interaction in MM-DiT.

Omission Signal Intervention

We posit that amplifying the omission signal induces the model to perceive missing concepts as more critical — compelling it to generate them.

- We compute a normalized direction vector, given by the difference between the mean "absent" and "present" representations from the probing dataset.
- We then intervene on the concept token's key vector by adding a scaled multiple of this direction at inference time.
- To minimize unintended perturbations, we selectively apply intervention to the top-K heads with the highest probing accuracy, over the early timestep interval.

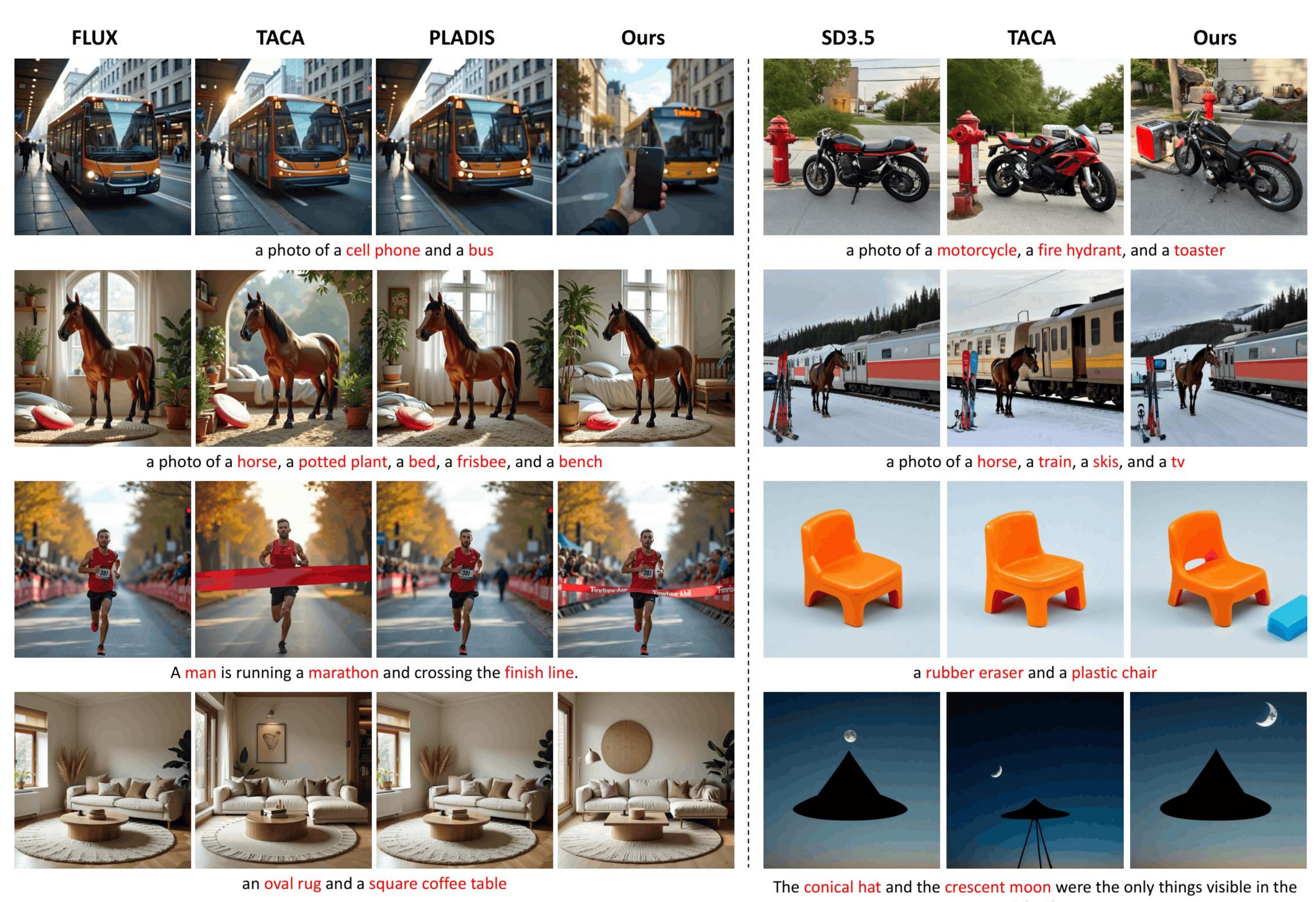
$$\delta^{(l,h)} = \mathbb{E}[\mathbf{k}^{(t,l,h)} \mid y=0] - \mathbb{E}[\mathbf{k}^{(t,l,h)} \mid y=1]$$

$$\theta^{(l,h)} = \frac{\delta^{(l,h)}}{\|\delta^{(l,h)}\|_2}$$

$$\mathbf{k}_c^{(t,l,h)} \leftarrow \mathbf{k}_c^{(t,l,h)} + \alpha\sigma^{(l,h)}\theta^{(l,h)}$$

Experiment

Across diverse multi-concept prompts, OSI consistently captures every specified object or attribute, while baselines and prior methods often omit some objects or attributes.



Our method achieves state-of-the-art performance across benchmarks with varying object counts (2–6), demonstrating strong generalization to complex multi-object scenarios.

Backbone	Method	GenEval						T2I-CompBench
		Two object	Three object	Four object	Five object	Six object	Avg.	Non-spatial
FLUX	Base	0.81	0.63	0.44	0.29	0.18	0.47	0.3069
	TACA	0.89	0.68	0.56	0.31	0.20	0.53	0.3078
	PLADIS	0.87	**0.71**	0.56	0.31	0.20	0.53	0.3075
	Ours	**0.92**	**0.71**	**0.64**	**0.40**	**0.40**	**0.61**	**0.3083**
SD3.5	Base	0.82	0.72	0.59	0.38	0.24	0.55	0.3155
	TACA	0.87	0.74	0.55	**0.47**	0.26	0.58	**0.3164**
	Ours	**0.89**	**0.80**	**0.68**	0.46	**0.35**	**0.64**	0.3159

Although our probes are trained on object-level omission labels, OSI generalizes to attribute neglect.

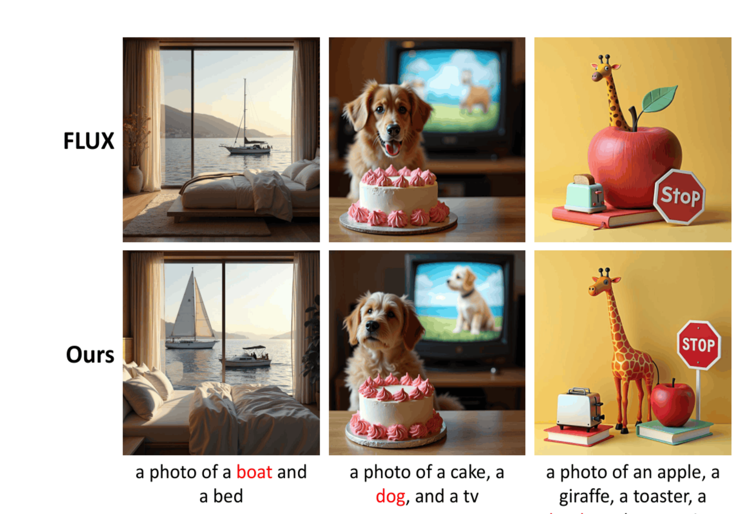
Backbone	Method	Color	Shape	Texture
FLUX	Base	0.7923	0.4995	0.6419
	TACA	0.7742	0.5118	0.6493
	PLADIS	0.7914	0.5108	0.6441
	Ours	**0.8014**	**0.5819**	**0.7039**
SD3.5	Base	0.7955	0.5820	0.7305
	TACA	**0.8159**	0.5948	0.7458
	Ours	0.8048	**0.6119**	**0.7480**

We conduct ablation studies on intervention direction and head selection. Our method achieves the best performance.

Setting		Two	Three	Four	Five	Six
Base (FLUX)		0.81	0.63	0.44	0.29	0.18
Direction (θ)	Opposite	0.72	0.40	0.15	0.06	0.02
	Random	0.84	0.65	0.53	0.32	0.30
Heads (K)	Bottom	0.81	0.60	0.47	0.22	0.15
	Random	0.82	0.67	0.55	0.31	0.17
	All	0.90	**0.74**	0.50	0.33	0.32
Ours		**0.92**	0.71	**0.64**	**0.40**	**0.40**

Limitations & Future Work

- Excessive intervention occasionally causes over-generation of certain concepts. Furthermore, the binary omission signal limits applicability to subjective or ambiguous concepts such as aesthetic qualities.
- Extending OSI to broader concept categories remains an open direction. Exploring its applicability to other generation domains, such as video and action generation, could also be a promising avenue.



Contact
Kanghyun Baek
qor6271@snu.ac.kr

Paper

Github Page