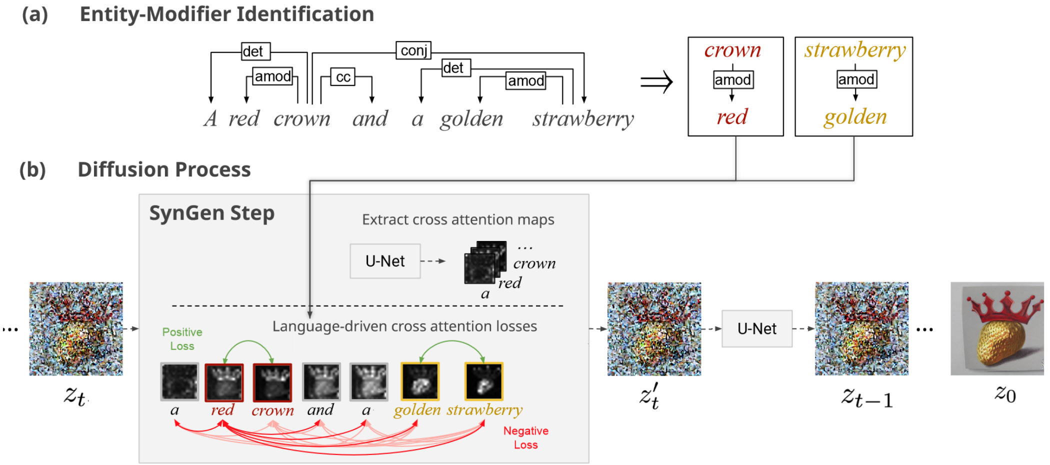
Token Merging for Training-Free Semantic Binding in Text-to-Image Synthesis

Taihang Hu[1], Linxuan Li[1], Joost van de Weijer[3], Hongcheng Gao[4]
Fahad Shahbaz Khan[5,6], Jian Yang[1], Ming-Ming Cheng[1,2], Kai Wang[3]*, Yaxing Wang[1,2]*
[1]VCIP, College of Computer Science, Nankai University, [2]NKIARI, Shenzhen Futian
[3]Computer Vision Center, Universitat Autònoma de Barcelona
[4]University of Chinese Academy of Sciences
[5]Mohamed bin Zayed University of AI, [6]Linkoping University

Introduction: Problem Definition

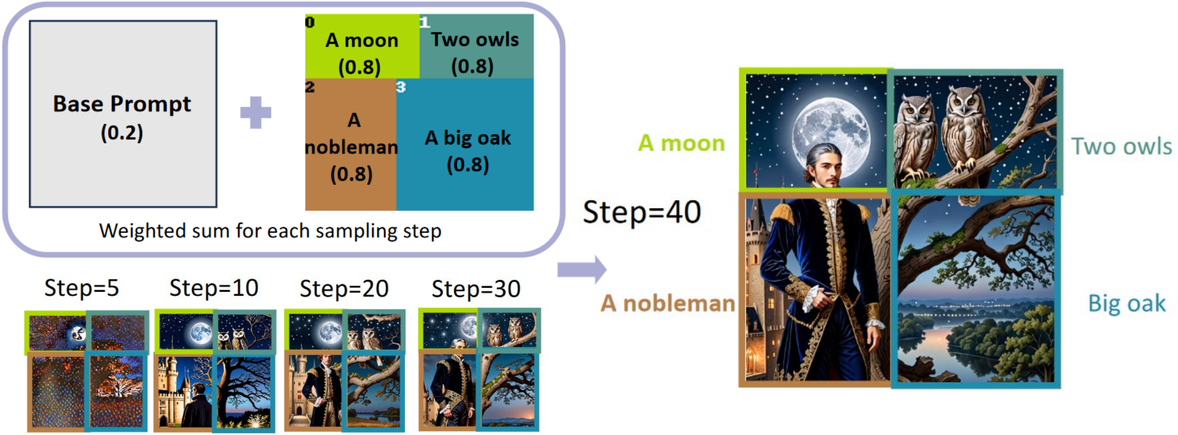
- Semantic Binding: associating an object with its attribute (attribute binding) or linking it to related sub-objects (object binding).
- Existing Issues: Incorrect binding and missing attributes.

Introduction: Related works



[1] Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map, NeurIPS 2023 oral

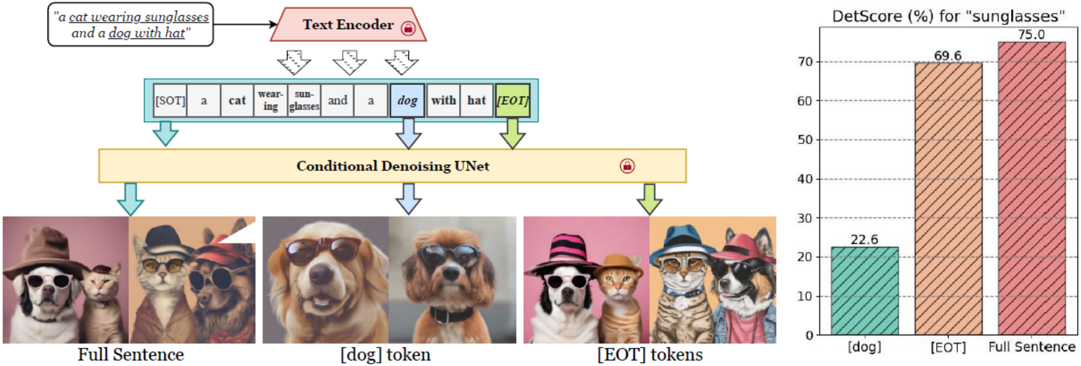
Introduction: Related works



[2] Mastering Text-to-Image Diffusion: Recaptioning, Planning, and Generating with Multimodal LLMs, ICML 2024

Motivation

- Text tokens exhibit information coupling, even a single token can couple with preceding information.
- EOT(End of Text) has the ability to contain all information.



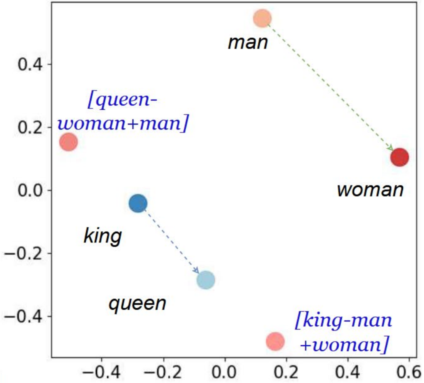
(a) T2I generation with various tokens (b) DetScore probabilities

Motivation

- Textual token embedding is additive, and the composite token obtained through element-wise addition has the ability to represent multiple objects.

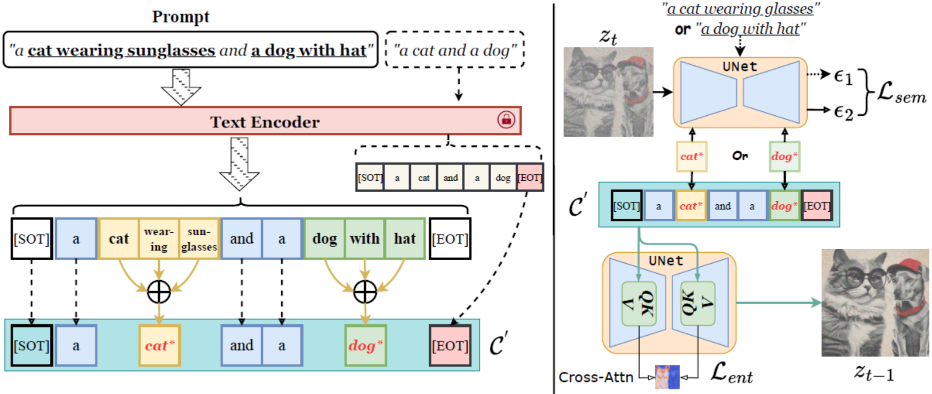
(a) Additivity Text-to-Image Generations

(b) PCA plot of text embeddings



Method

- Token merging: aggregating relevant tokens into a single composite token, aligning the object, its attributes, and sub-objects in the same cross-attention map



(a) Token Merging and End Token Substitution (b) Iterative Composite Token Update

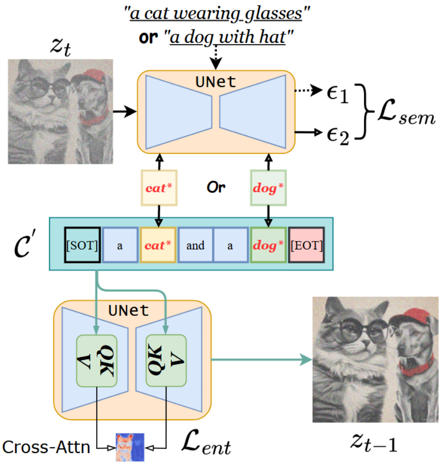
Method

- **Semantic binding loss**: Using a clean prompt as supervisory to eliminate any irrelevant semantic information within the composite token

$$\mathcal{L}_{sem} = \sum_{k\in[1,K]} \|\epsilon_\theta(z_t, \hat{c}_k, t) - \epsilon_\theta(z_t, \mathcal{C}, t)\|_2^2$$

- **Entropy loss**: Ensure that tokens focus exclusively on their designated regions, preventing the cross-attention map from becoming overly divergent

$$\mathcal{L}_{ent} = \sum_{k\in[1,K]} \sum_{p_i\in A_k} -p_i \log(p_i)$$



Experiments

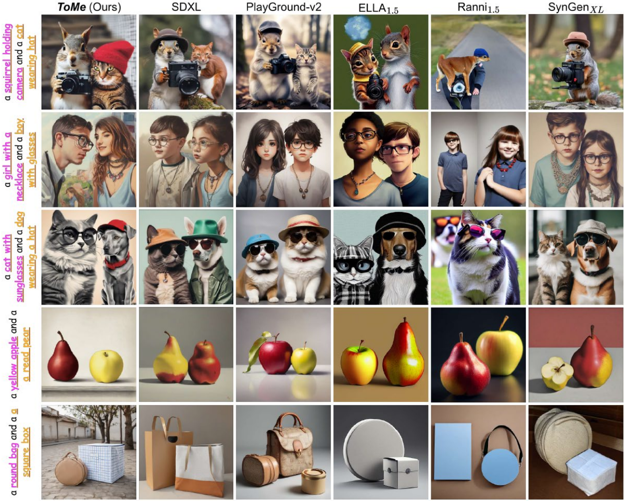


Figure 5: Qualitative comparison among various T2I generation methods with complex prompts.

Experiments

Table 1: Quantitative results for semantic binding assessment on various benchmarking subsets. We denote the best score in blue, and the second-best score in green.

Method	Base Model	Train	BLIP-VQA ↑ Color	BLIP-VQA ↑ Texture	BLIP-VQA ↑ Shape	Human-preference ↑ Color	Human-preference ↑ Texture	Human-preference ↑ Shape	GPT-4o ↑
SDXL[53]	-	✓	0.6369	0.5637	0.5408	0.7798	0.5140	0.4029	0.4907
PlayG-v2[37]	-	✓	0.6208	0.6125	0.5087	-	-	-	0.5417
Ranni[21]	SD1.5	✓	0.2414	0.3029	0.2857	-0.8554	-0.6853	-0.8051	0.4166
ELLA[30]	SD1.5	✓	0.6911	0.6308	0.4938	0.6586	0.2963	0.0565	0.6481
SynGen[58]	SD1.5	✗	0.6619	0.6451	0.4661	0.4326	0.5072	0.0426	0.5545
CoMat[34]	SD1.5	✓	0.6561	0.6190	0.4975	-	-	-	-
Ranni[21]	SDXL	✓	0.6893	0.6325	0.4934	-	-	-	-
ELLA[30]	SDXL	✓	0.7260	0.6686	0.5634	-	-	-	-
SynGen[58]	SDXL	✗	0.7010	0.6044	0.5069	1.016	0.7867	0.4016	0.6458
CoMat[34]	SDXL	✓	0.7774	0.6591	0.5262	-	-	-	-
ToMe (Ours)	SDXL	✗	0.7656	0.6894	0.6051	1.074	0.9281	0.5916	0.9549

Experiments

Experiments: Ablation Study

- Semantic binding loss filters out irrelevant information in tokens;
- Entropy loss makes the cross-attn map more focused.

Table 2: Ablation Study conducted on the T2I-CompBench benchmark.

Conf.	*ToMe*	$\mathcal{L}_{ent}$	$\mathcal{L}_{sem}$	BLIP-VQA Color	BLIP-VQA Texture	BLIP-VQA Shape
A	×	×	×	0.6369	0.5637	0.5408
B	✓	×	×	0.6577	0.5828	0.5437
C	✓	✓	×	0.7525	0.6775	0.5797
D	×	✓	✓	0.5881	0.6194	0.5386
E	×	✓	×	0.5983	0.5798	0.5125
F	✓	×	✓	0.6804	0.6263	0.5645
Ours	✓	✓	✓	**0.7656**	**0.6894**	**0.6051**

Figure 6: Text-to-Image generation with various configurations.

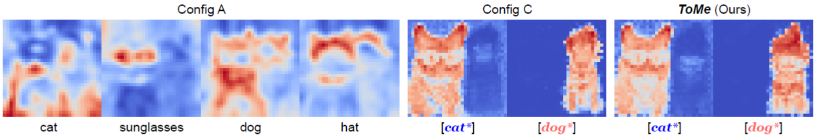


Figure 7: Cross-Attention maps visualization with various configurations.

Experiments: Additional Applications

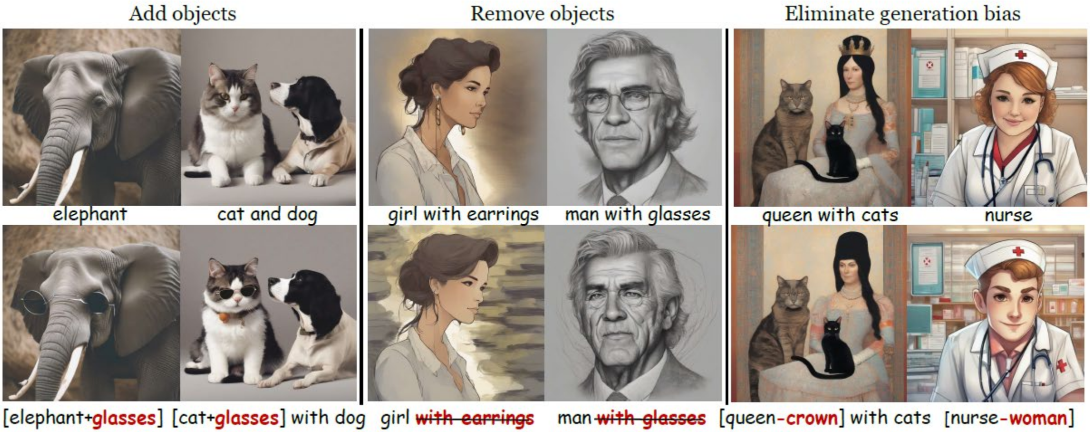


Figure 8: Additional applications of semantic additivity in text embedding.

Conclusion

- The paper introduces Token Merging (ToMe), a training-free method that enhances semantic binding without the need for additional datasets, large language models, or extensive fine-tuning.
- By merging tokens for objects and their attributes into a single composite token, ToMe ensures that the generated image maintains coherent cross-attention, aligning the visual output closely with the intended semantics of the text prompt.

Code: https://github.com/hutaihang/Tome