

ISAC: Training-Free Instance-to-Semantic Attention Control for Multi-Instance Generation

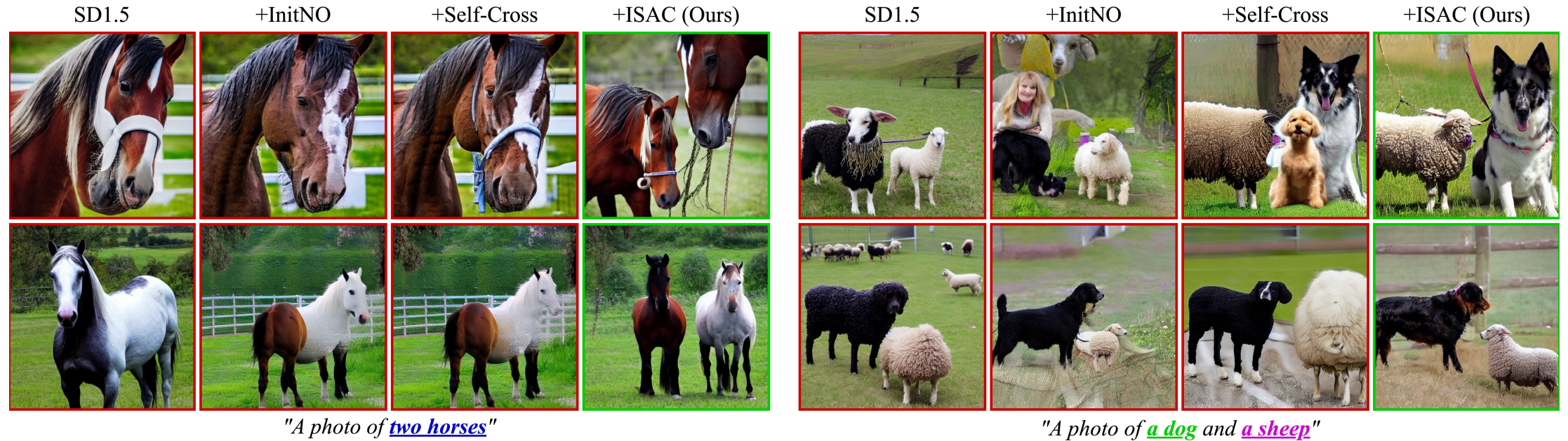
Sanghyun Jo^{1,2*†}, Wooyeol Lee^{2*}, Ziseok Lee^{2*}, Jonghyun Choi²,
Jaesik Park², Kyungsu Kim^{2†}

¹OGQ, Seoul, Korea

²Seoul National University, Seoul, Korea



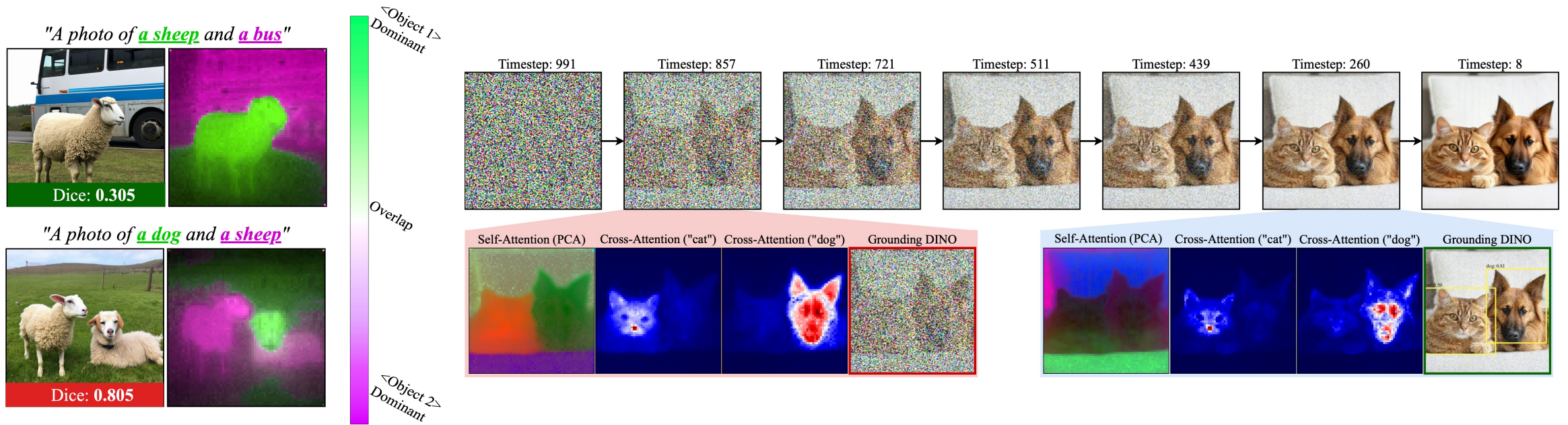
Existing Methods Merge Instances and Mix Attributes



Base diffusion models and existing training-free guidance still exhibit:

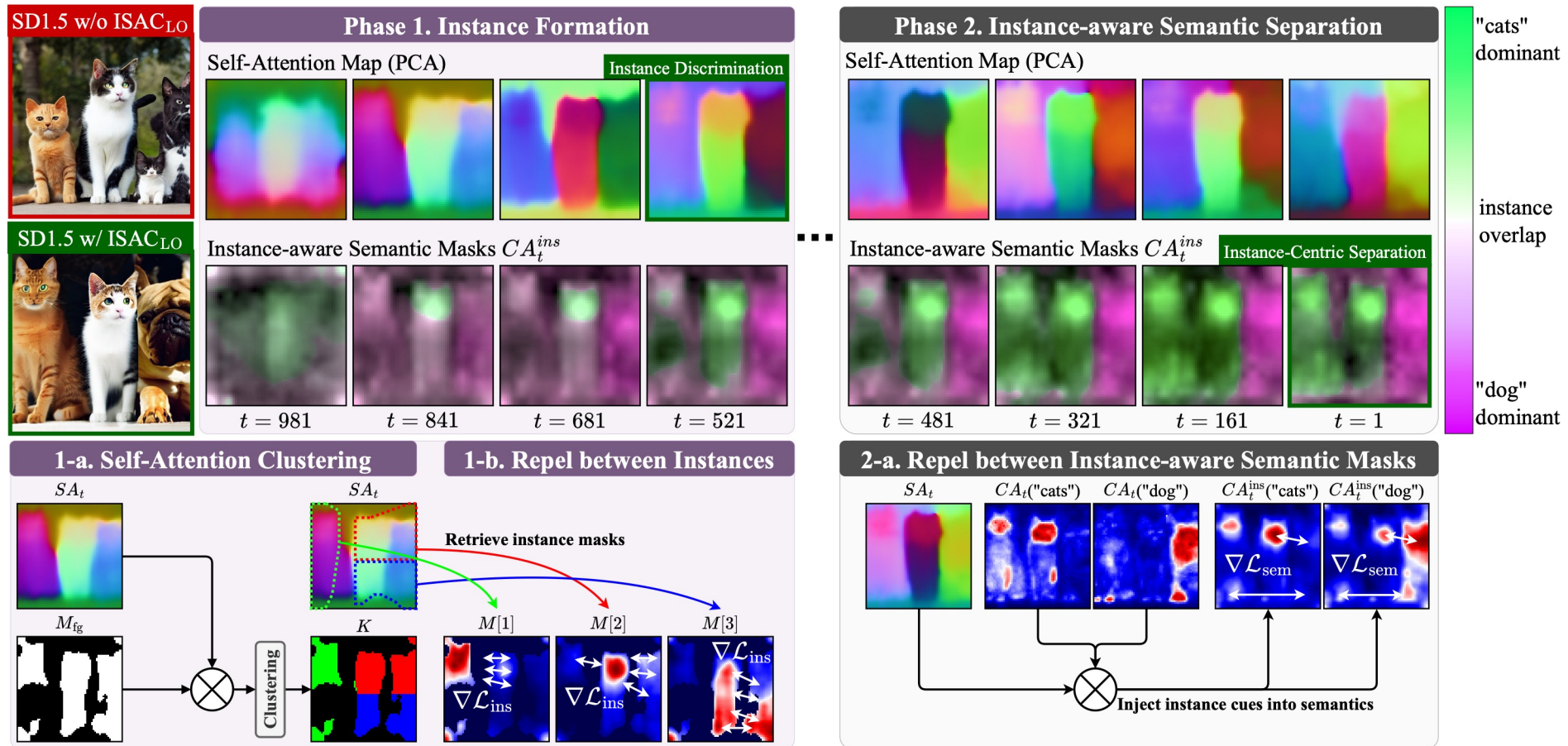
- (1) **count failures**: models often omit or merge instances
- (2) **semantic mixing**: attributes leak across instances

Instance Structure Should Be Stabilized Before Semantics



- Similar semantics produce strongly overlapping masks obtained from cross-attention (Dice 0.305 $\rightarrow$ 0.805).
- When one semantic mask covers multiple objects, token-level guidance, which the prior works belong to, cannot determine which pixels belong to which instance.
- Early self-attention already reveals class-agnostic instance layouts.
- Natural design principle: stabilize the class-agnostic instance structure first, and then bind semantic tokens to each instance.

ISAC Separates Instances First, Then Binds Semantics



1. **Structure first.** Cluster self-attention into N instance masks.
2. **Semantics second.** Bind tokens within each instance and repel semantics across instances.
3. **Training-free, backbone-agnostic.** Optimize only the latent. Model weights remain frozen. Can be easily applied to various backbones.

ISAC improves multi-class generation in a backbone agnostic manner

Backbone	Metric	Base	Best Prior	ISAC
SD1.5	HRS-Bench Spatial	0.094	0.135	0.263
	IntraCompBench Multi-Class Avg.	8%	20%	36%
SD3.5-M	HRS-Bench Spatial	0.264	0.279	0.350
	IntraCompBench Multi-Class Avg.	25%	34%	52%

Best prior is selected per metric; see the paper for full results.

Results: Multi-Instance Generation Examples

SD1.5 +ISAC_{LO} (Ours) +InitNO +Self-Cross SD3.5-M +ISAC_{LO} (Ours) +InitNO +Self-Cross



"A photo of a broccoli, a hot dog, and a pizza."



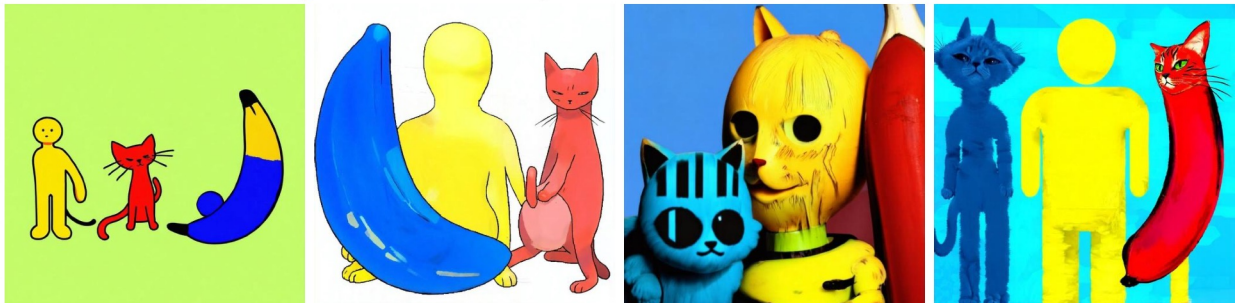
"a orange horse and a green car."



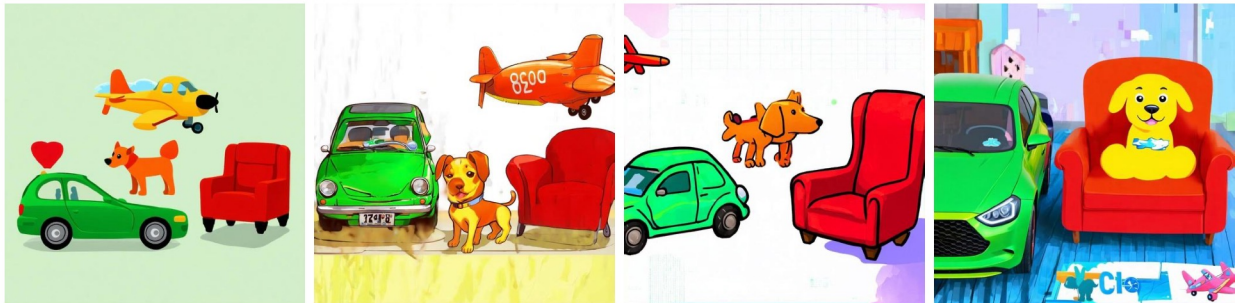
"a red banana and a green suitcase"



"A photo of a skateboard, a snowboard, and a tennis racket."



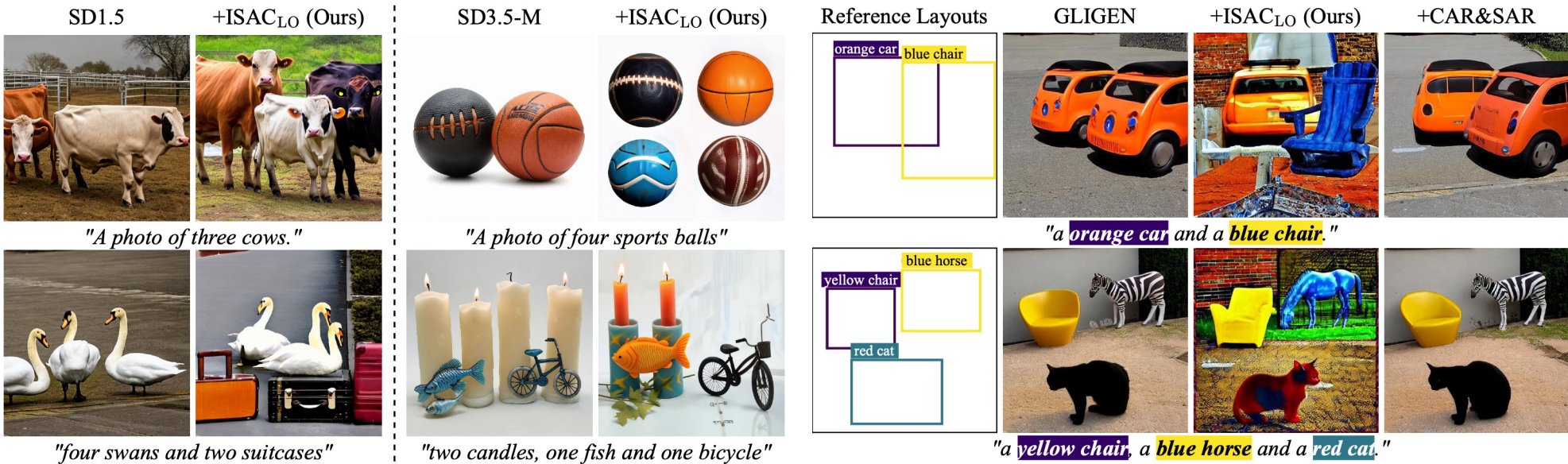
"a yellow person, a red cat, and a blue banana."



"a green car, a red chair, a orange airplane, and a yellow dog."

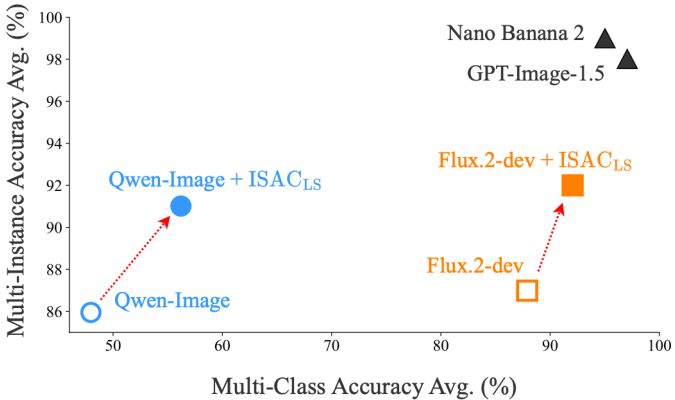
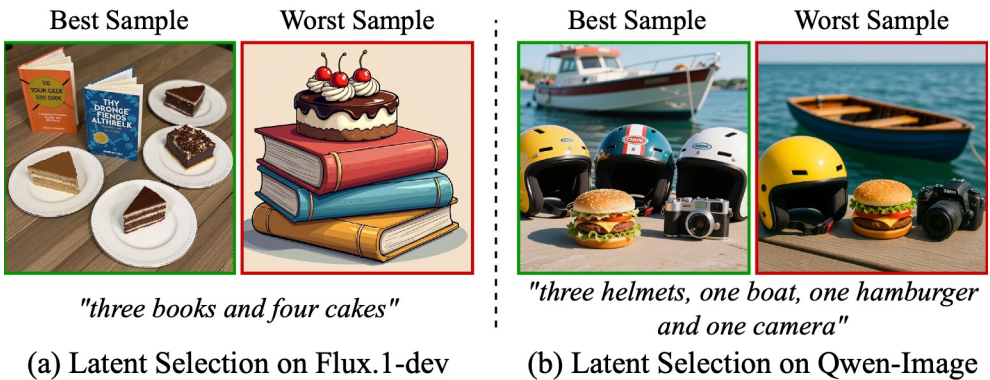
Results: Broader Applicability

Count Control & Layout-to-Image



Latent Selection for Large Backbones

ISAC narrows the gap between open-weight models and commercial models.



Conclusion

Separate instances first. Bind semantics second.

ISAC enables training-free, multi-instance generation with improved count and semantic separation.

More details are available via the QR codes.

