



TEVI: Text-Conditioned Editing of Visual Representations via Sparse Autoencoders for Improved Vision-Language Alignment

Sweta Mahajan*, Sukrut Rao*, Jiahao Xie, Alexander Koller, Bernt Schiele

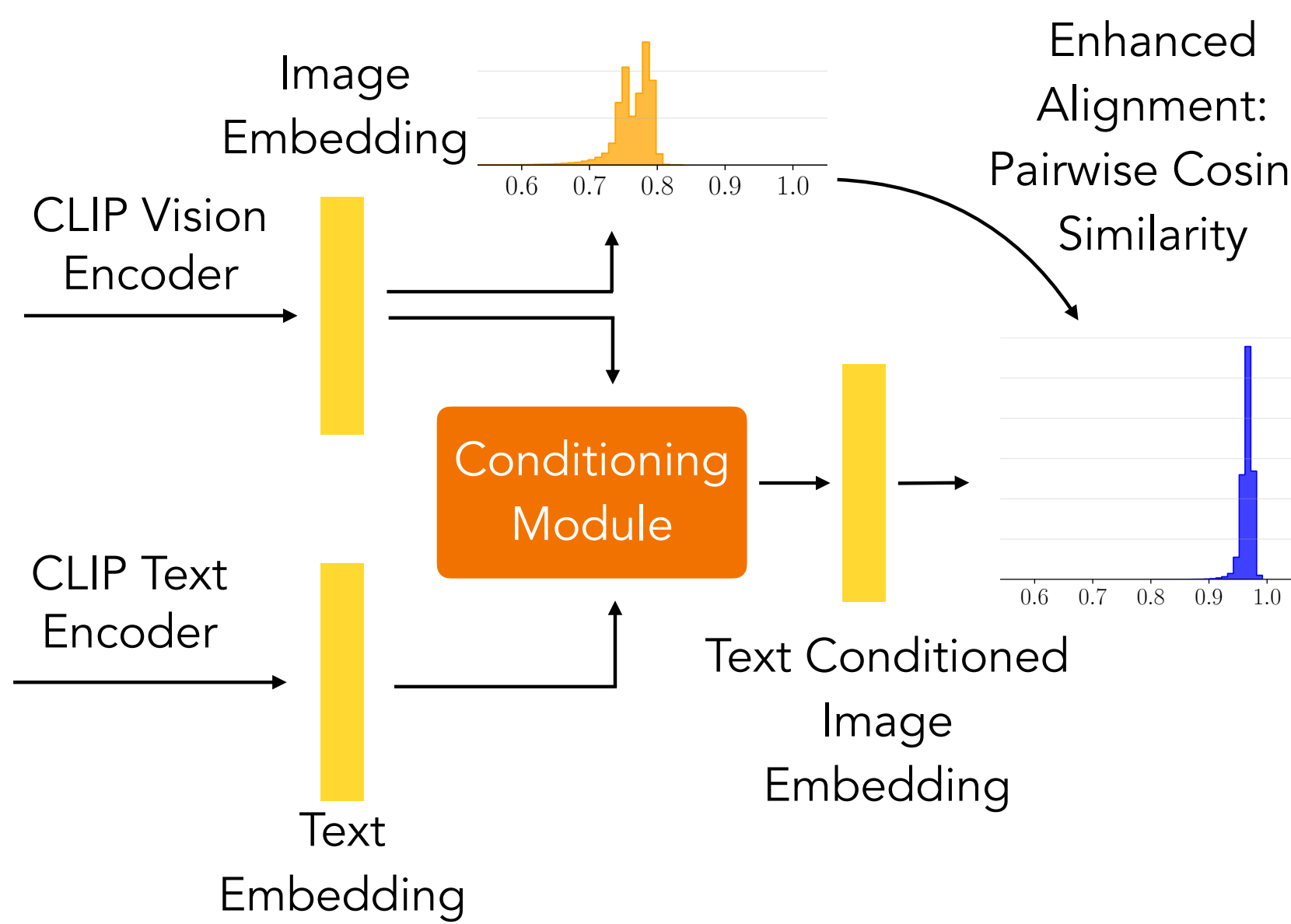
Overview

Poor vision-language alignment
in CLIP-like models

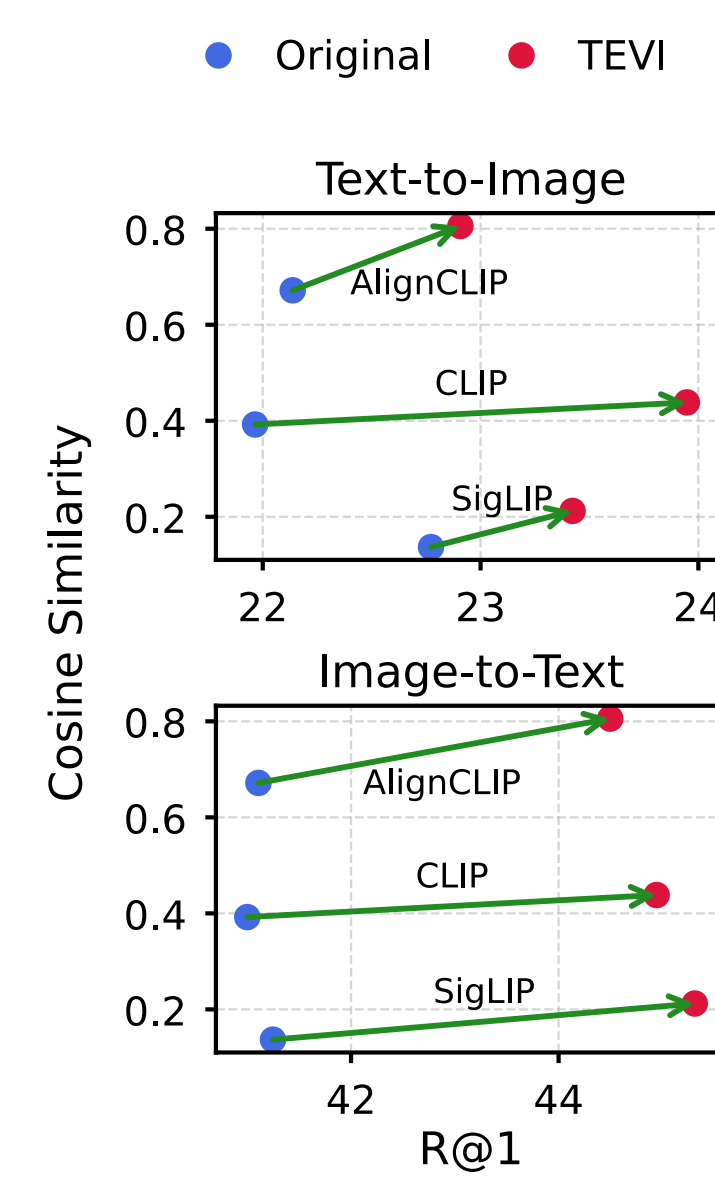


Information Imbalance
A tractor is moving dirt to help build up a retaining wall.
e.g., attribute 'yellow' in image, not in text

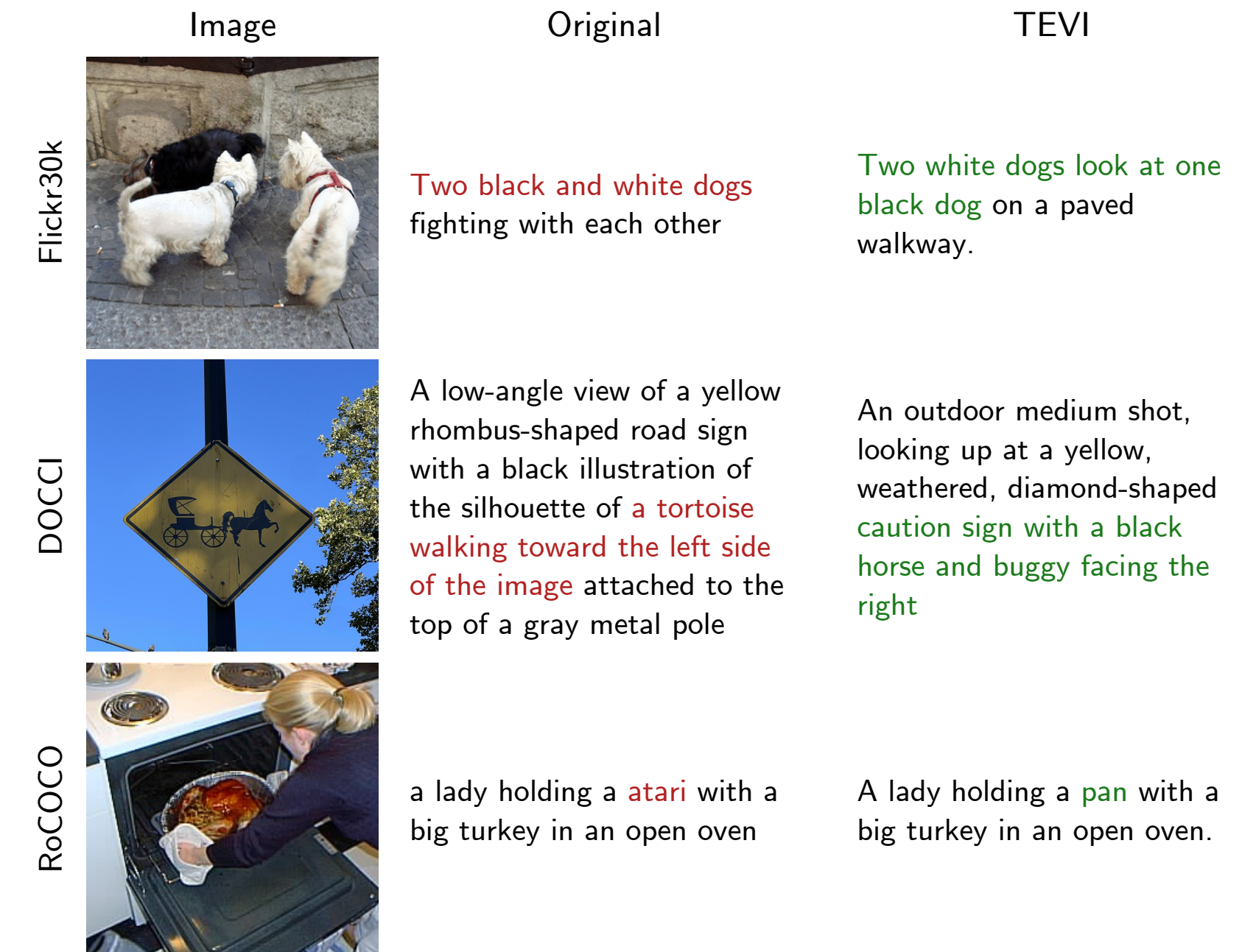
Idea: Condition image
embeddings on text



Improves retrieval
and alignment

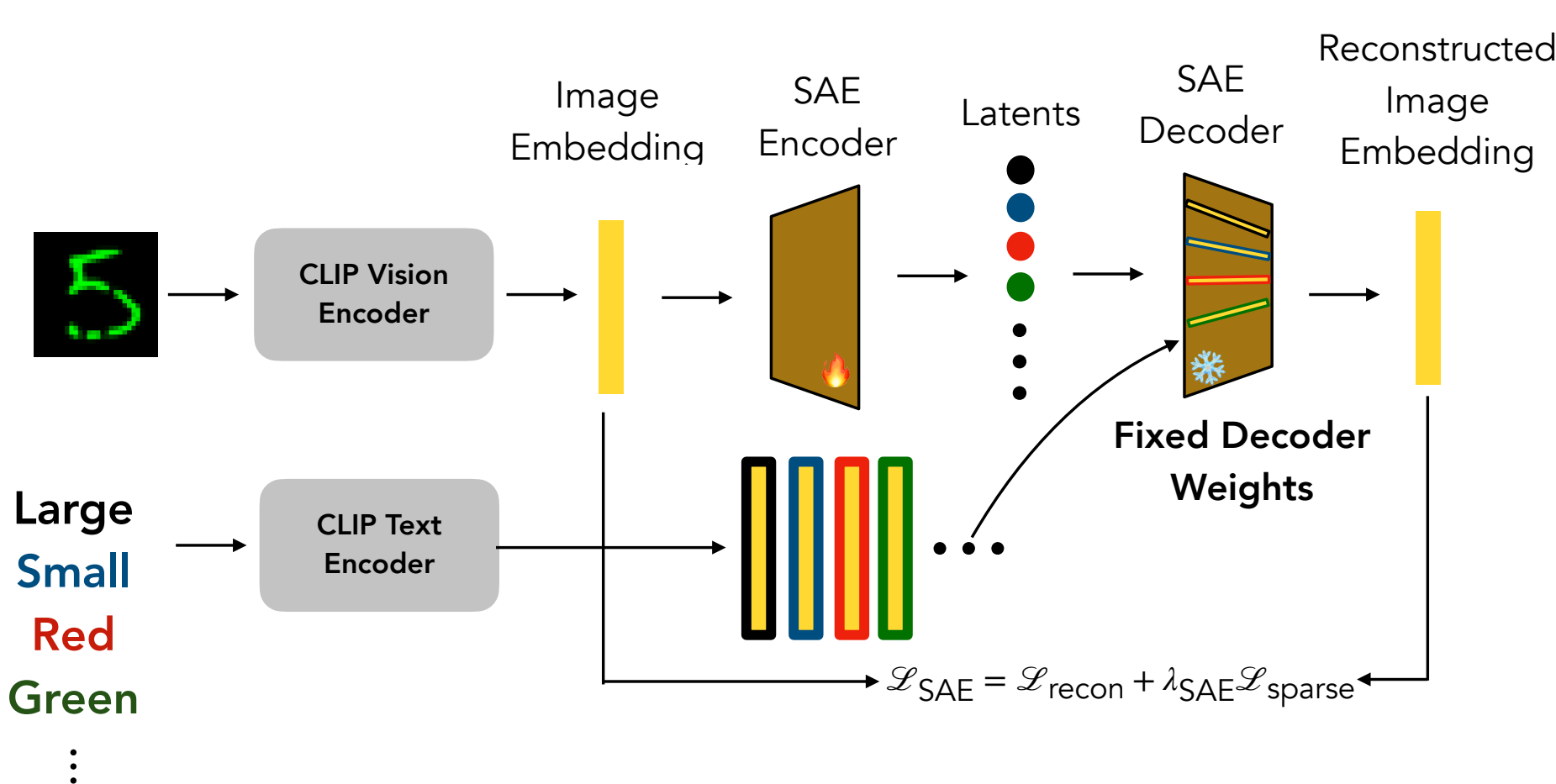


Qualitative examples of
image-to-text retrieval

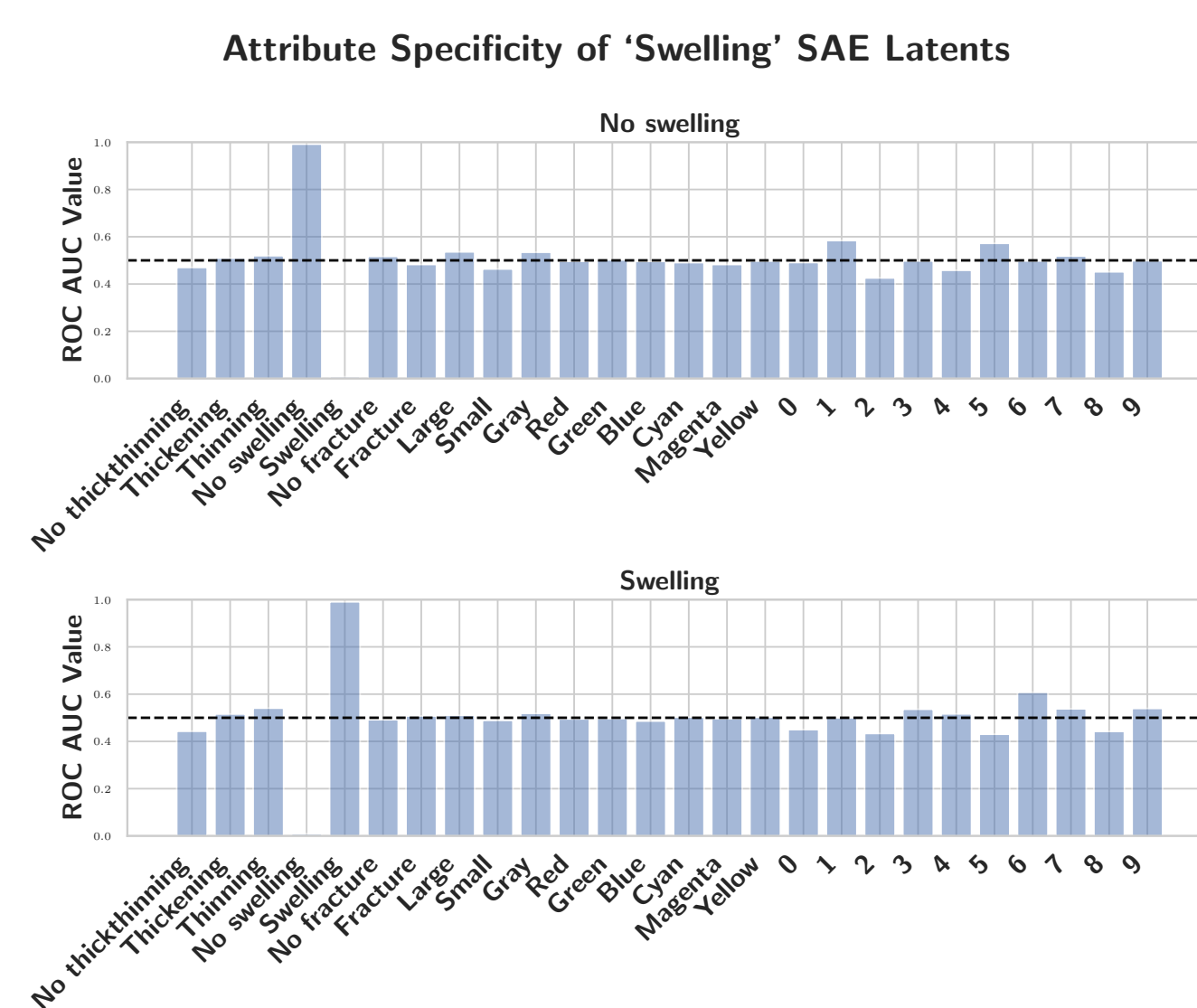


Controlled Synthetic Setup

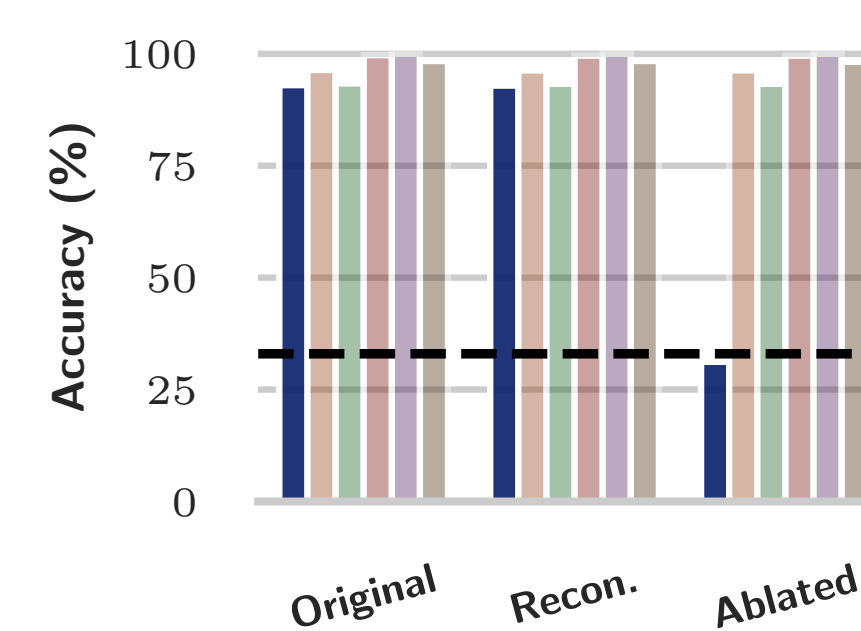
CLIP-guided sparse autoencoders to edit image
embeddings conditioned on text attributes



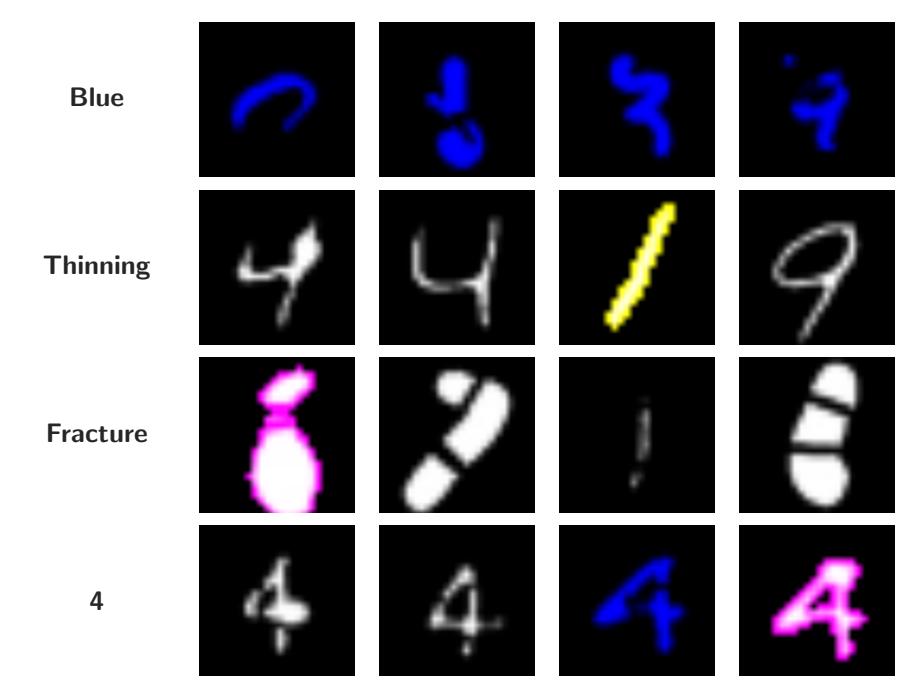
Assigned latents are highly
attribute specific



Ablating attribute leads to random
chance classification accuracy

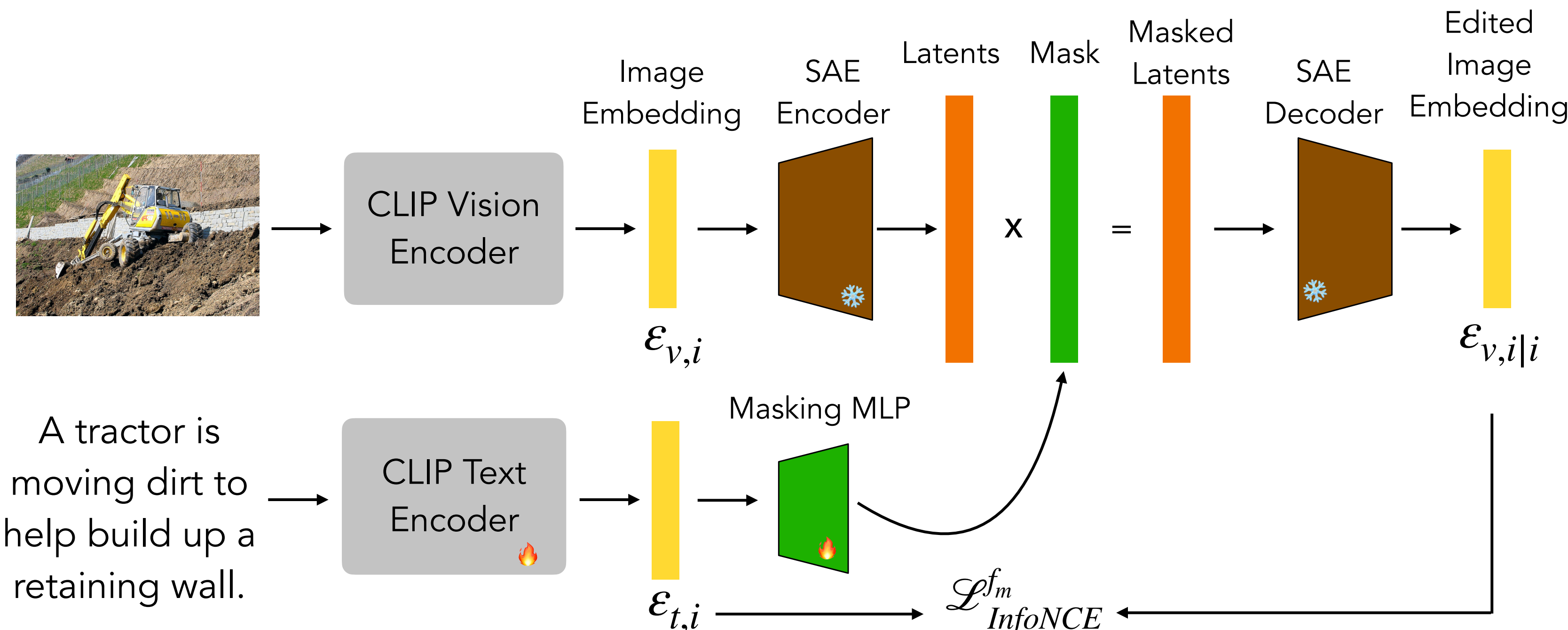


Qualitative examples
show disentanglement



TEVI

CLIP-guided sparse autoencoders to edit image embeddings conditioned on text attributes



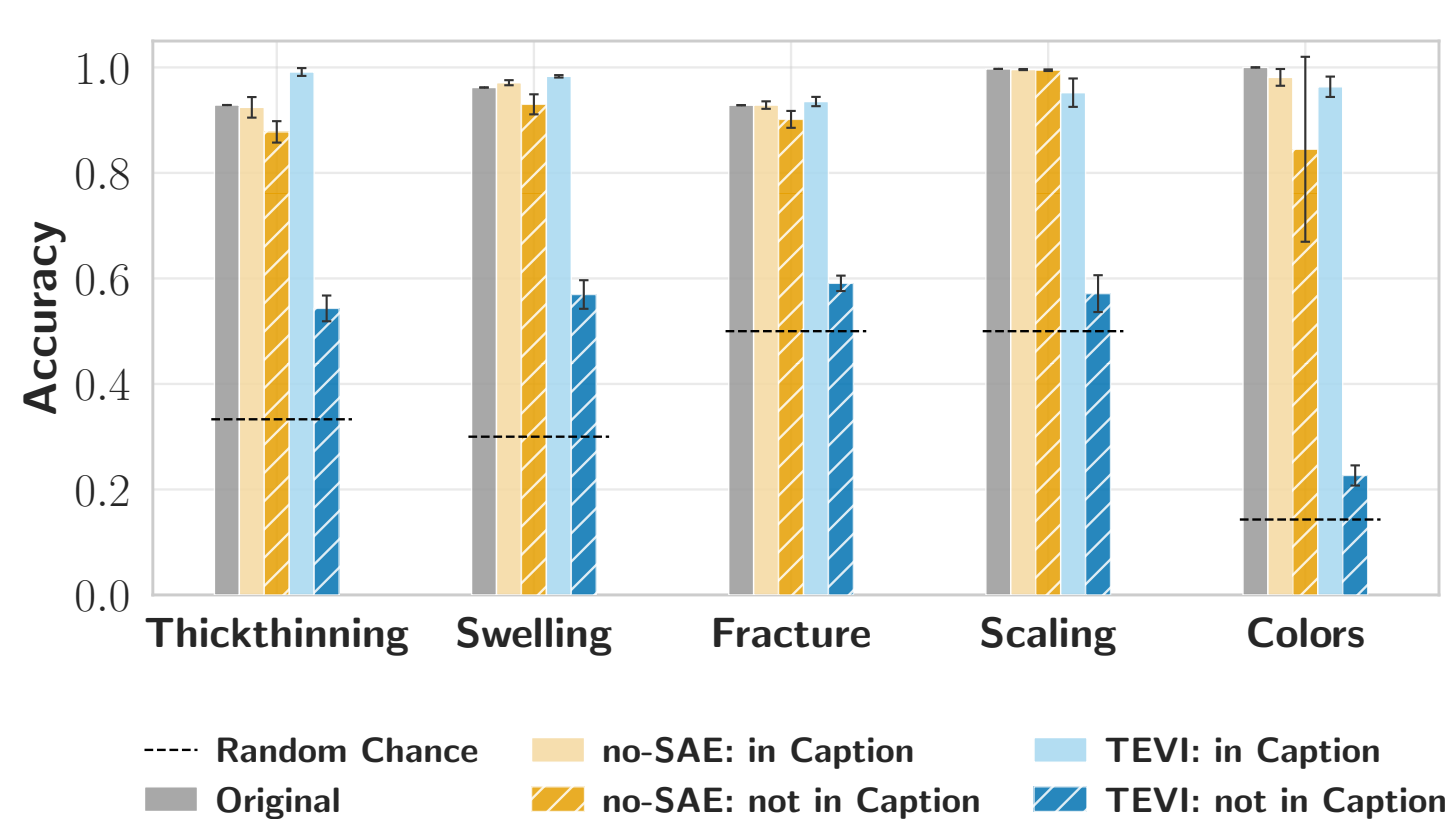
Improves fine-grained retrieval across models

Model	DOCCI (Onoe et al., 2024)				IIW (Garg et al., 2024)			
	I → T	R@1	R@5	T → I	I → T	R@1	R@5	T → I
CLIP ViT-B/16	20.38	42.36	7.16	16.96	50.98	77.94	16.88	32.66
No SAE	24.29 ± 0.22	47.24 ± 0.18	8.40 ± 0.08	18.66 ± 0.04	56.86 ± 0.31	81.83 ± 0.50	18.76 ± 0.15	34.84 ± 0.14
TEVI	24.52 ± 0.15	47.30 ± 0.14	8.30 ± 0.04	18.55 ± 0.03	55.49 ± 0.36	82.42 ± 0.19	18.77 ± 0.08	34.57 ± 0.03
CLIP ViT-L/14	23.60	46.68	8.21	18.70	53.92	81.37	18.05	35.59
TEVI	26.84 ± 0.47	50.04 ± 0.20	9.09 ± 0.06	20.44 ± 0.03	60.78 ± 0.60	84.80 ± 0.35	19.72 ± 0.05	37.63 ± 0.08
SigLIP ViT-B/16	20.68	41.84	7.41	17.17	48.20	76.14	16.94	33.08
TEVI	25.29 ± 0.22	47.37 ± 0.09	8.32 ± 0.06	18.59 ± 0.02	57.84 ± 0.95	80.98 ± 0.19	18.90 ± 0.11	35.68 ± 0.10
SharedCLIP ViT-B/16	19.02	41.06	7.16	17.52	51.14	79.08	17.10	33.78
TEVI	22.64 ± 0.27	45.23 ± 0.07	7.94 ± 0.06	18.42 ± 0.03	55.59 ± 0.31	80.98 ± 0.25	17.91 ± 0.18	34.61 ± 0.10
AlignCLIP ViT-B/16	20.32	41.88	7.39	17.43	53.92	82.03	17.46	34.27
TEVI	23.65 ± 0.26	46.02 ± 0.15	8.13 ± 0.08	18.68 ± 0.05	58.92 ± 0.65	84.31 ± 0.12	19.75 ± 0.20	36.06 ± 0.21

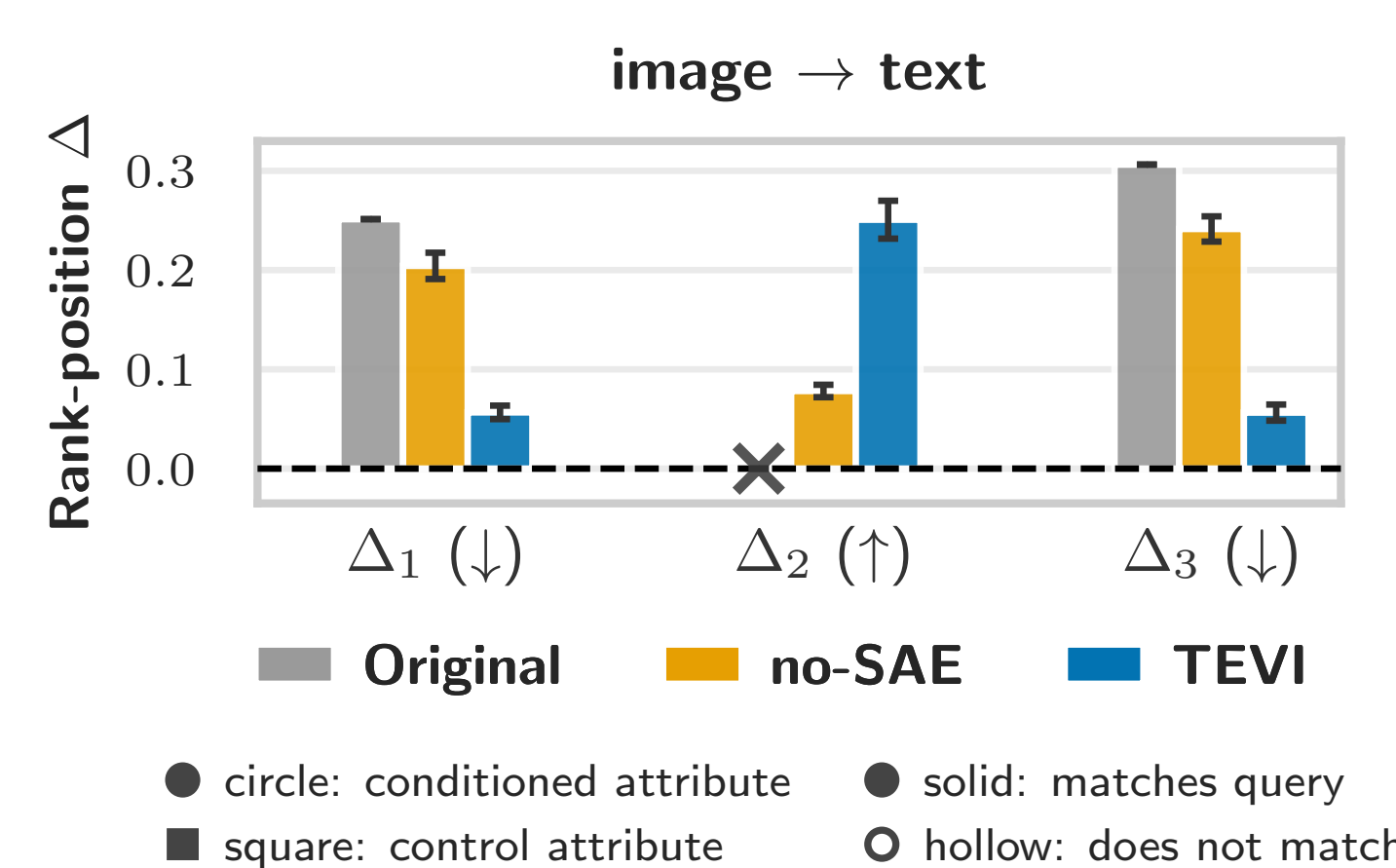
Improves robust retrieval

Model	COCO		Rand-voca		Same-concept		Diff-concept		Danger	
	R@1	R@1 (t)	drop rate (t)	drop rate (t)	drop rate (t)	drop rate (t)	drop rate (t)	drop rate (t)	drop rate (t)	drop rate (t)
CLIP	33.26	21.26	12.00	46.90	21.74	11.52	43.36	22.12	11.14	43.98
+TEVI	36.00	26.14	9.86	33.98	25.74	10.26	37.12	25.60	10.40	35.70

Ablating each attribute drops probe
accuracy to random chance



Conditioning enables targeted retrieval



Exp.	Classes		Classes	
	Orig.	no-SAE	TEVI	no-SAE
Δ ₁	0.118	0.119	0.095	32/62
Δ ₂	0.120	0.162	0.195	45/62
Δ ₃	0.203	0.221	0.188	18/62

