



MAX PLANCK INSTITUTE
FOR INFORMATICS

SIC Saarland Informatics
Campus

NEURO
EXPLICIT
MODELS



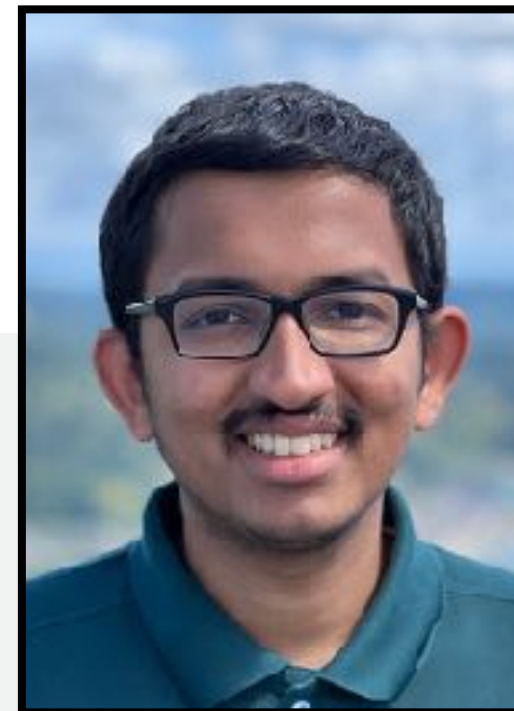
UNIVERSITÄT
DES
SAARLANDES

EMNLP2026
BUDAPEST
OCTOBER 24-29

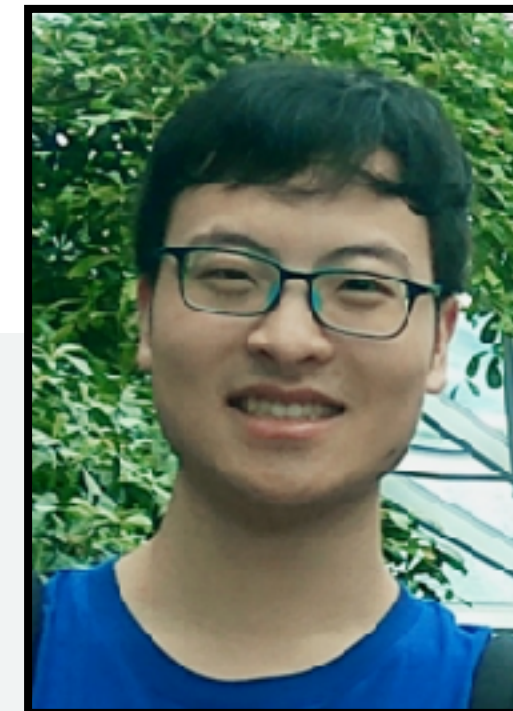
TEVI: Text-Conditioned Editing of Visual Representations via Sparse Autoencoders for Improved Vision-Language Alignment



Sweta Mahajan*



Sukrut Rao*



Jiahao Xie



Alexander Koller



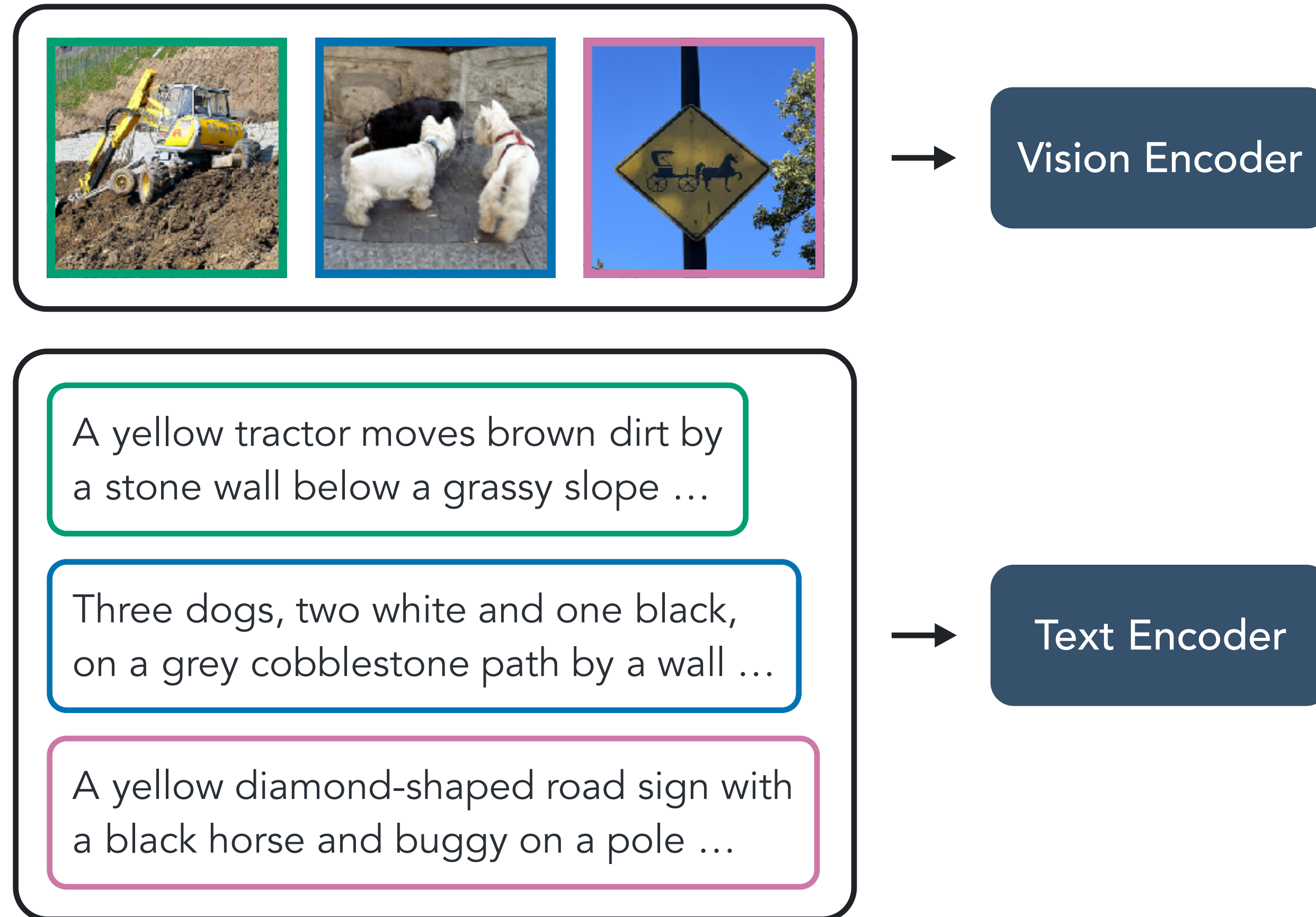
Bernt Schiele

CLIP aligns images and text in a shared embedding space

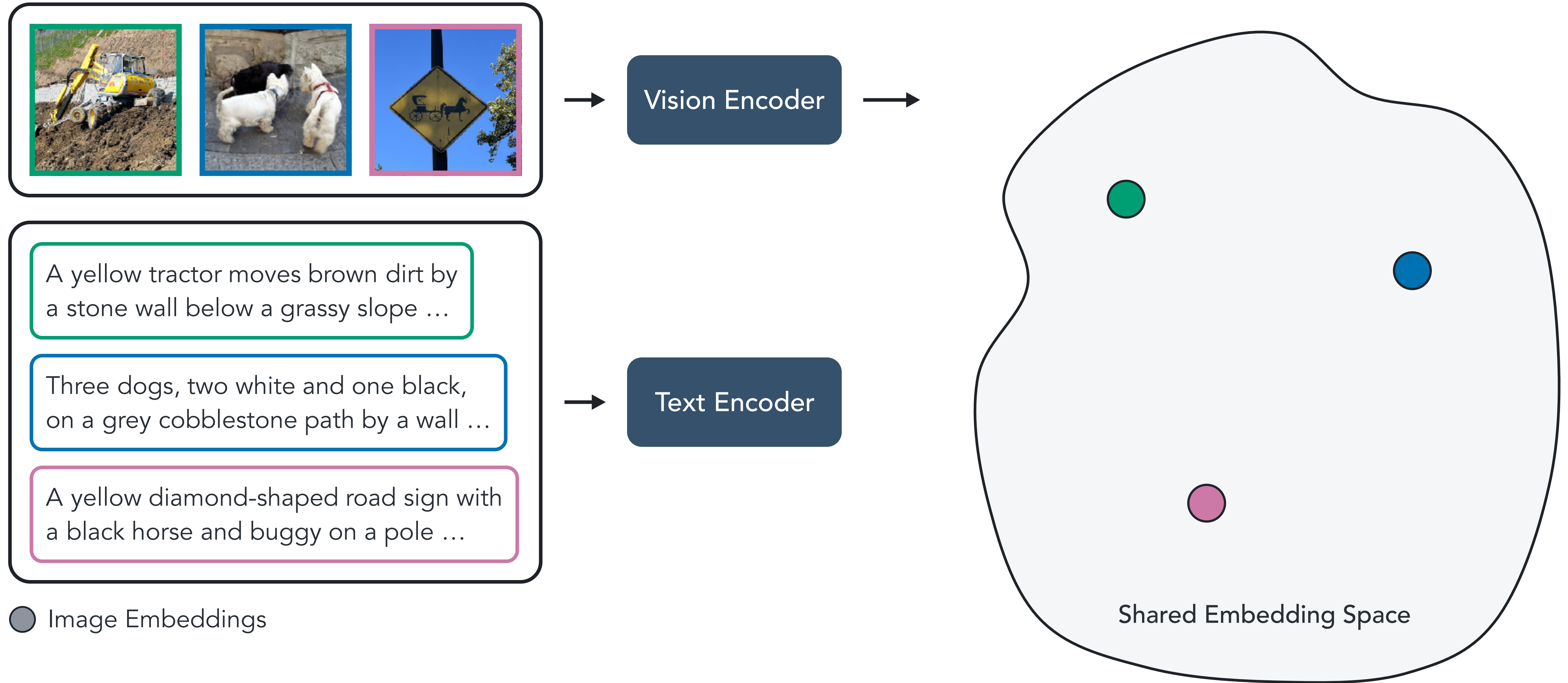
Vision Encoder

Text Encoder

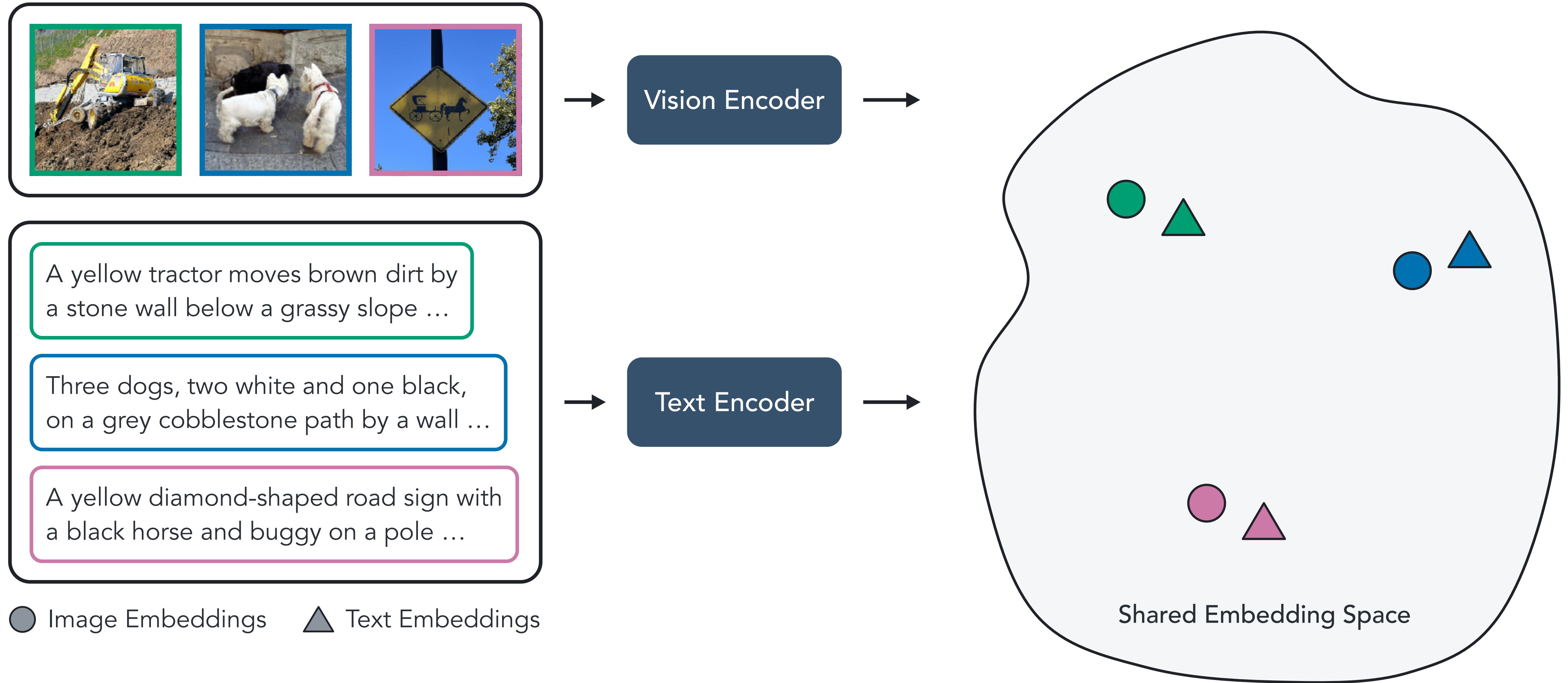
CLIP aligns images and text in a shared embedding space



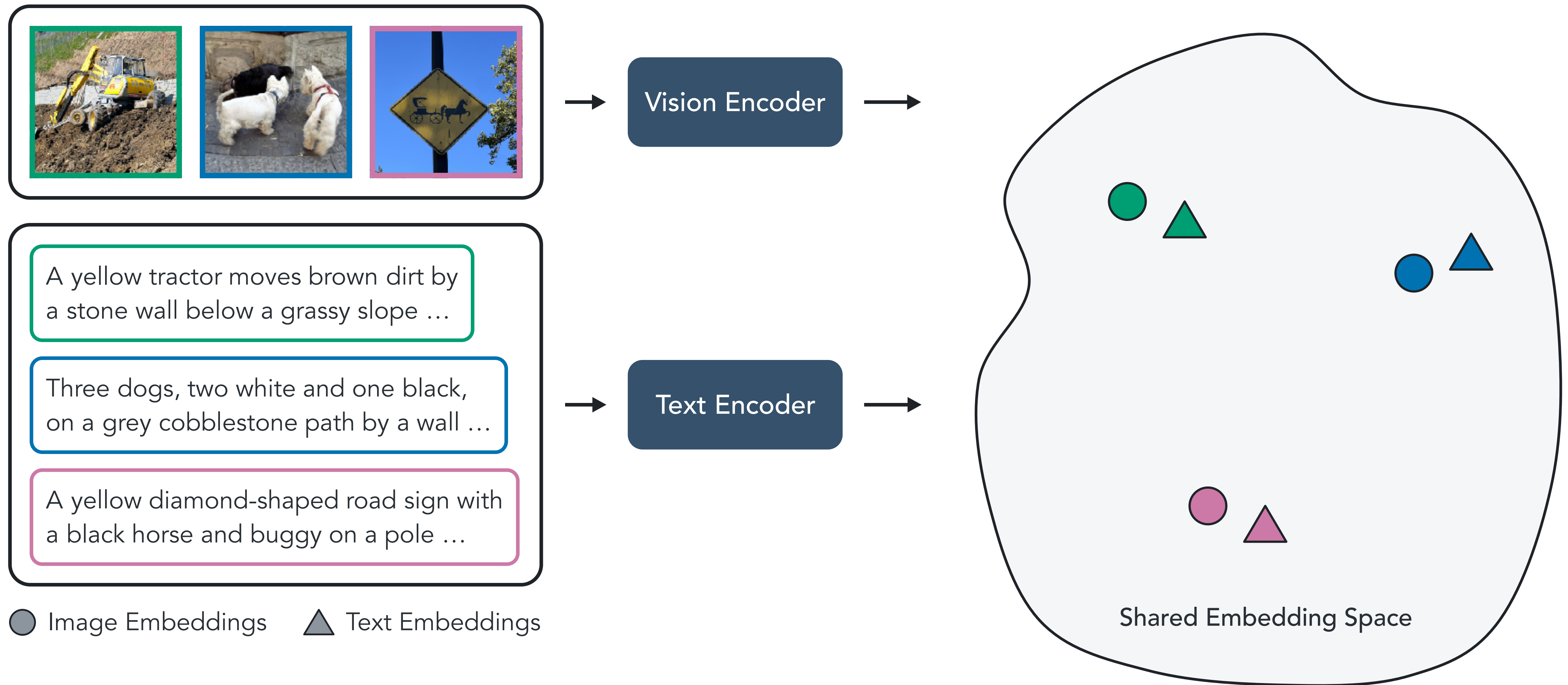
CLIP aligns images and text in a shared embedding space



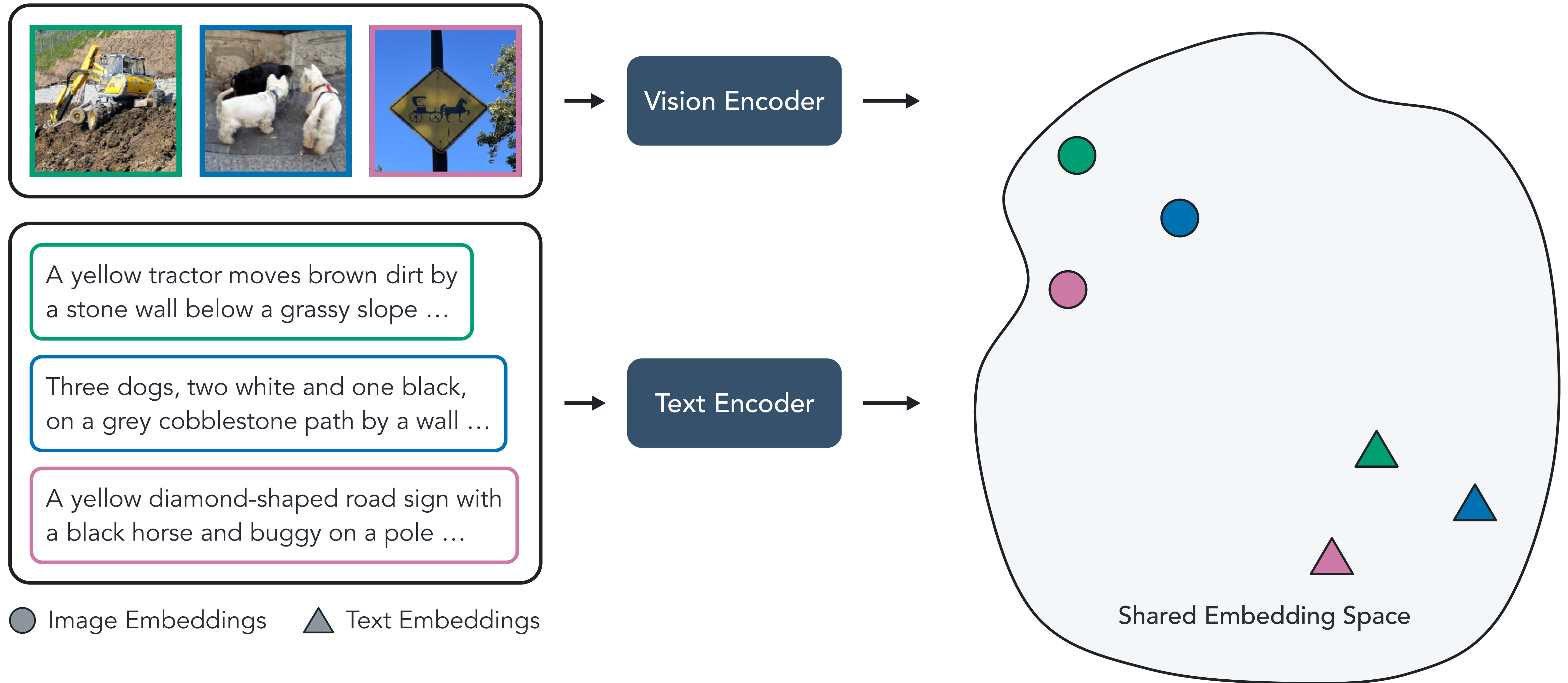
CLIP aligns images and text in a shared embedding space



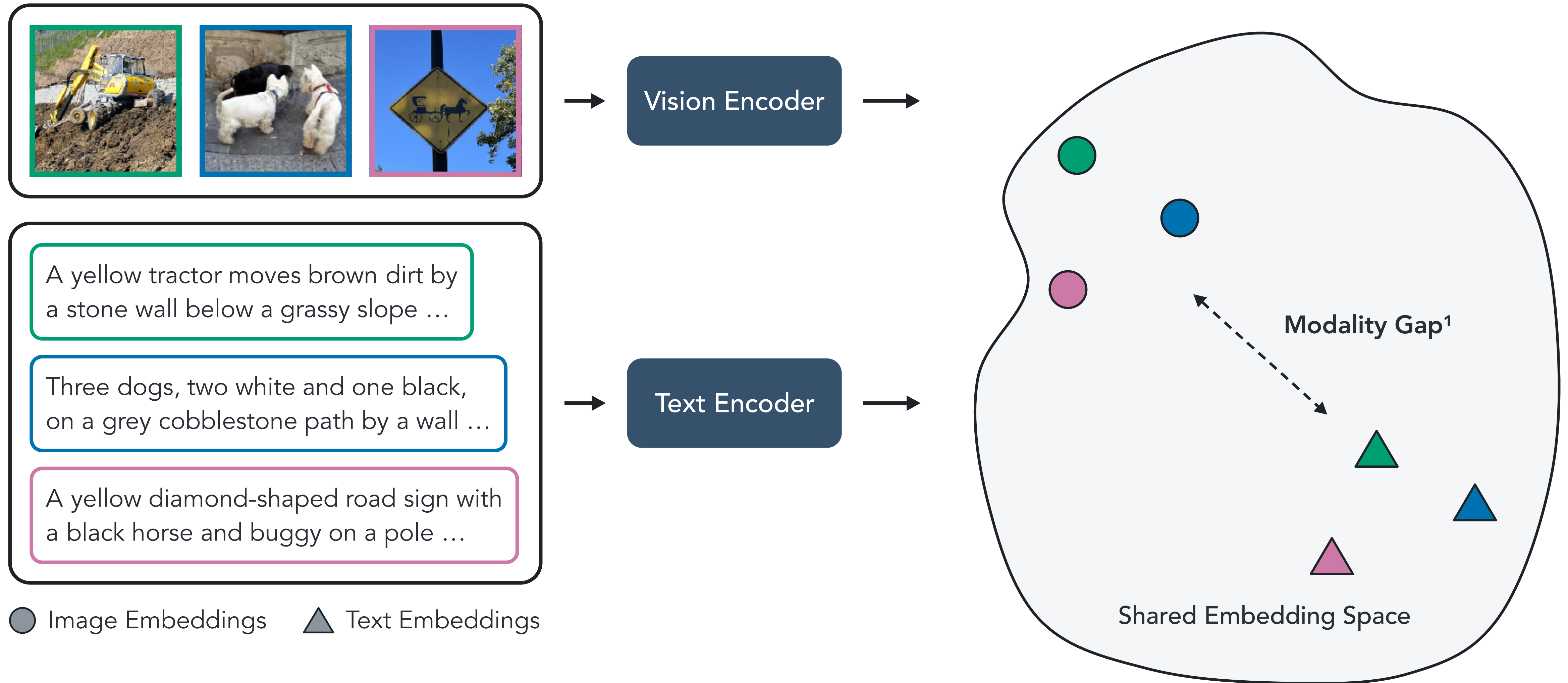
However, image and text embeddings lie apart



However, image and text embeddings lie apart

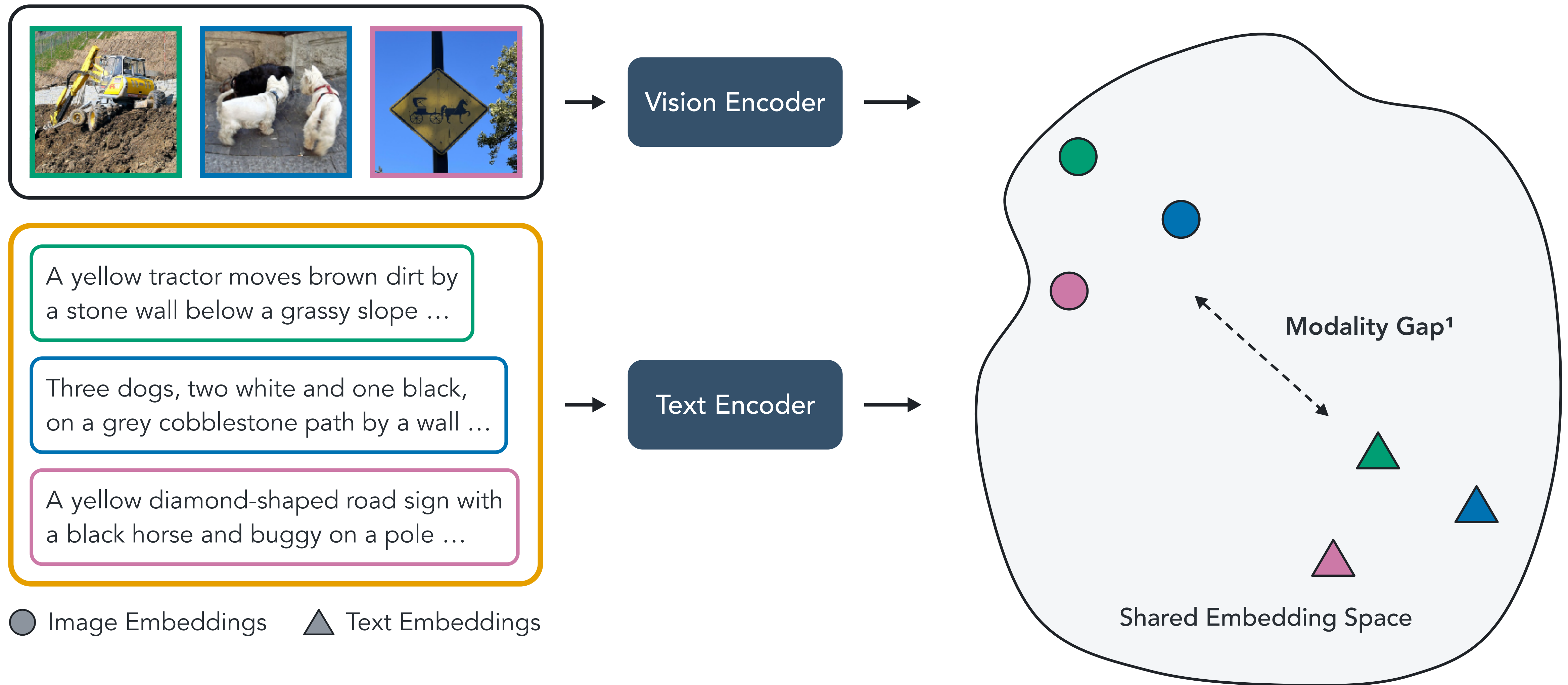


However, image and text embeddings lie apart



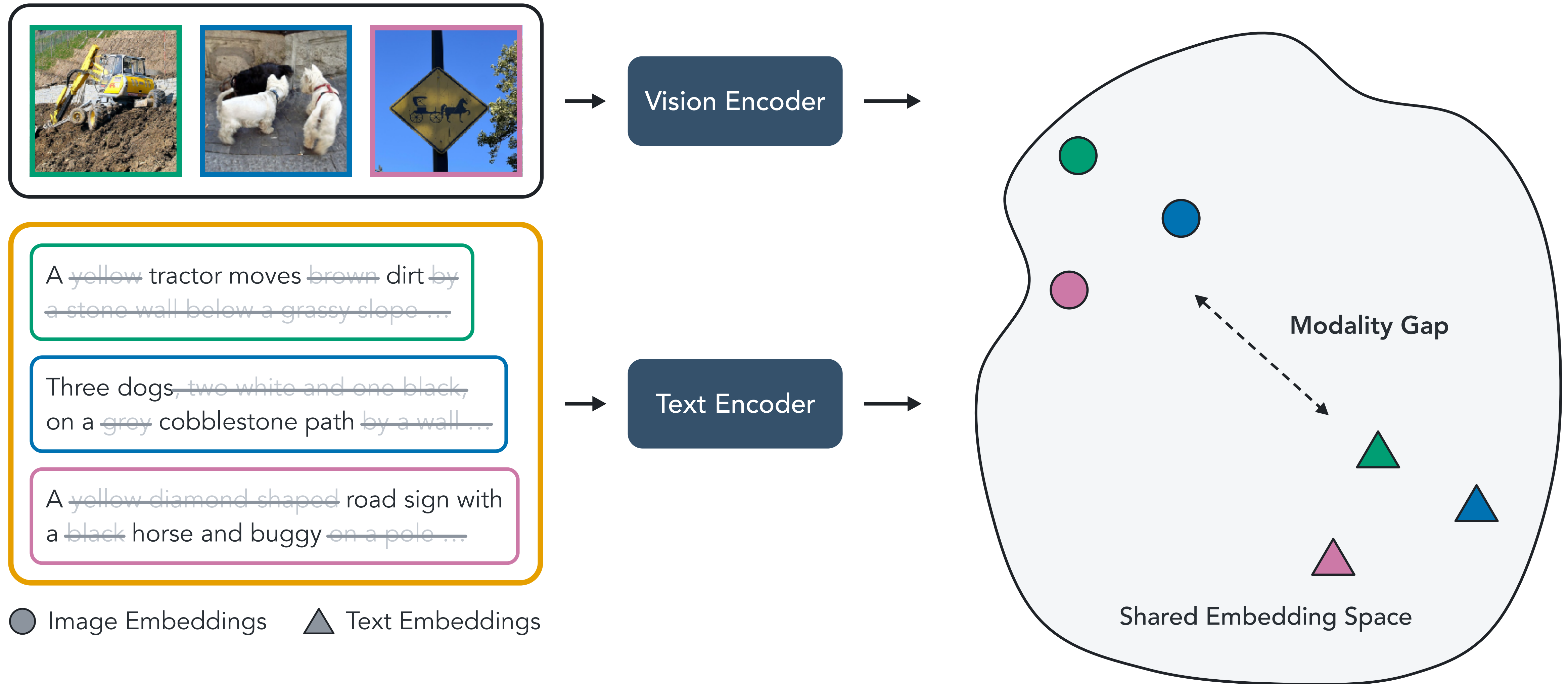
¹Liang et al., NeurIPS 2022

However, image and text embeddings lie apart



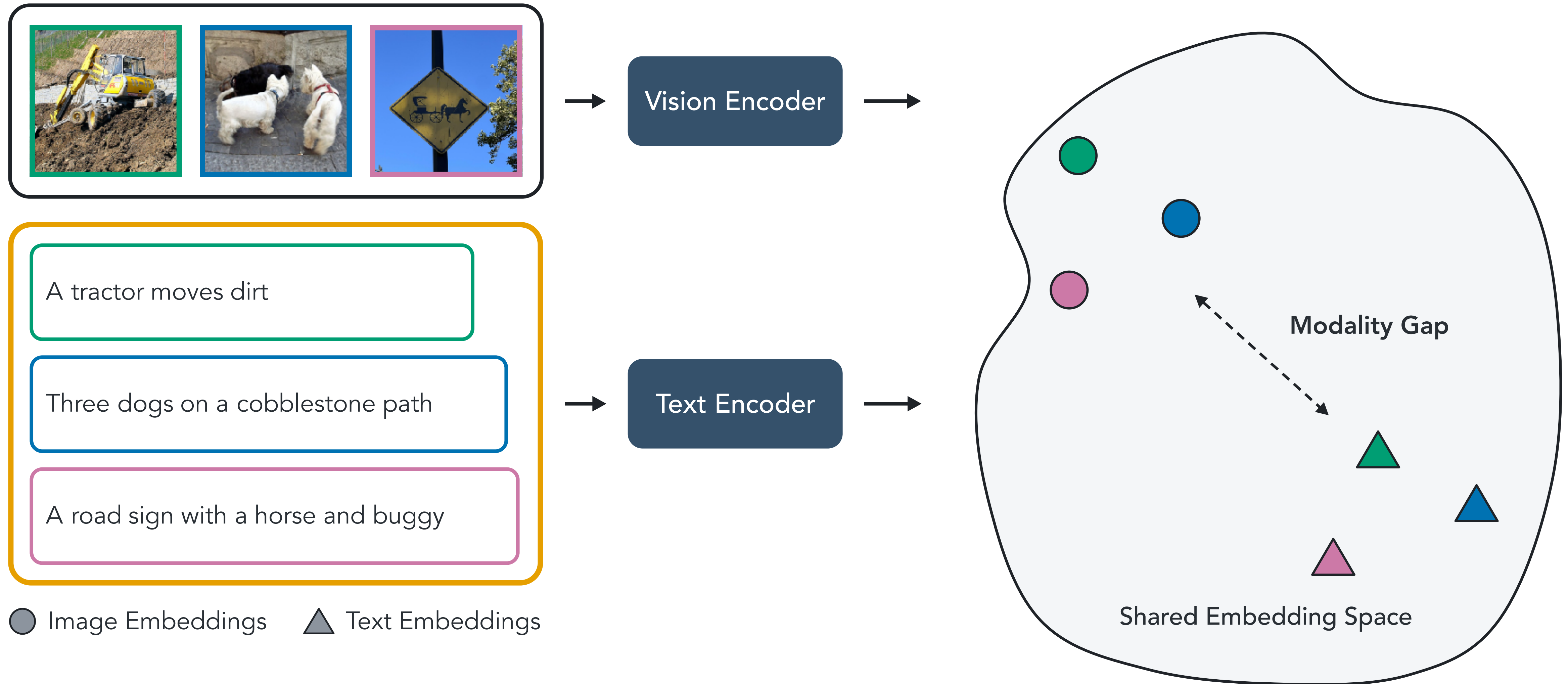
¹Liang et al., NeurIPS 2022

Hypothesis²: captions hold less information than images



²Schrodi et al., ICLR 2025

Hypothesis²: captions hold less information than images



²Schrodi et al., ICLR 2025

Hypothesis²: captions hold less information than images



A tractor moves dirt

²Schrodi et al., ICLR 2025

Hypothesis²: captions hold less information than images



A tractor moves dirt

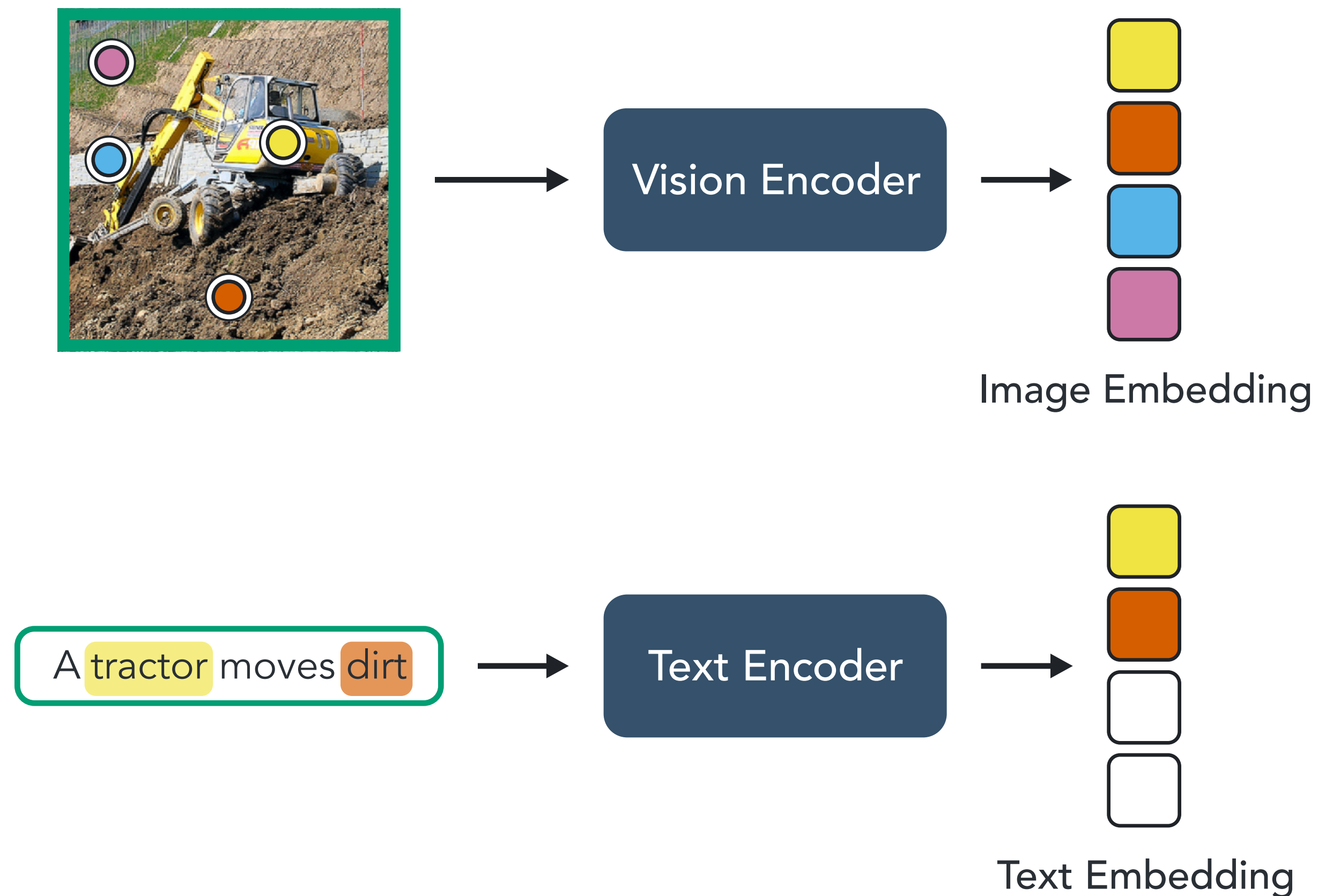
²Schrodi et al., ICLR 2025

The text embedding holds part of the image information

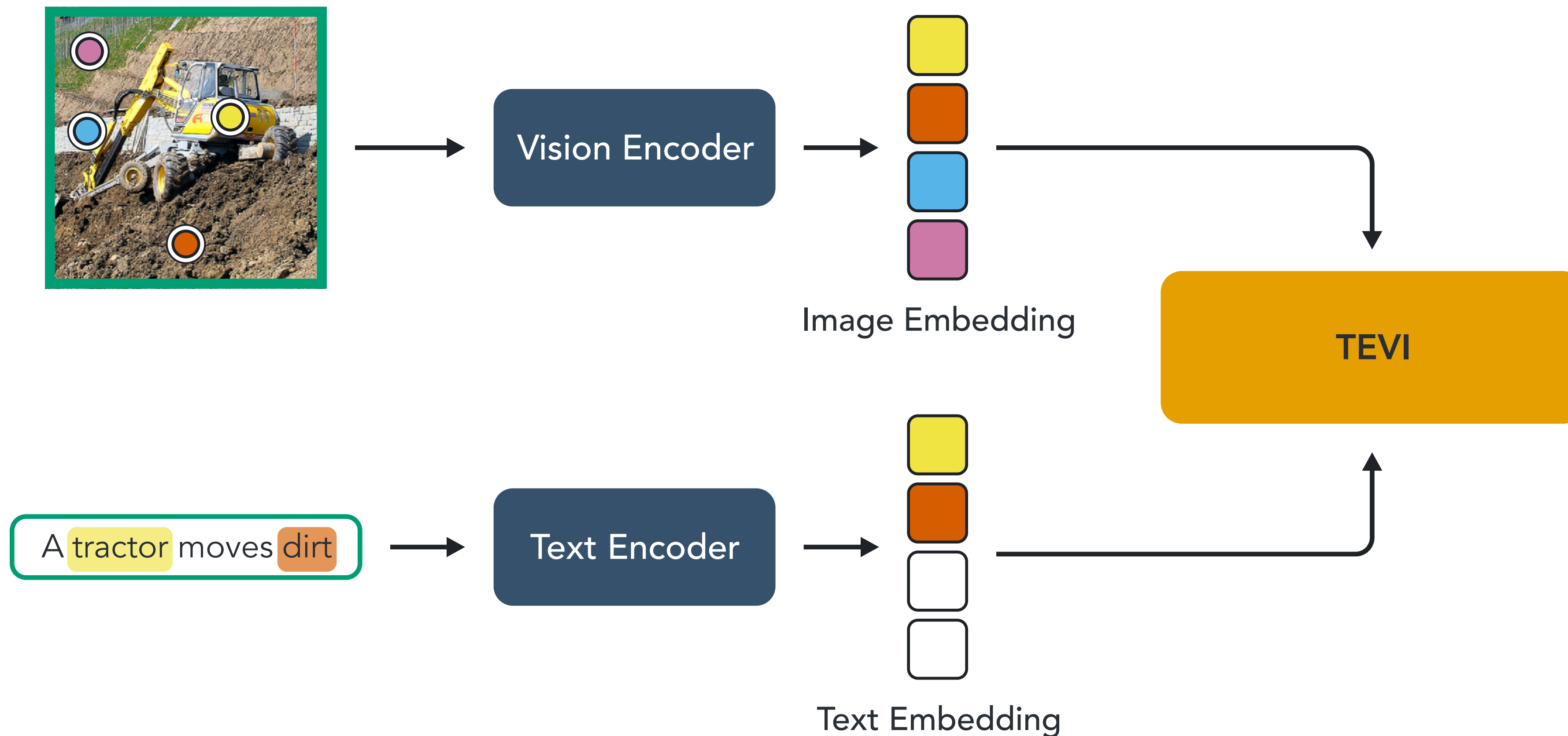


A tractor moves dirt

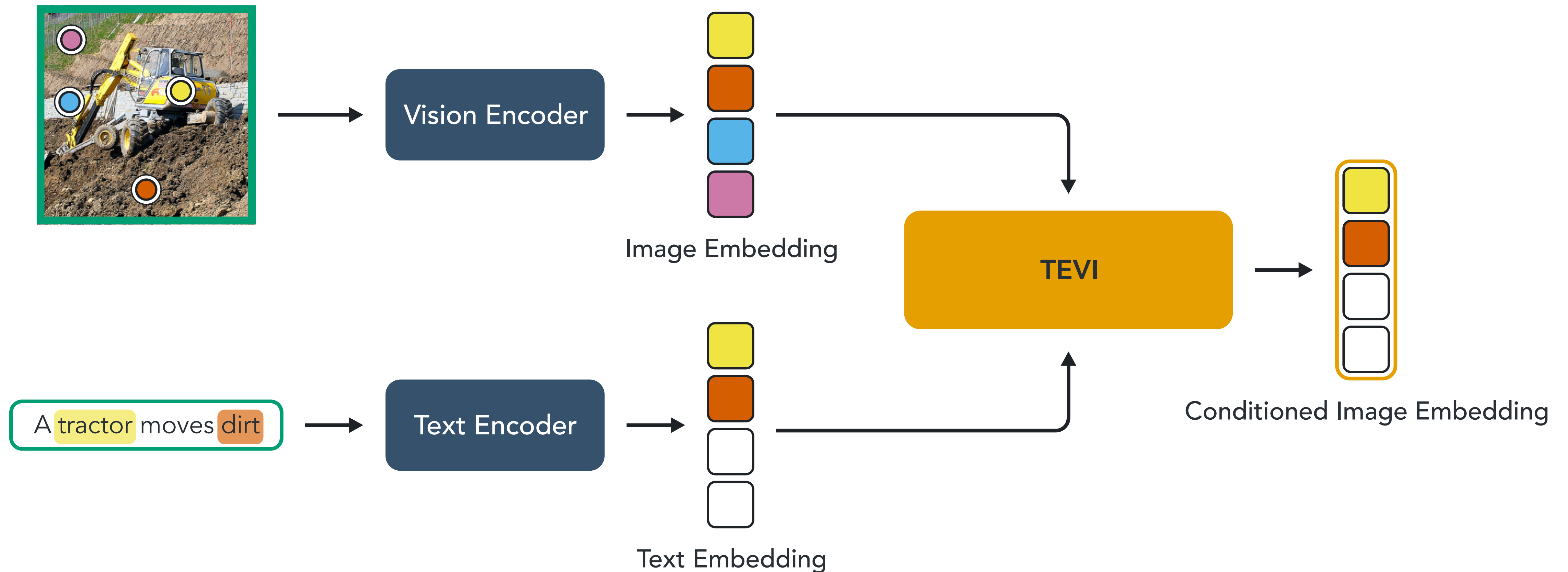
The text embedding holds part of the image information



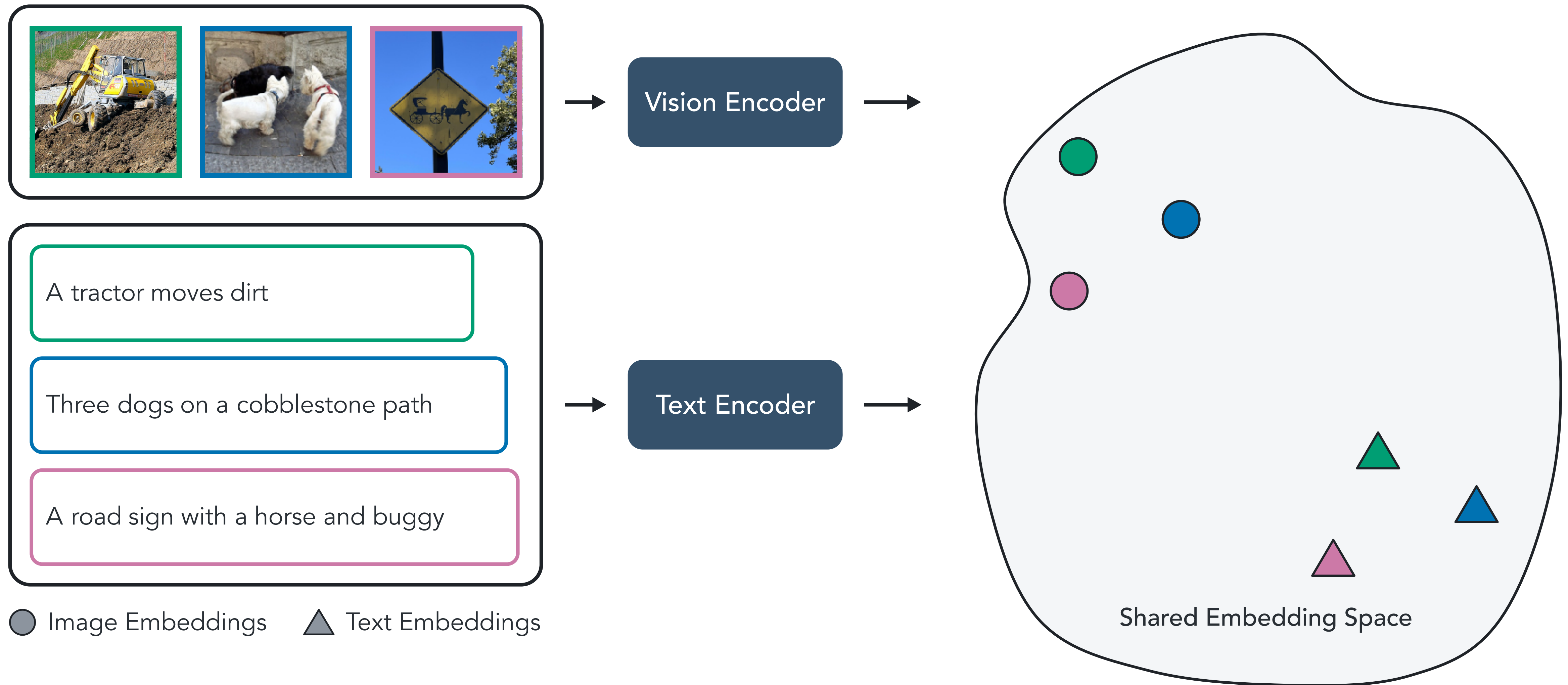
TEVI uses captions as a signal for what to retain



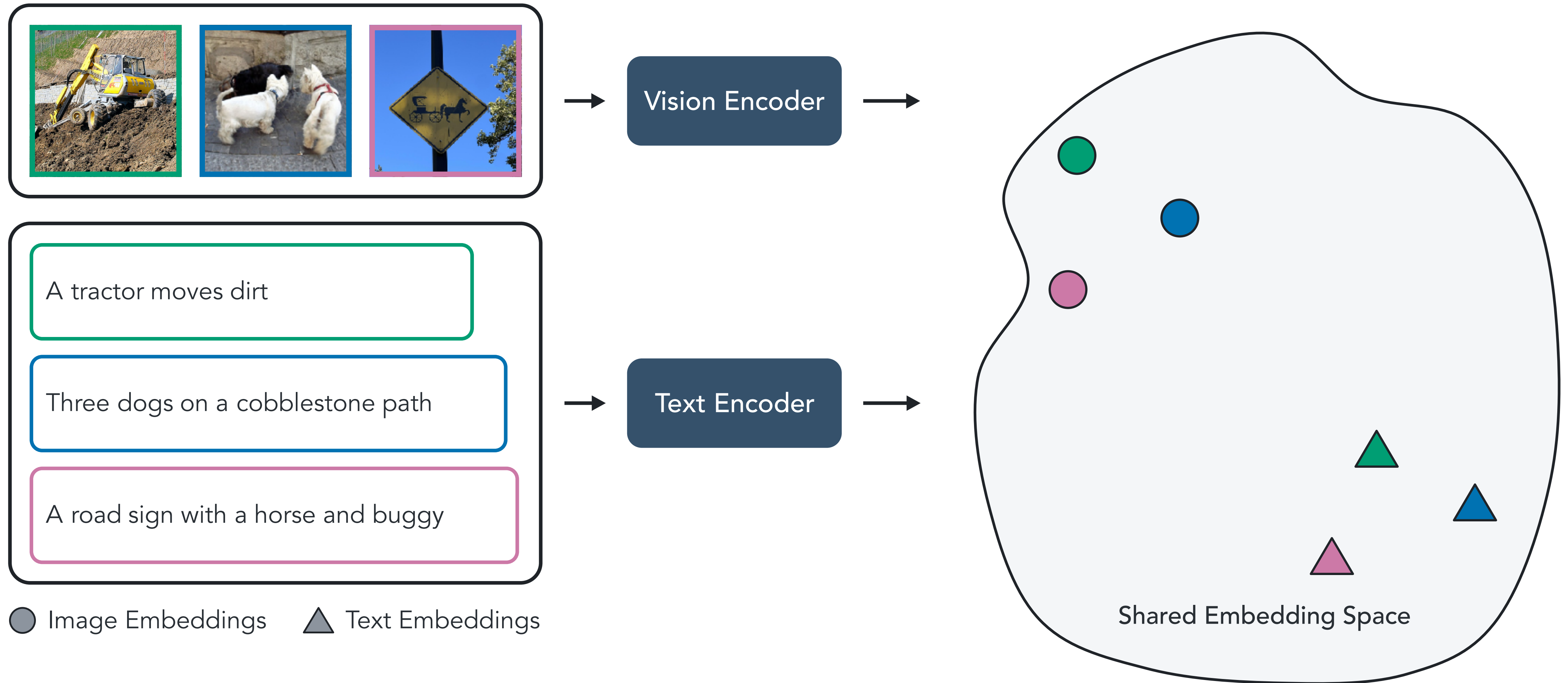
TEVI uses captions as a signal for what to retain



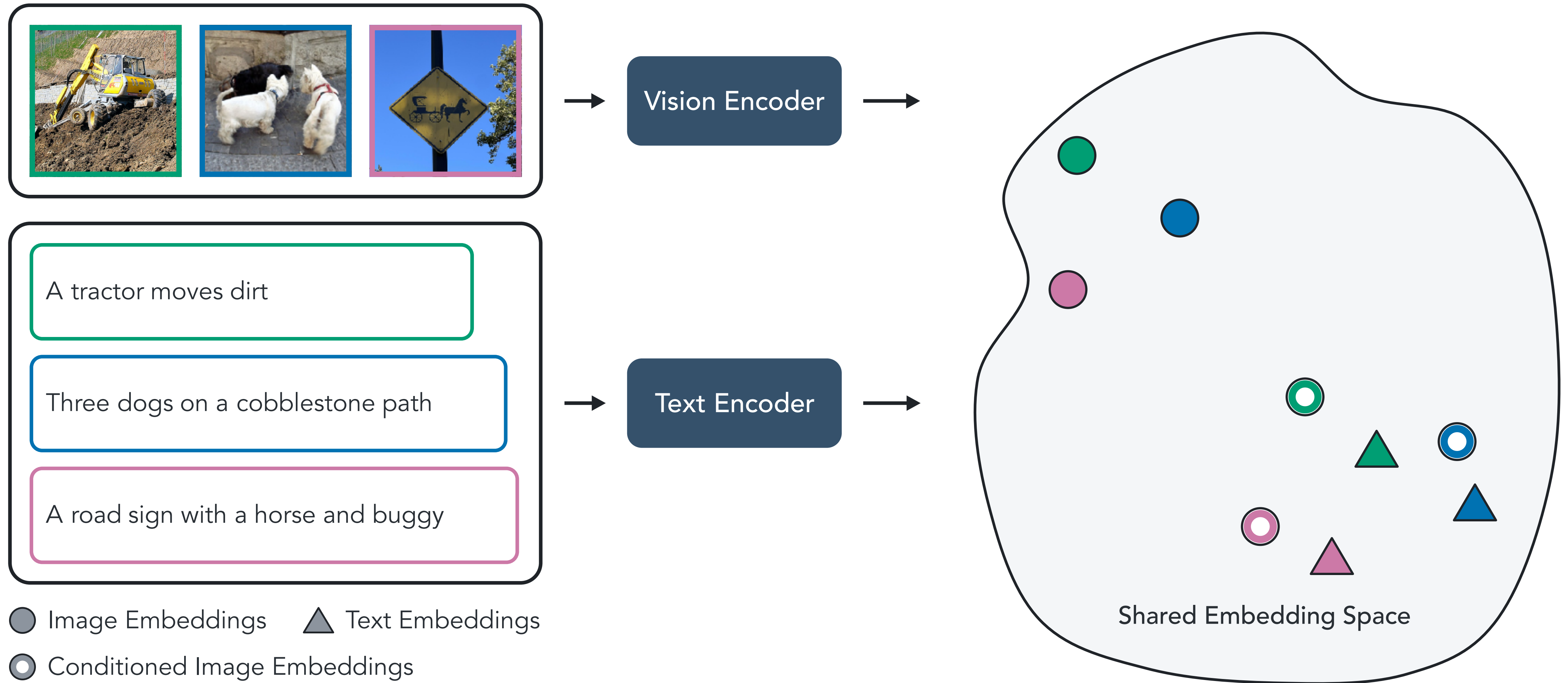
TEVI uses captions as a signal for what to retain



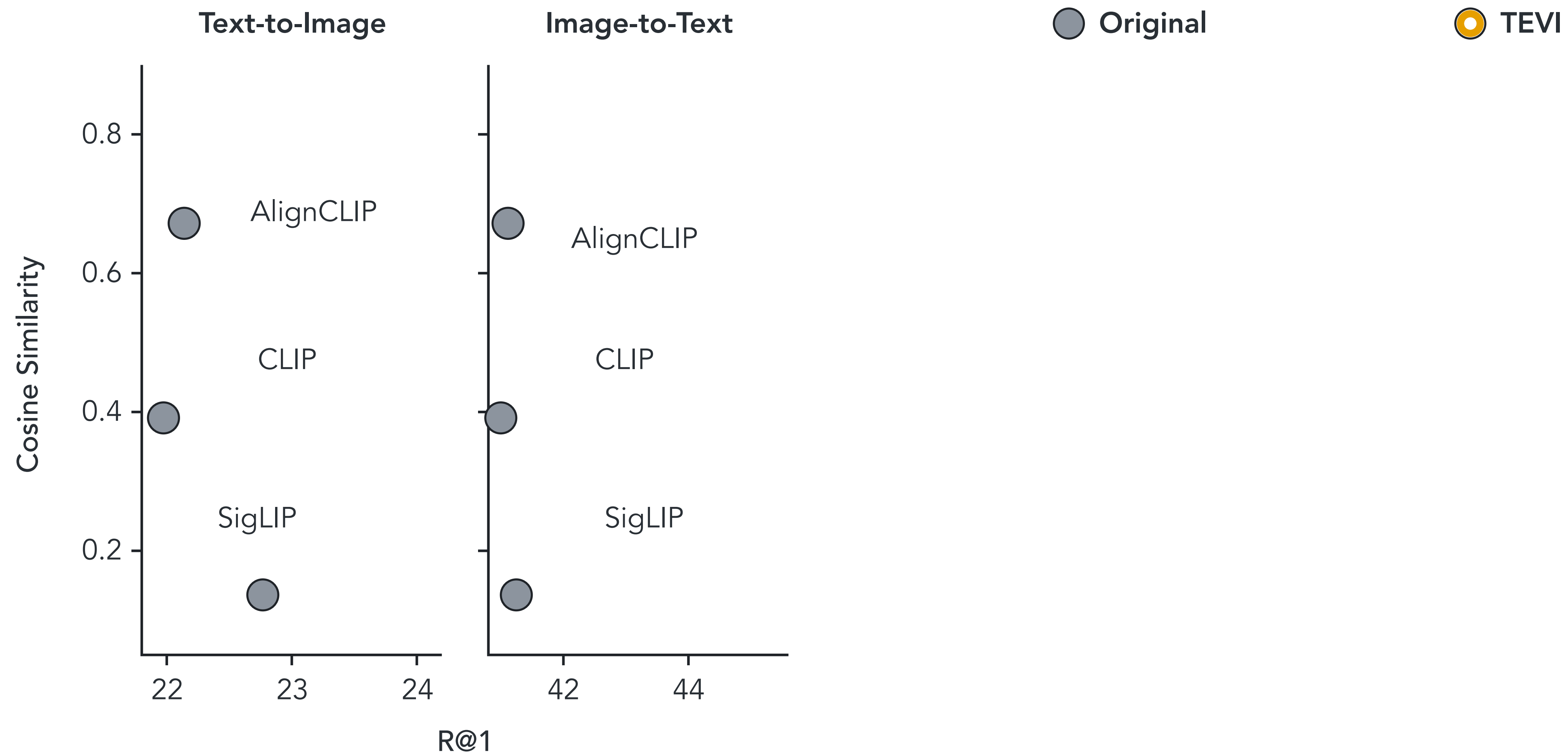
Conditioned embeddings lie closer to their captions



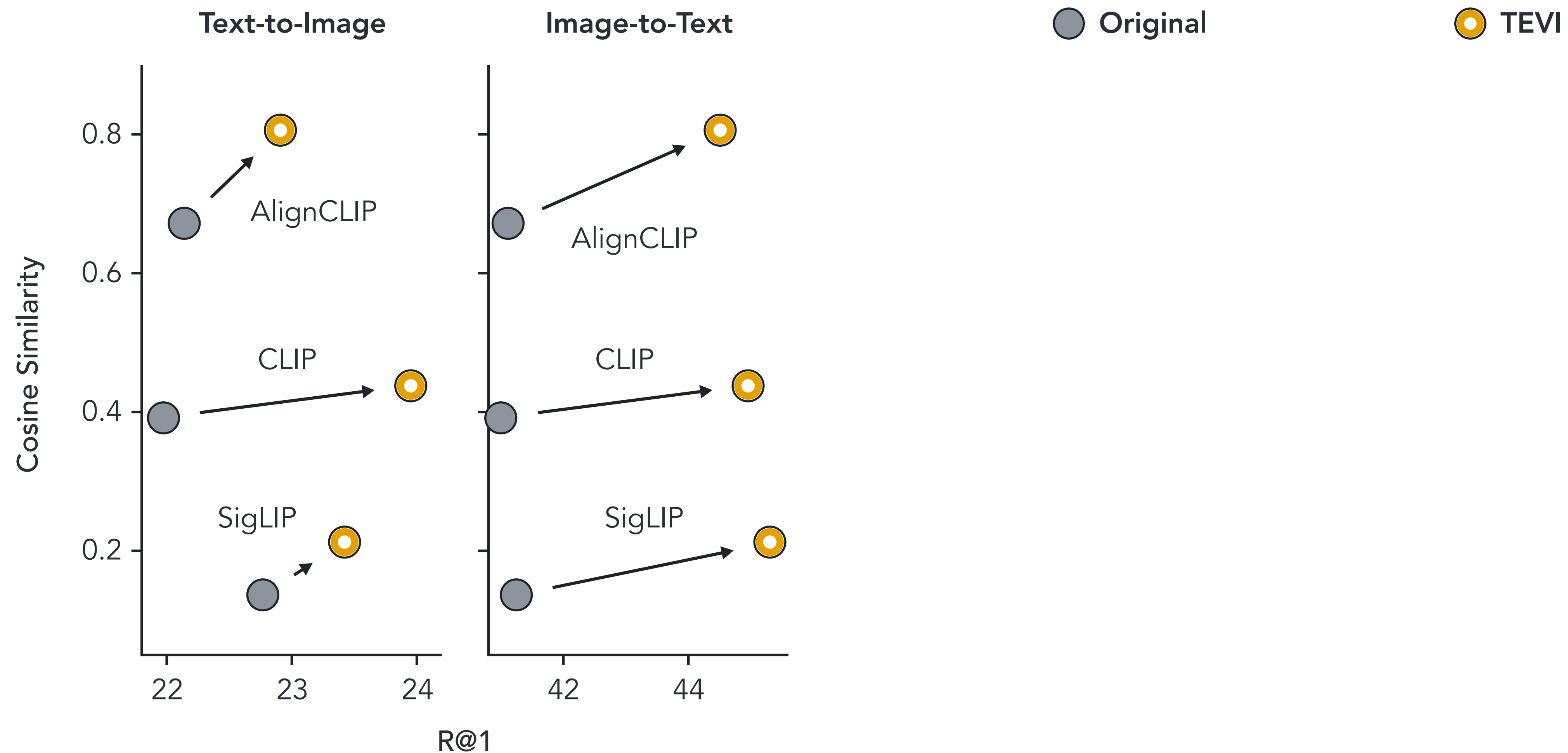
Conditioned embeddings lie closer to their captions



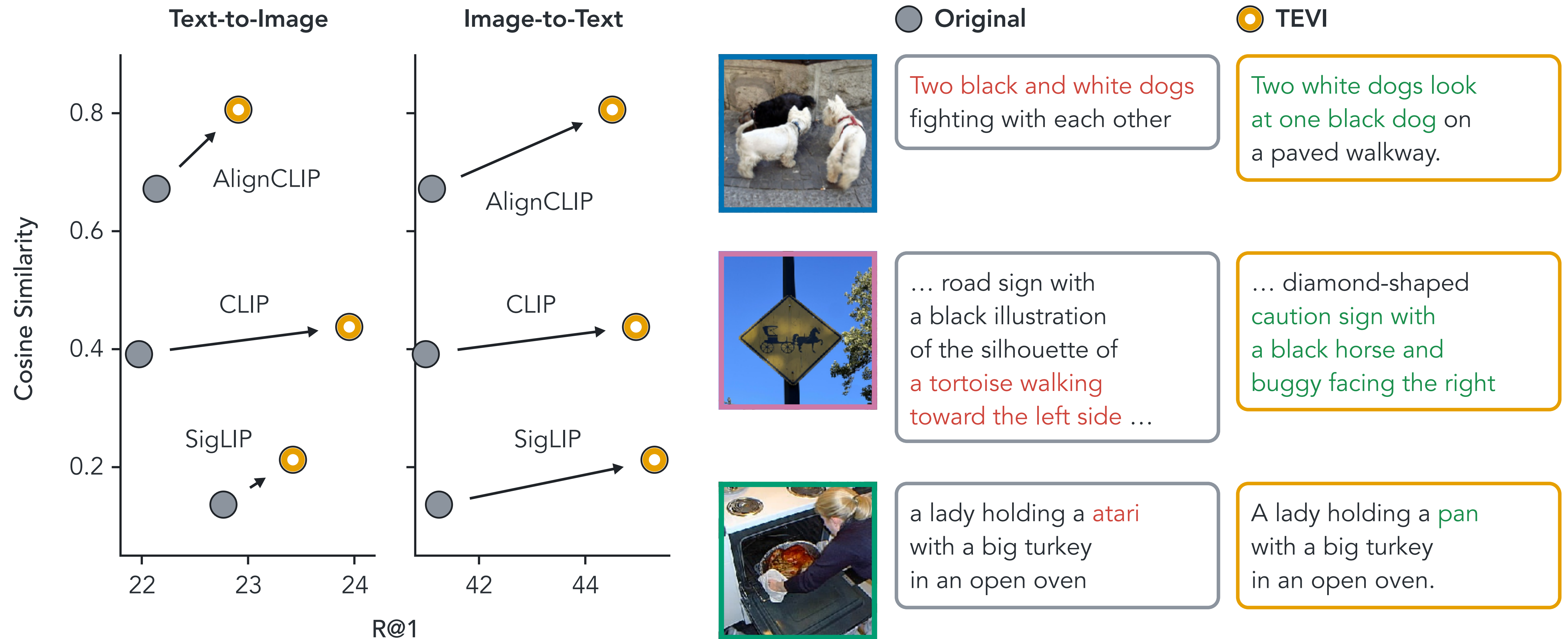
TEVI improves cross-modal alignment and retrieval



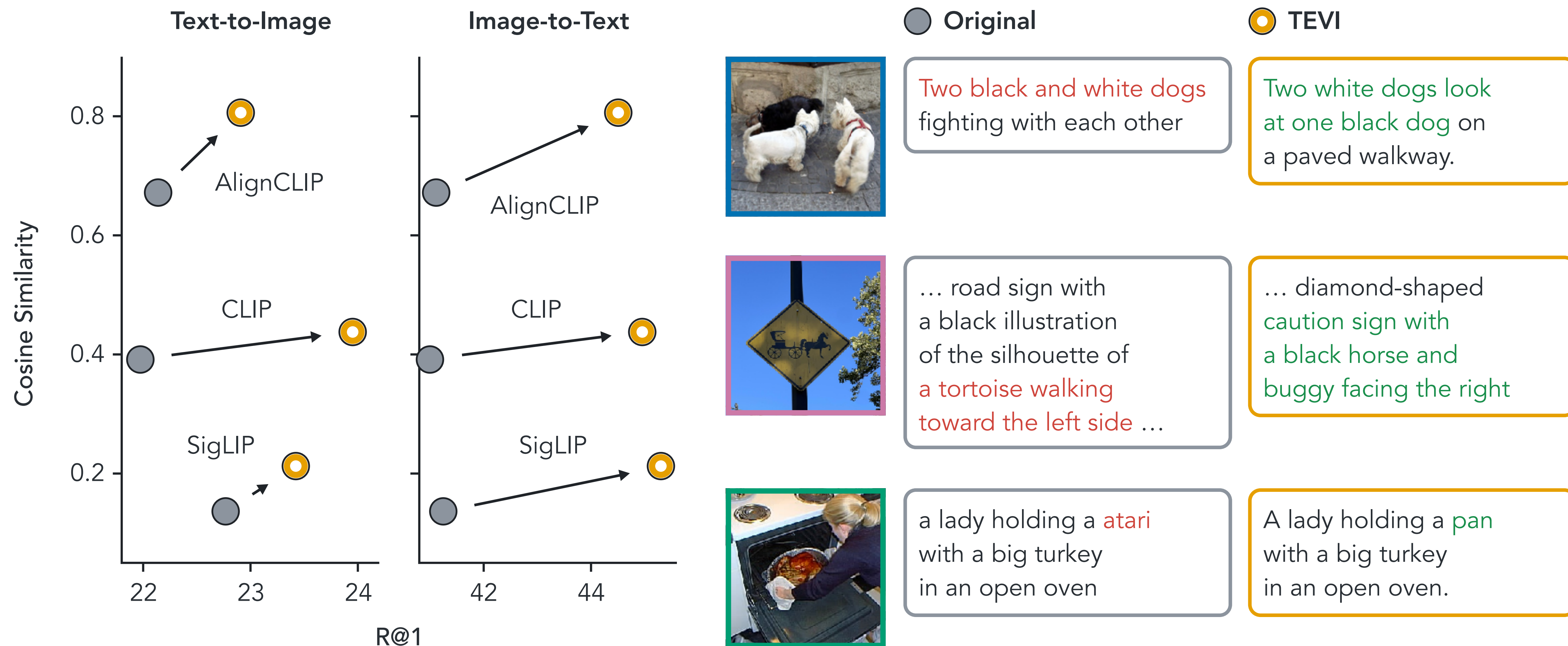
TEVI improves cross-modal alignment and retrieval



TEVI improves cross-modal alignment and retrieval

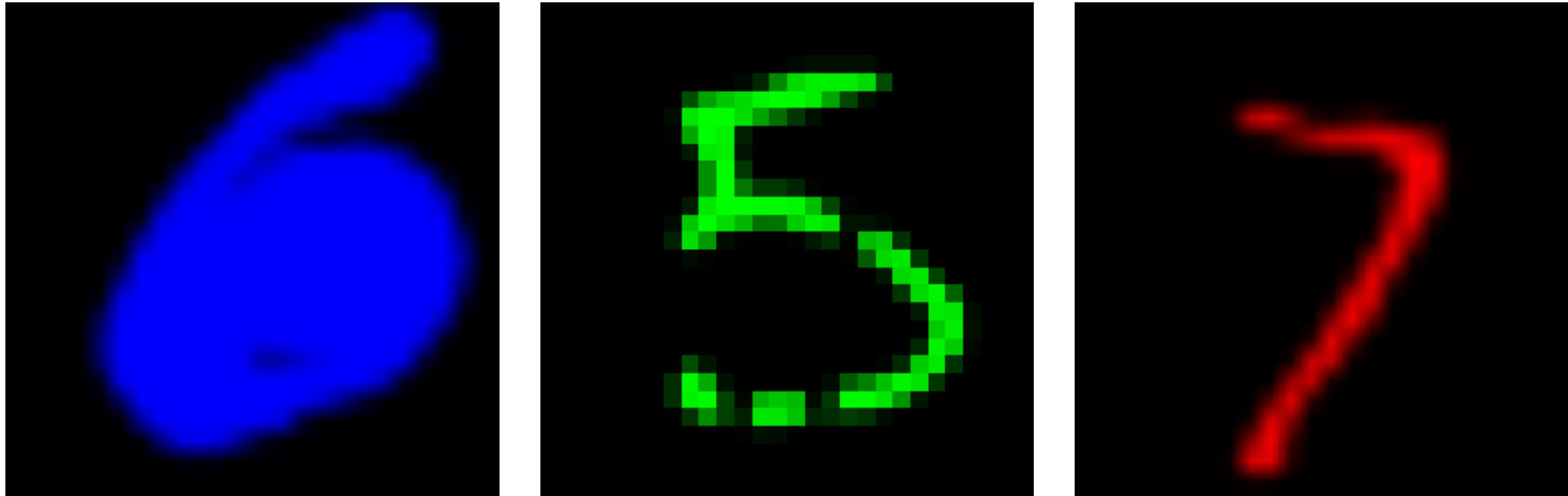


We first use a controlled setup with known attributes



We first use a controlled setup with known attributes

Controlled Setup



Synthetic Digits with Known Attributes



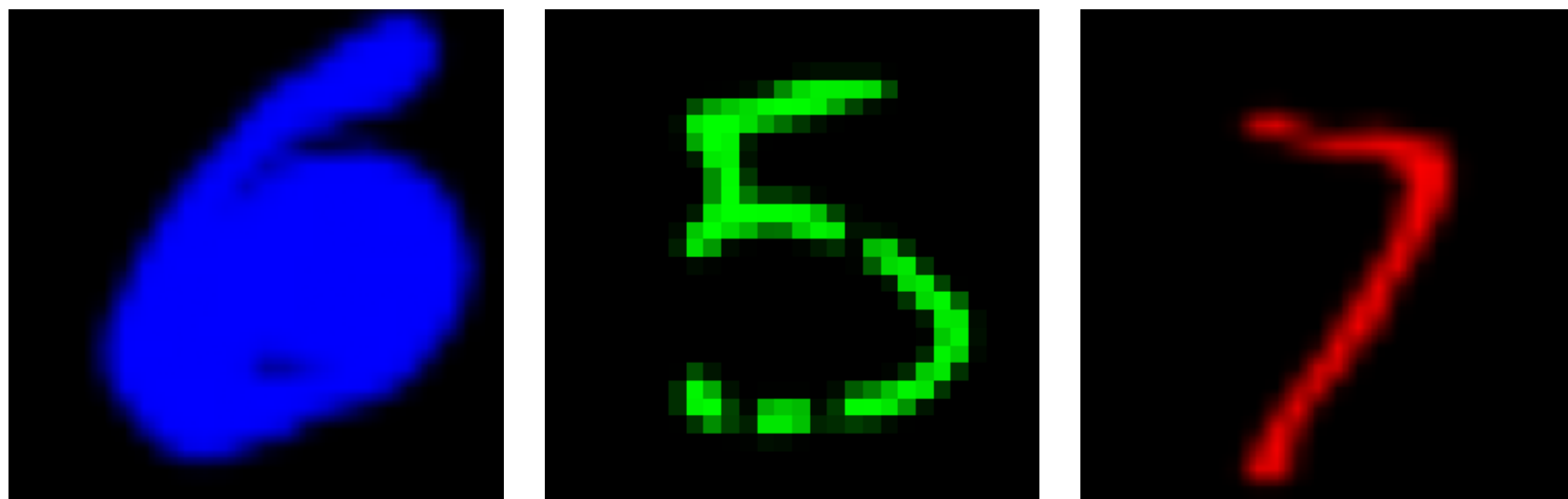
Natural Images



CLIP Models Trained on CC12M

Controlled setup: MAD dataset

Controlled Setup



Synthetic Digits with Known Attributes

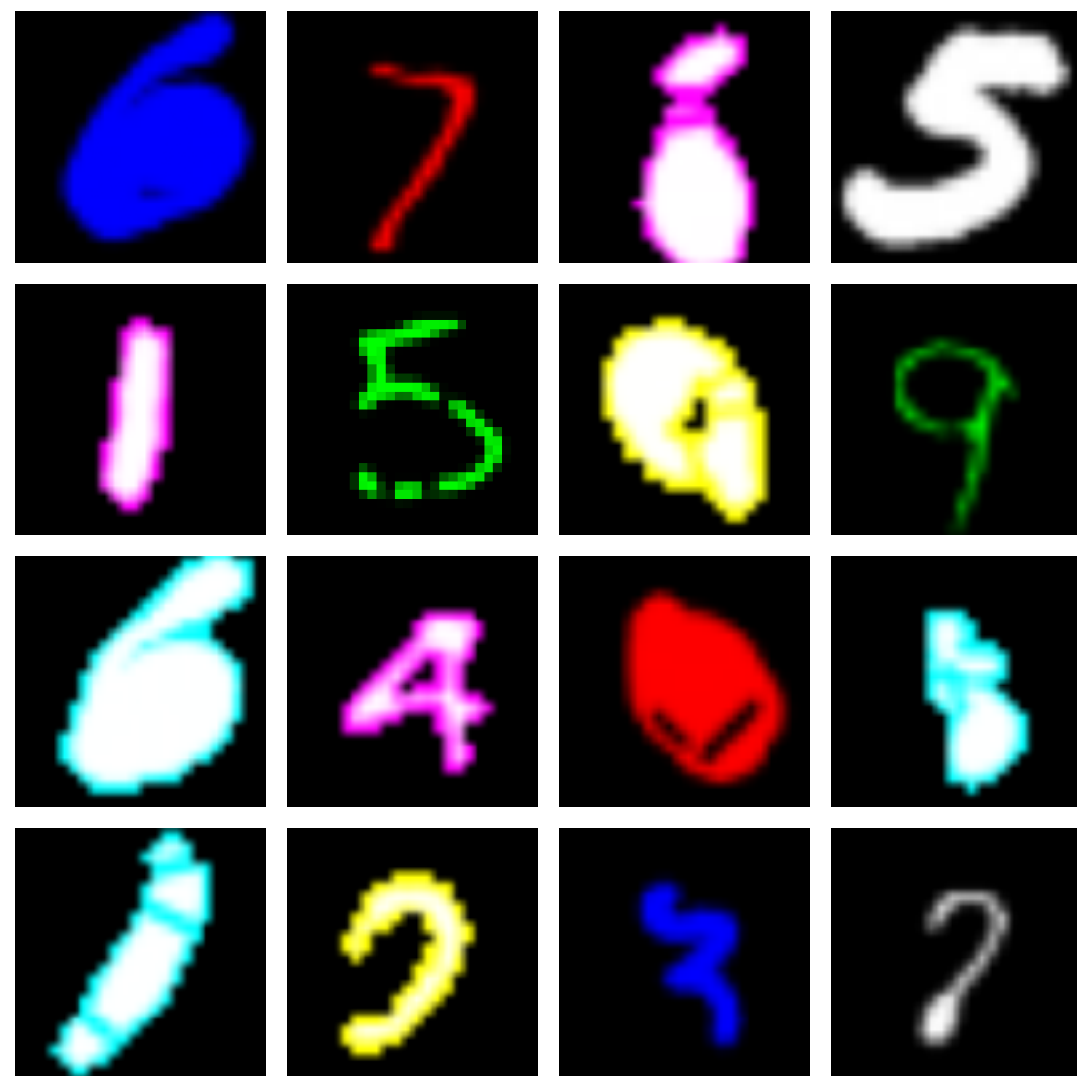


Natural Images



CLIP Models Trained on CC12M

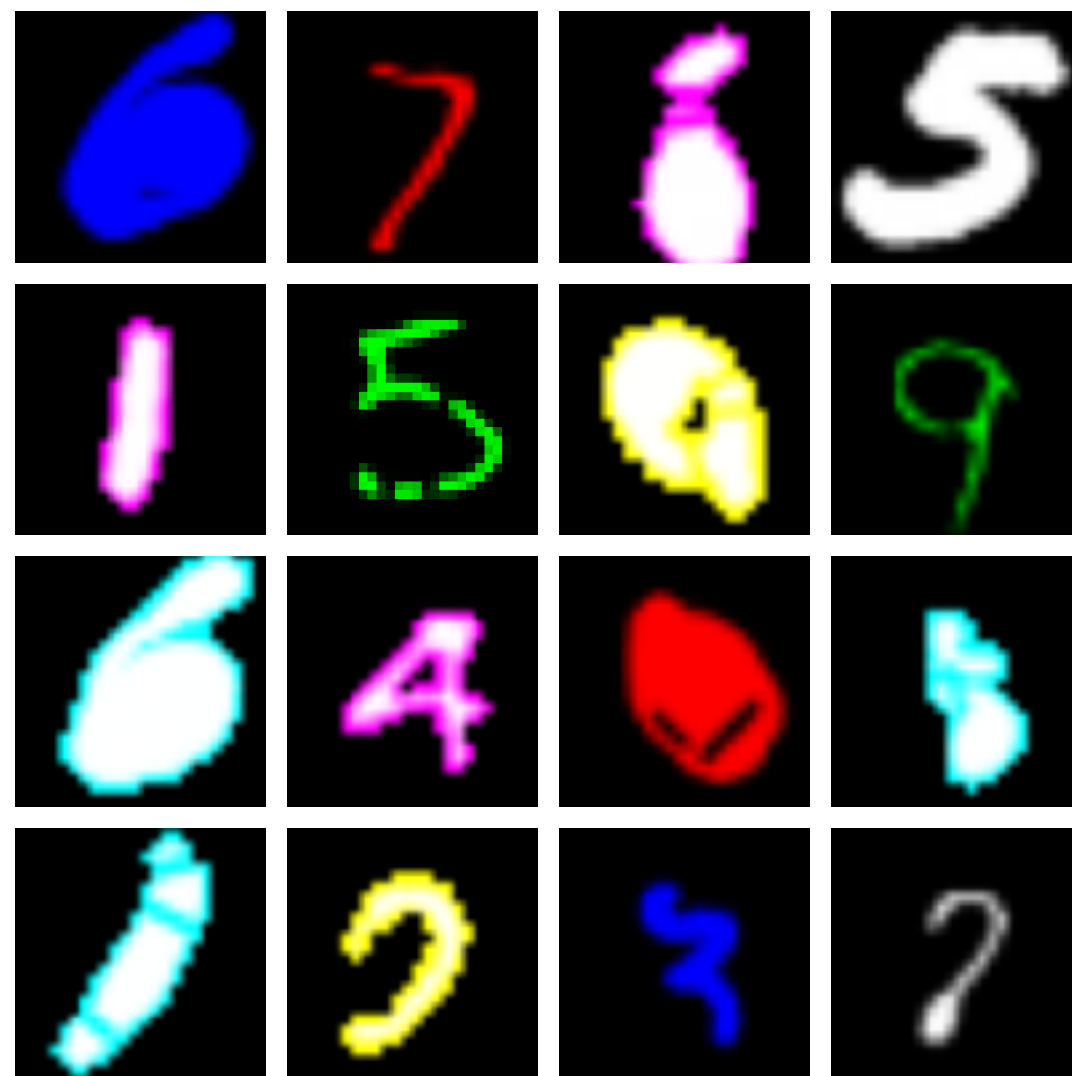
Controlled setup: MAD dataset



- **Digit** 0 · 1 · 2 · 3 · 4 · 5 · 6 · 7 · 8 · 9
- **Color** gray · red · green · blue · cyan · magenta · yellow
- **Thickthinning** thickening · thinning · no thickthinning
- **Swelling** swelling · no swelling
- **Fracture** fracture · no fracture
- **Scaling** large · small

MNIST with Morphological Transforms

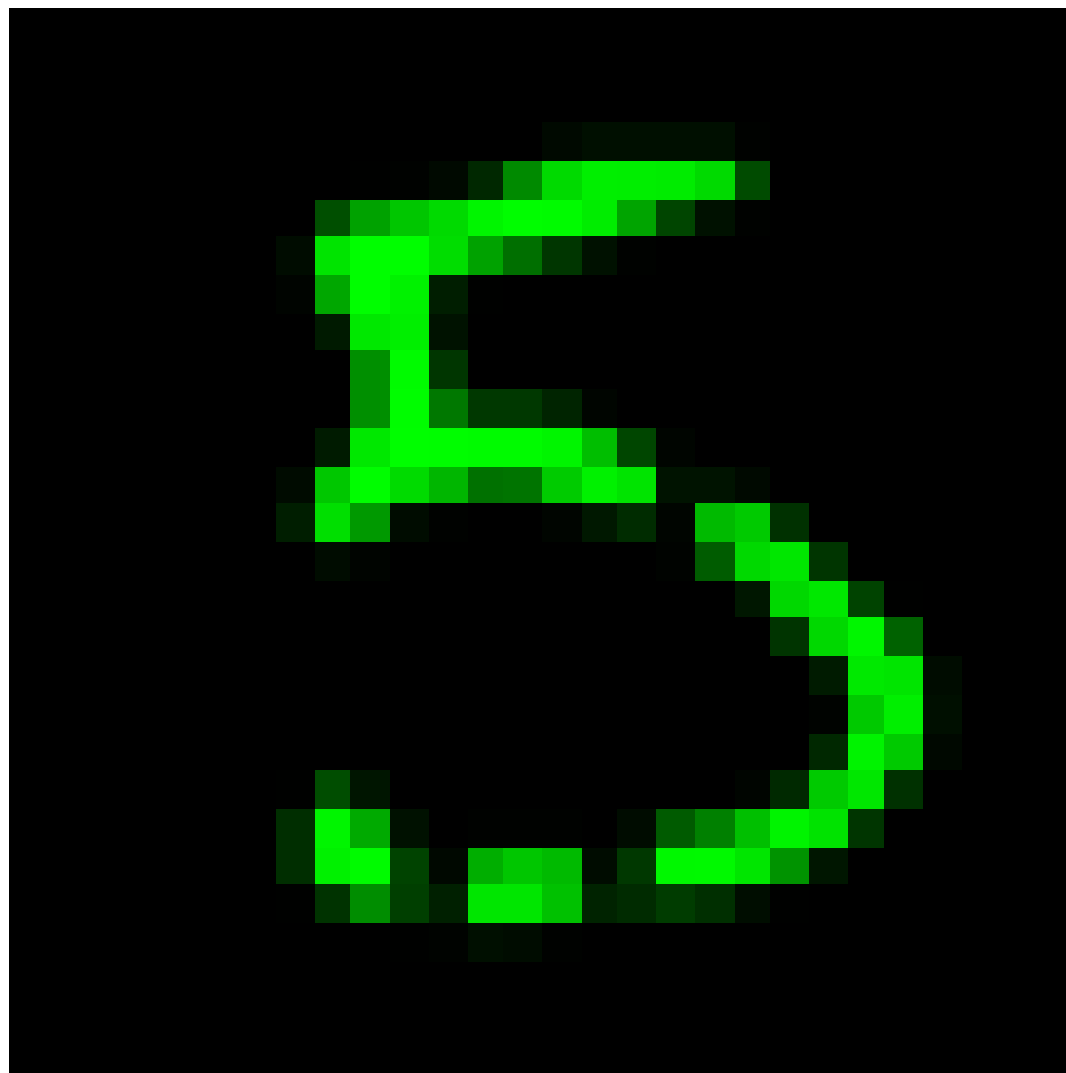
Every MAD image is characterized by six attributes



- **Digit** 0 · 1 · 2 · 3 · 4 · 5 · 6 · 7 · 8 · 9
- **Color** gray · red · green · blue · cyan · magenta · yellow
- **Thickthinning** thickening · thinning · no thickthinning
- **Swelling** swelling · no swelling
- **Fracture** fracture · no fracture
- **Scaling** large · small

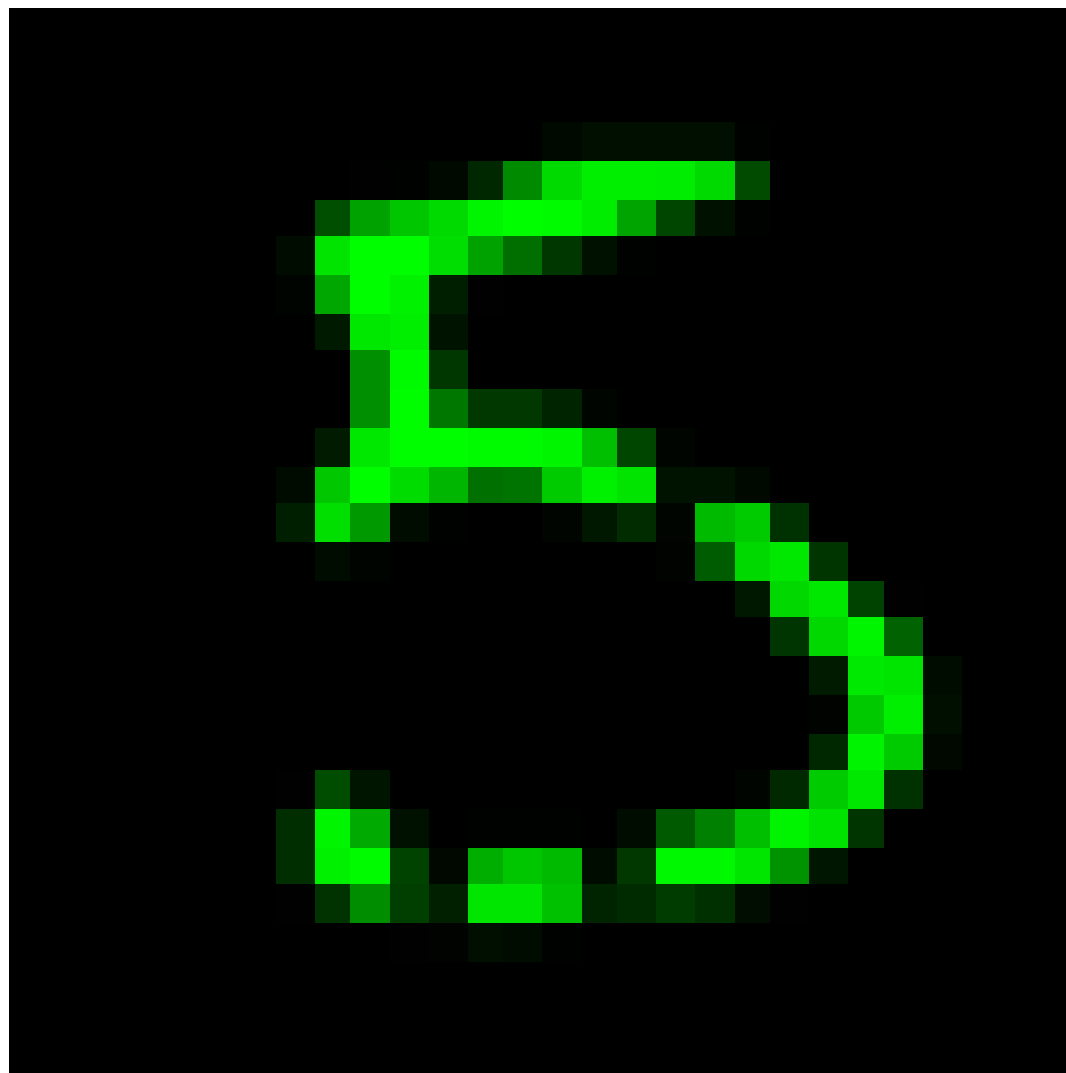
MNIST with Morphological Transforms

Every MAD image is characterized by six attributes



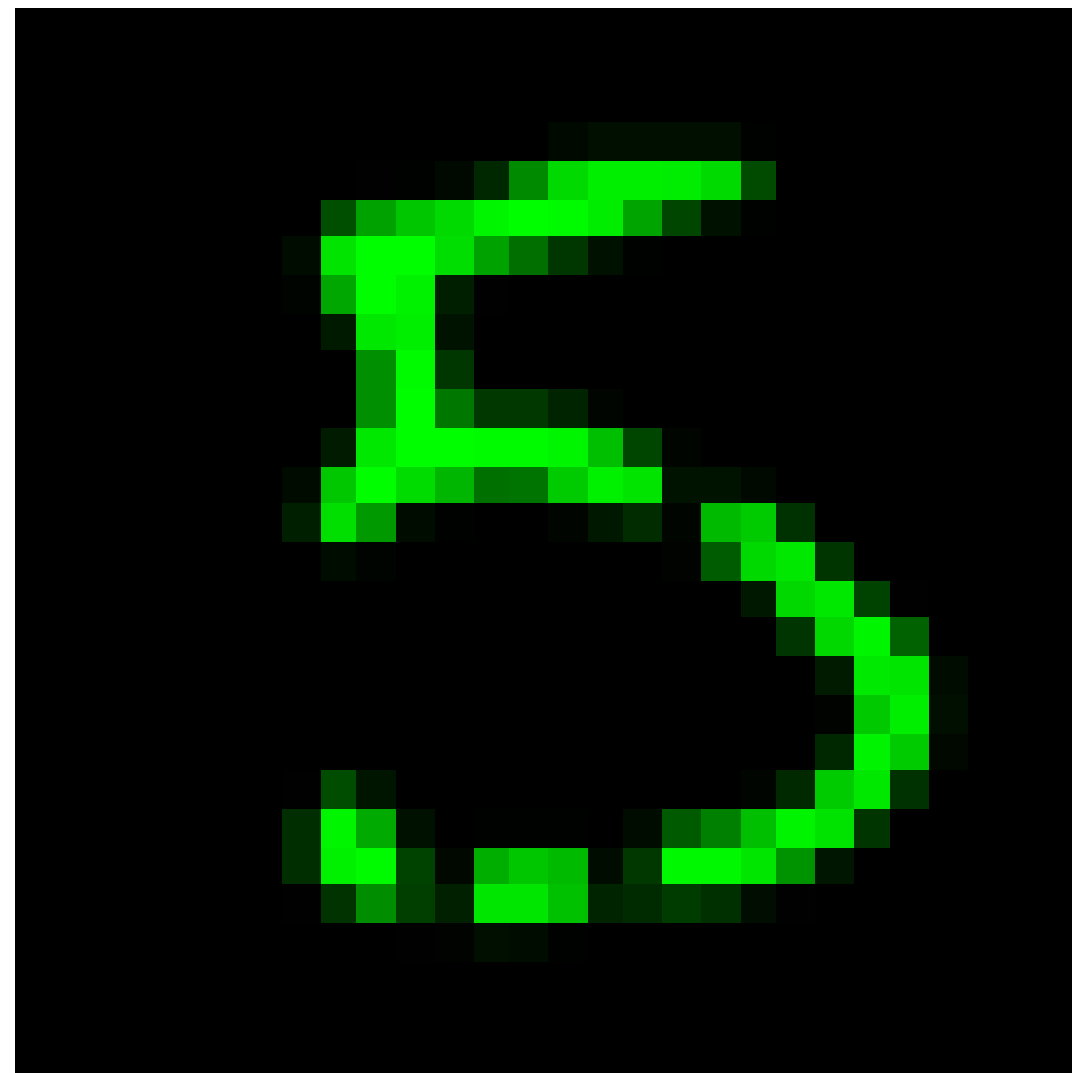
- **Digit** 0 · 1 · 2 · 3 · 4 · 5 · 6 · 7 · 8 · 9
- **Color** gray · red · green · blue · cyan · magenta · yellow
- **Thickthinning** thickening · thinning · no thickthinning
- **Swelling** swelling · no swelling
- **Fracture** fracture · no fracture
- **Scaling** large · small

Every MAD image is characterized by six attributes



- Digit 0 · 1 · 2 · 3 · 4 · 5 · 6 · 7 · 8 · 9
- Color gray · red · green · blue · cyan · magenta · yellow
- Thickthinning thickening · thinning · no thickthinning
- Swelling swelling · no swelling
- Fracture fracture · no fracture
- Scaling large · small

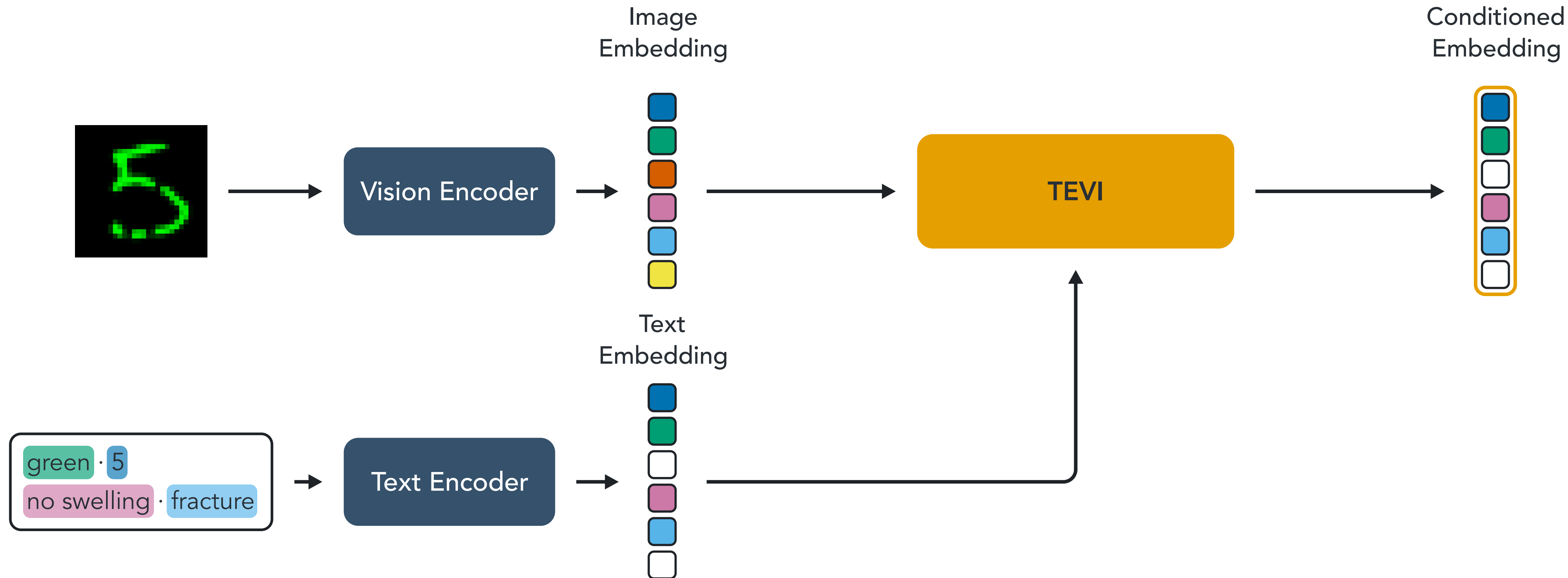
Captions contain the digit and three random attributes



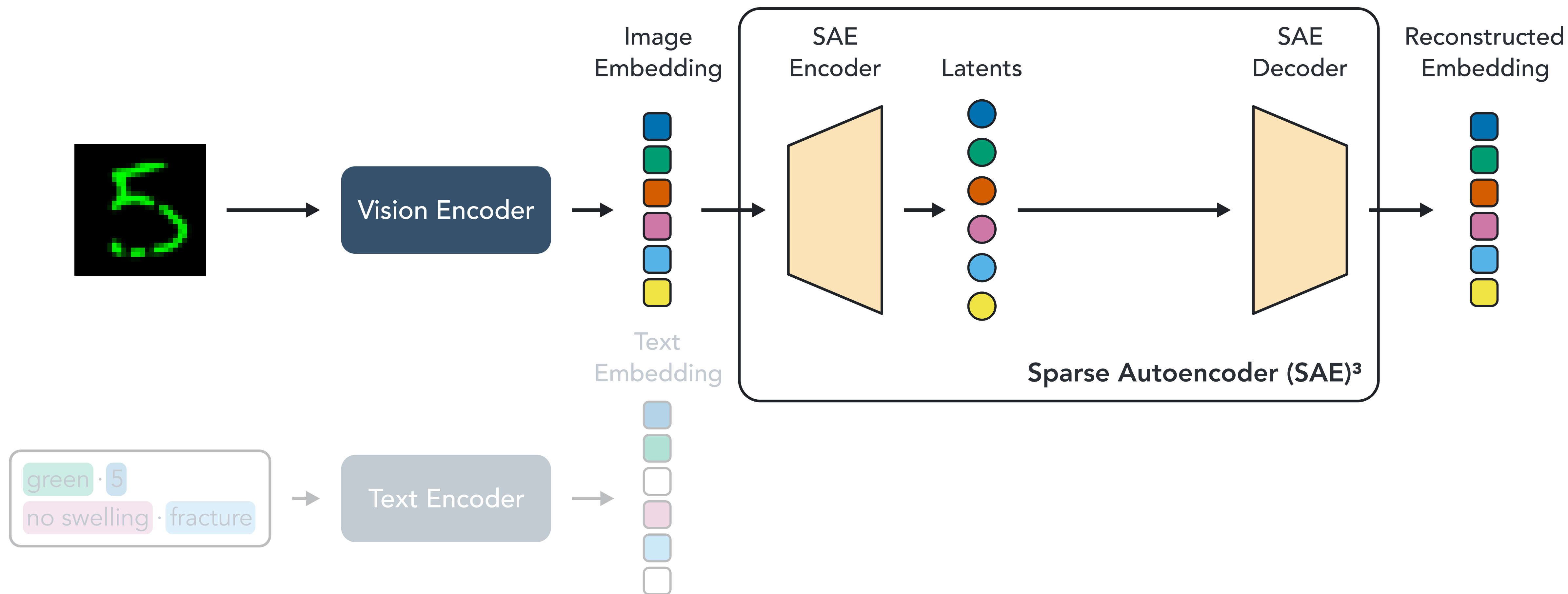
		In Caption
● Digit	0 · 1 · 2 · 3 · 4 · 5 · 6 · 7 · 8 · 9	✓
● Color	gray · red · green · blue · cyan · magenta · yellow	✓
● Thickthinning	thickening · thinning · no thickthinning	✗
● Swelling	swelling · no swelling	✓
● Fracture	fracture · no fracture	✓
● Scaling	large · small	✗

green · 5 · no swelling · fracture

TEVI uses captions as a signal for what to retain

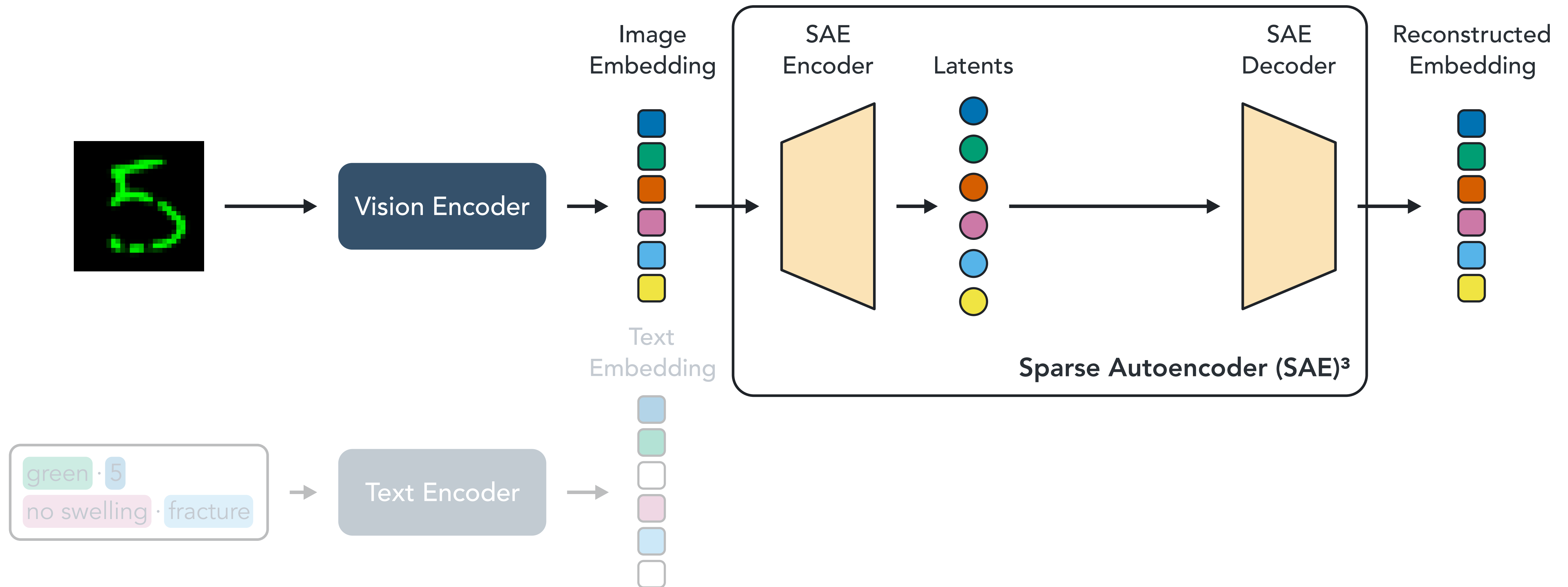


An SAE decomposes the image embedding into concepts



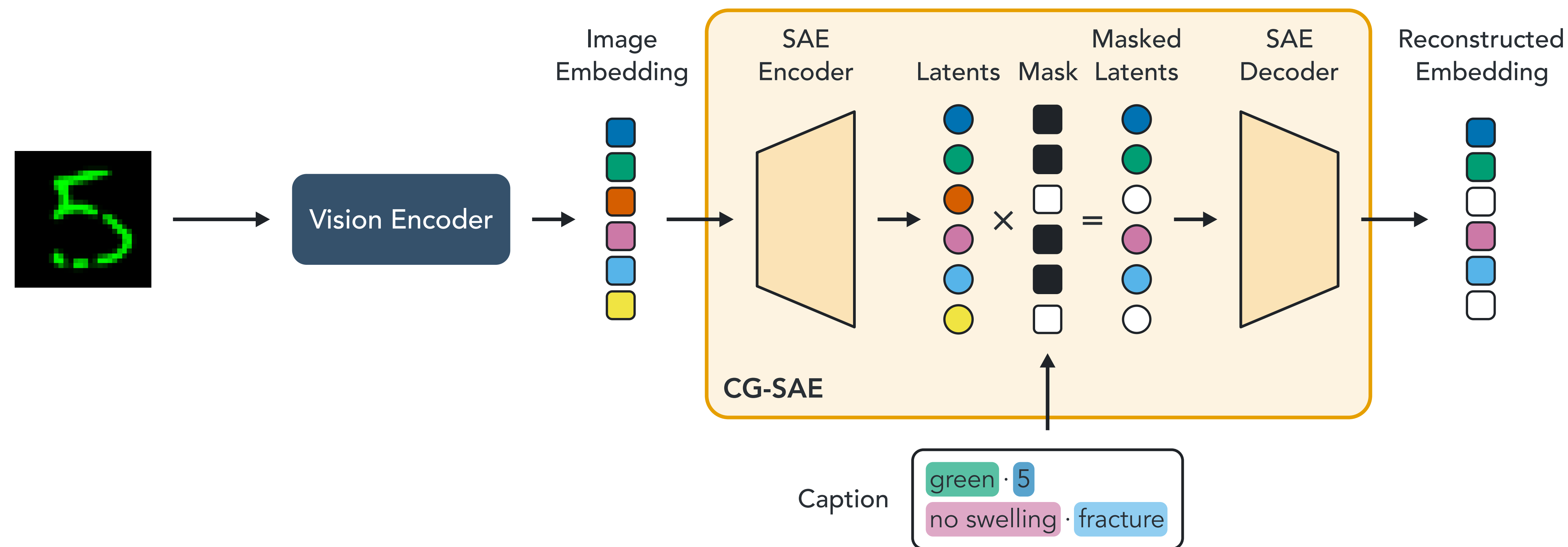
³Bricken et al., Transformer Circuits Thread 2023

With known attributes, the caption can set the mask directly

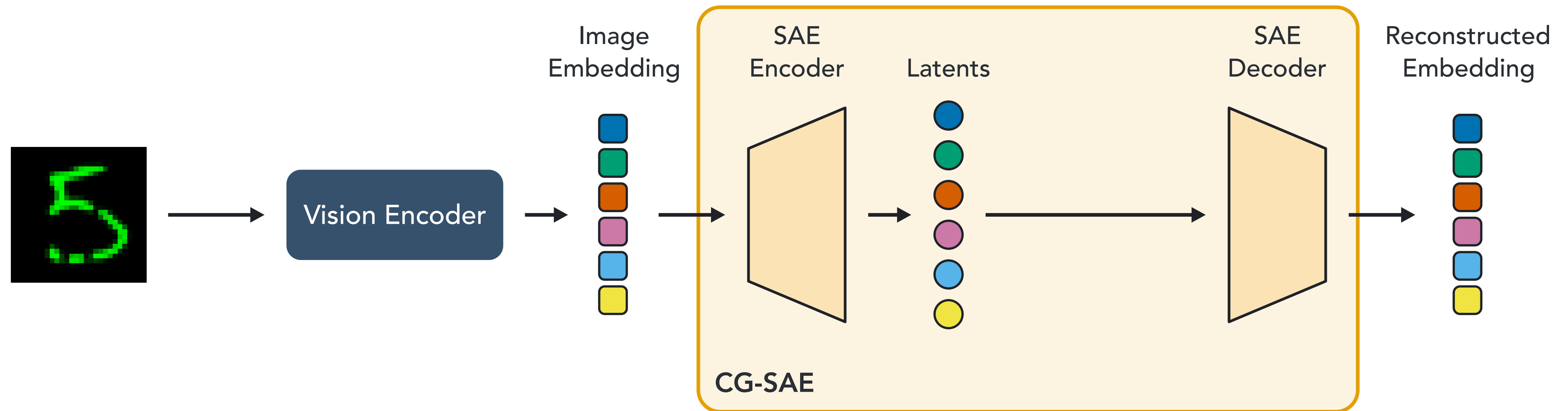


³Bricken et al., Transformer Circuits Thread 2023

With known attributes, the caption can set the mask directly



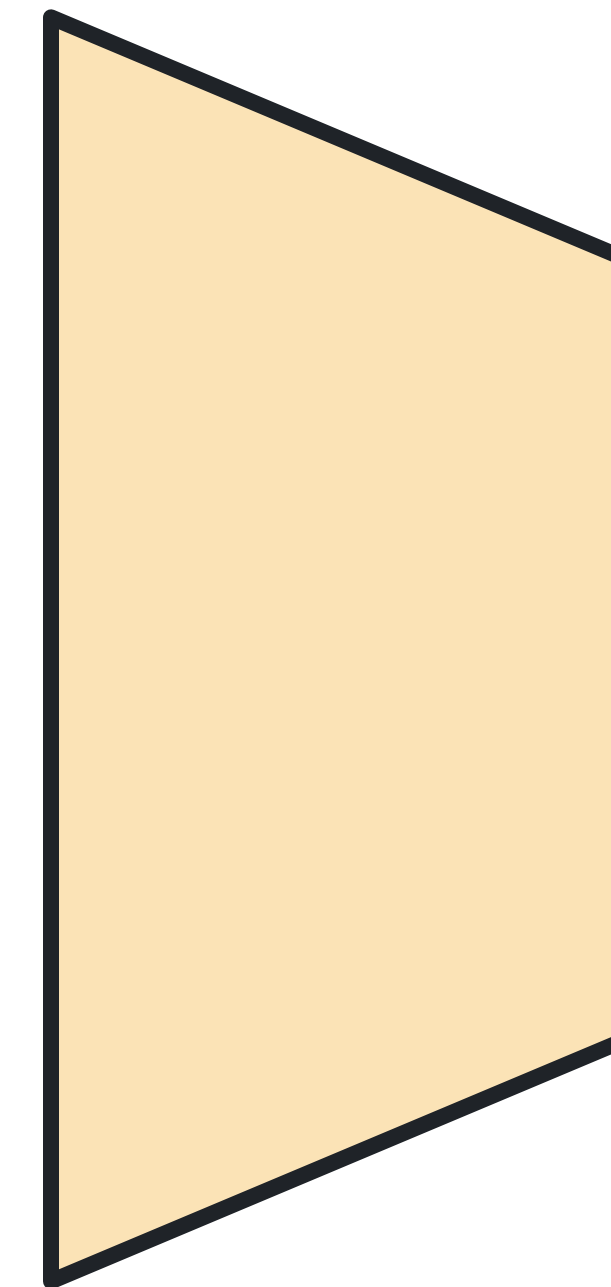
Each SAE latent represents one of the 26 attributes



Each SAE latent represents one of the 26 attributes

- 0 · 1 · 2 · 3 · 4 · 5 · 6 · 7 · 8 · 9
- gray · red · green · blue · cyan · magenta · yellow
- thickening · thinning · no thickthinning
- swelling · no swelling
- fracture · no fracture
- large · small

SAE
Decoder



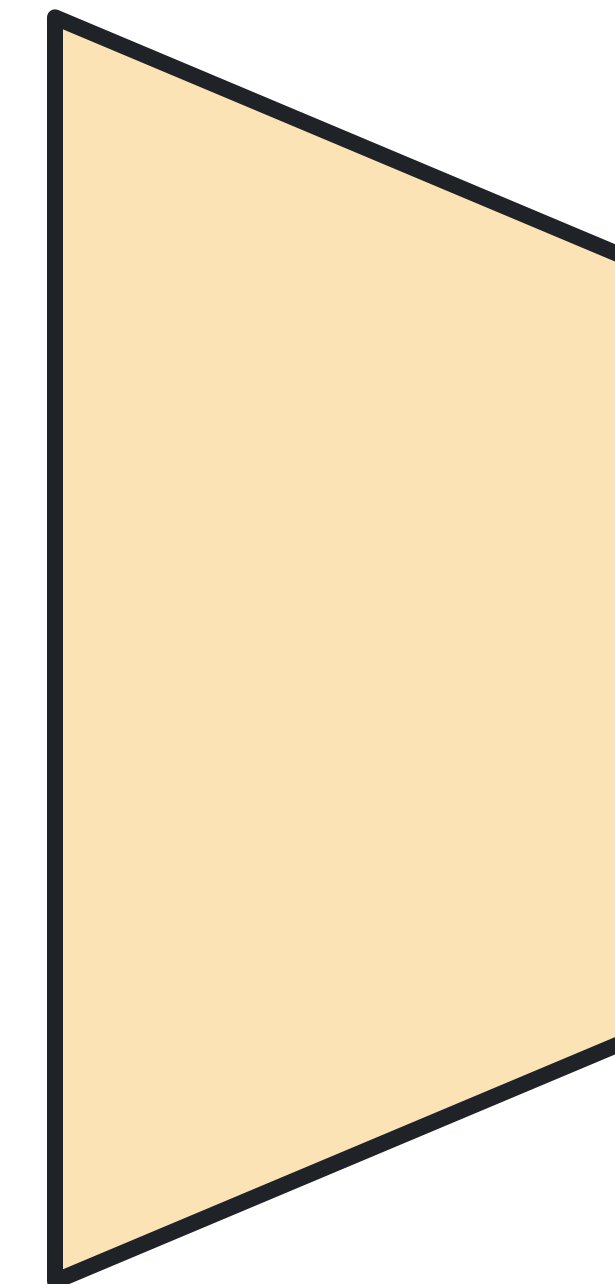
Each SAE latent represents one of the 26 attributes

- 0 · 1 · 2 · 3 · 4 · 5 · 6 · 7 · 8 · 9
- gray · red · green · blue · cyan · magenta · yellow
- thickening · thinning · no thickthinning
- swelling · no swelling
- fracture · no fracture
- large · small



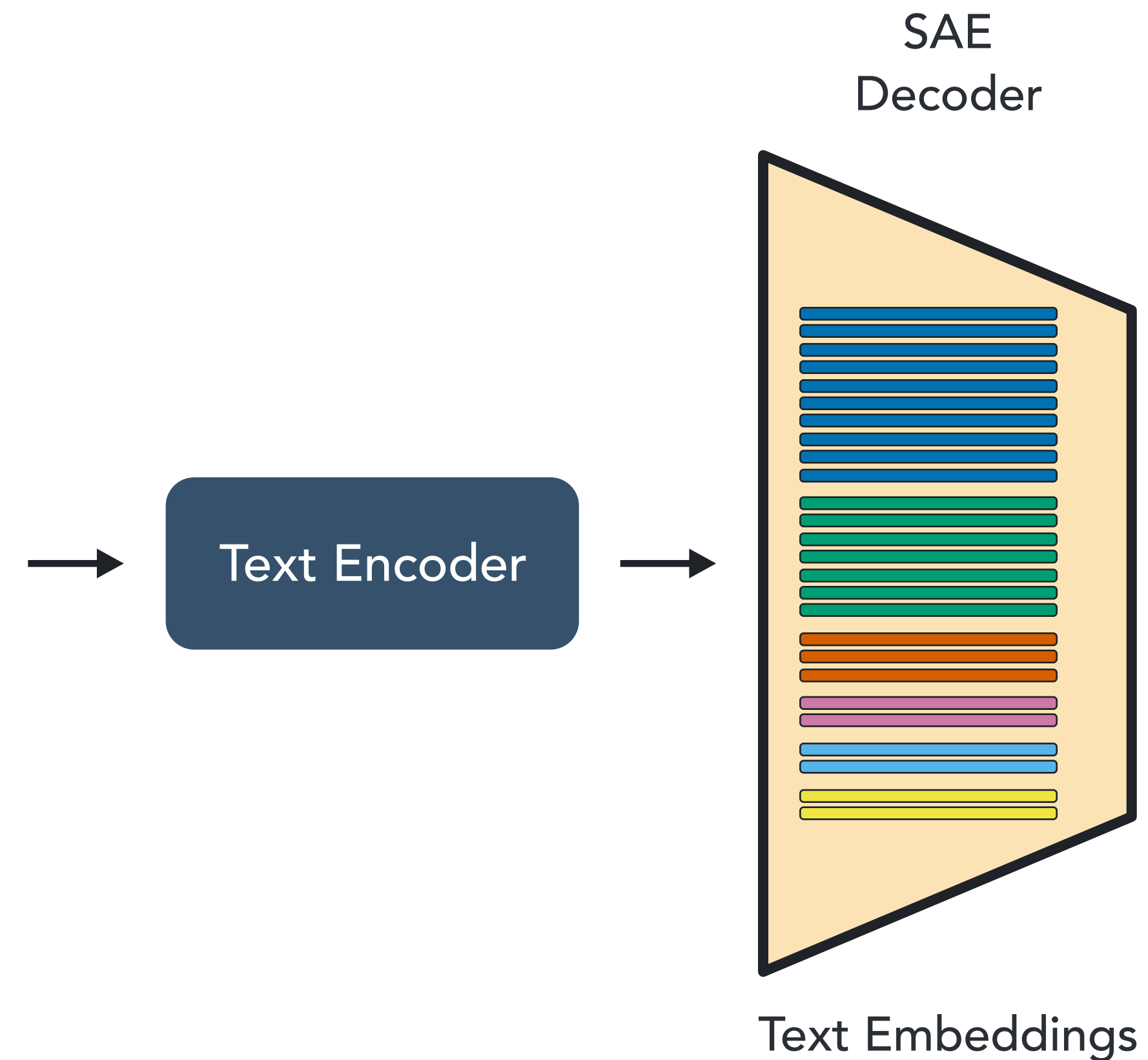
Text Encoder

SAE
Decoder

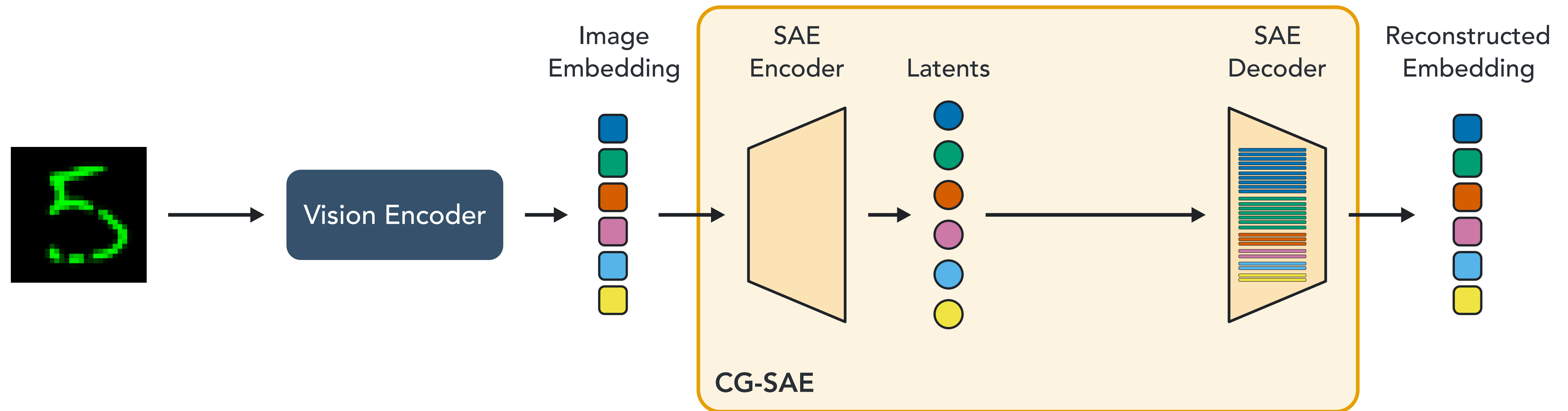


Each SAE latent represents one of the 26 attributes

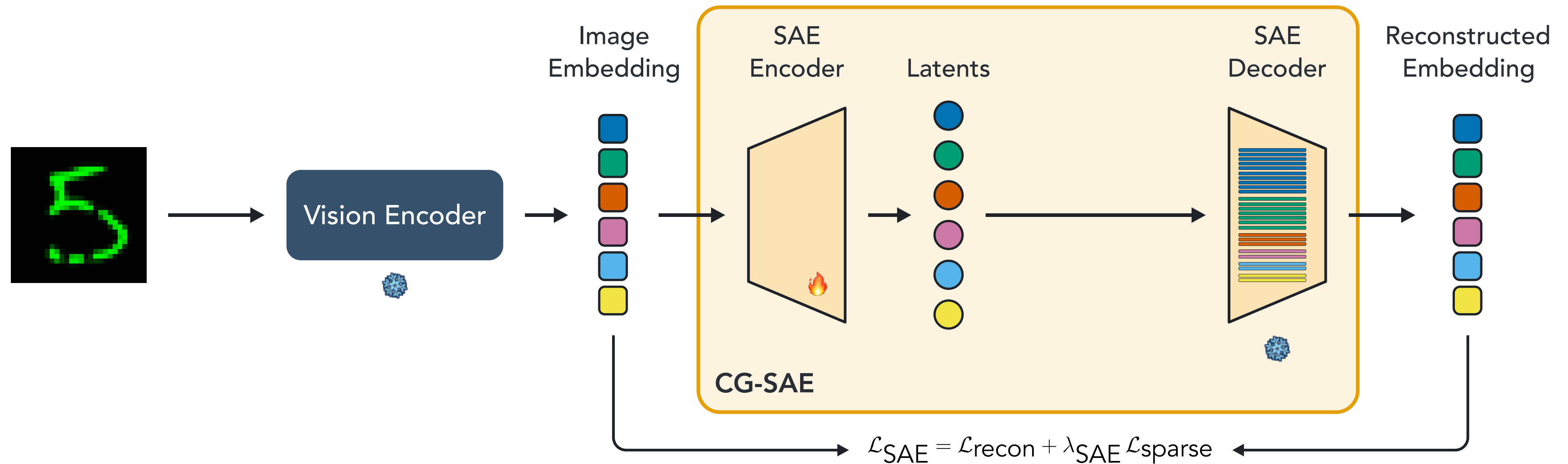
- 0 · 1 · 2 · 3 · 4 · 5 · 6 · 7 · 8 · 9
- gray · red · green · blue · cyan · magenta · yellow
- thickening · thinning · no thickthinning
- swelling · no swelling
- fracture · no fracture
- large · small



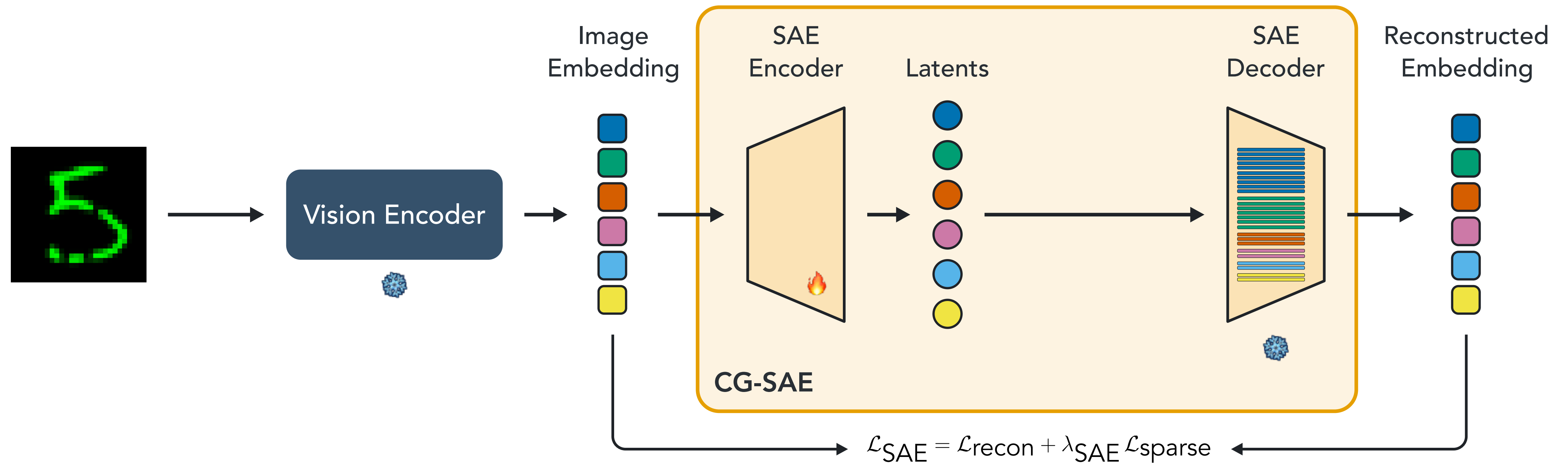
Each SAE latent represents one of the 26 attributes



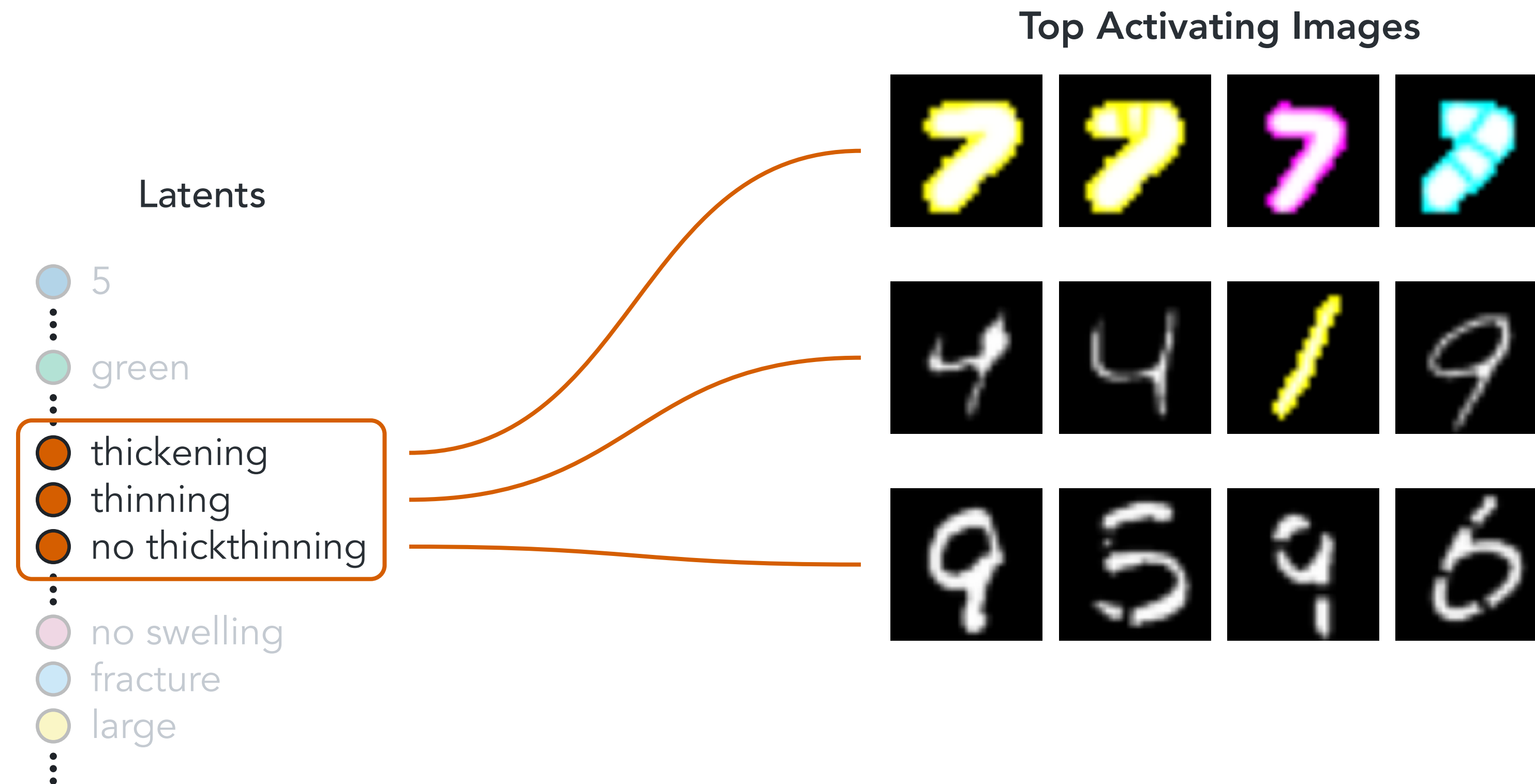
With the decoder fixed, only the SAE encoder is trained



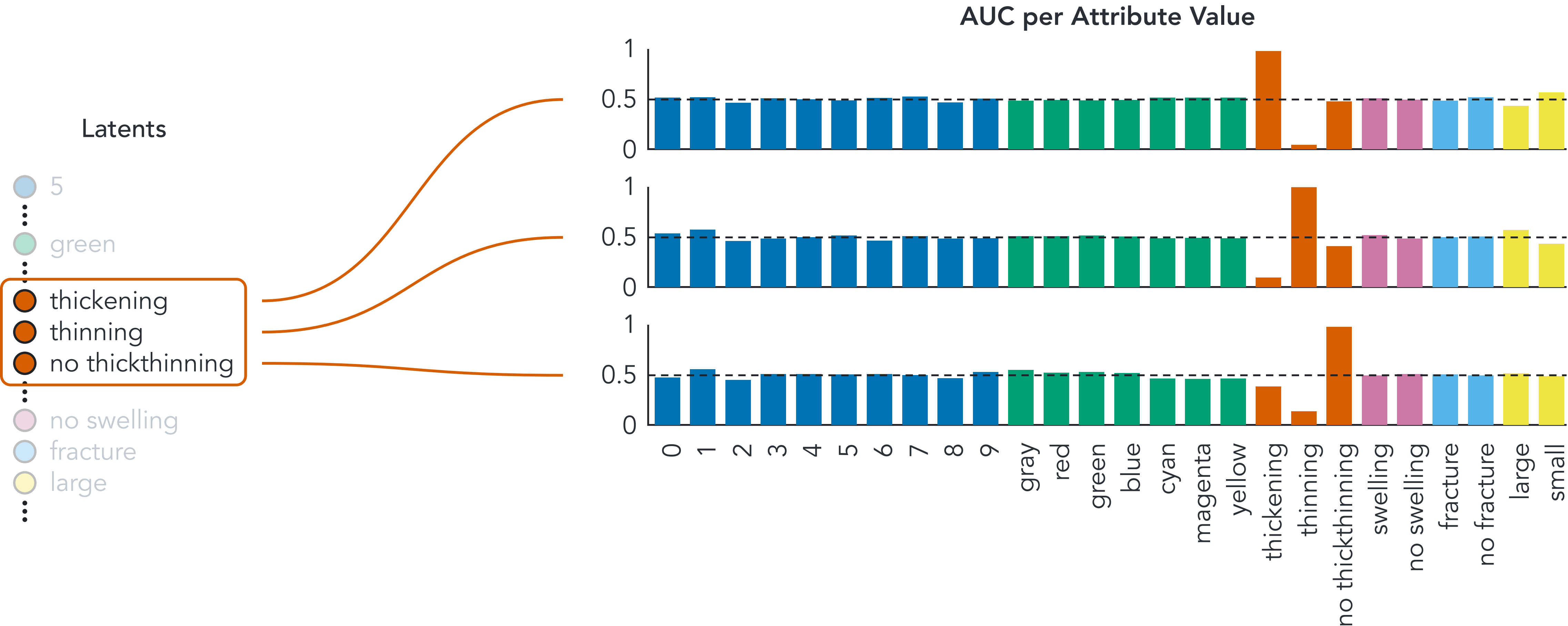
Each latent encodes the attribute assigned to it



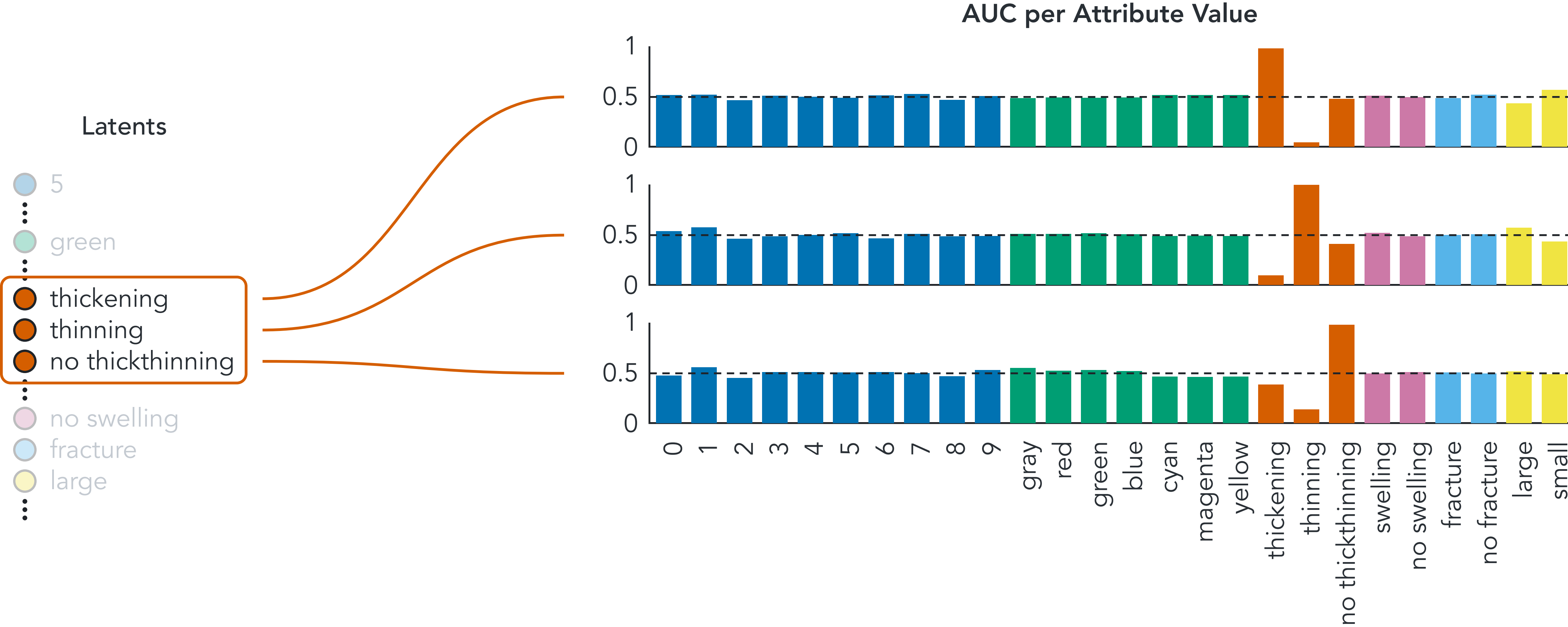
Each latent encodes the attribute assigned to it



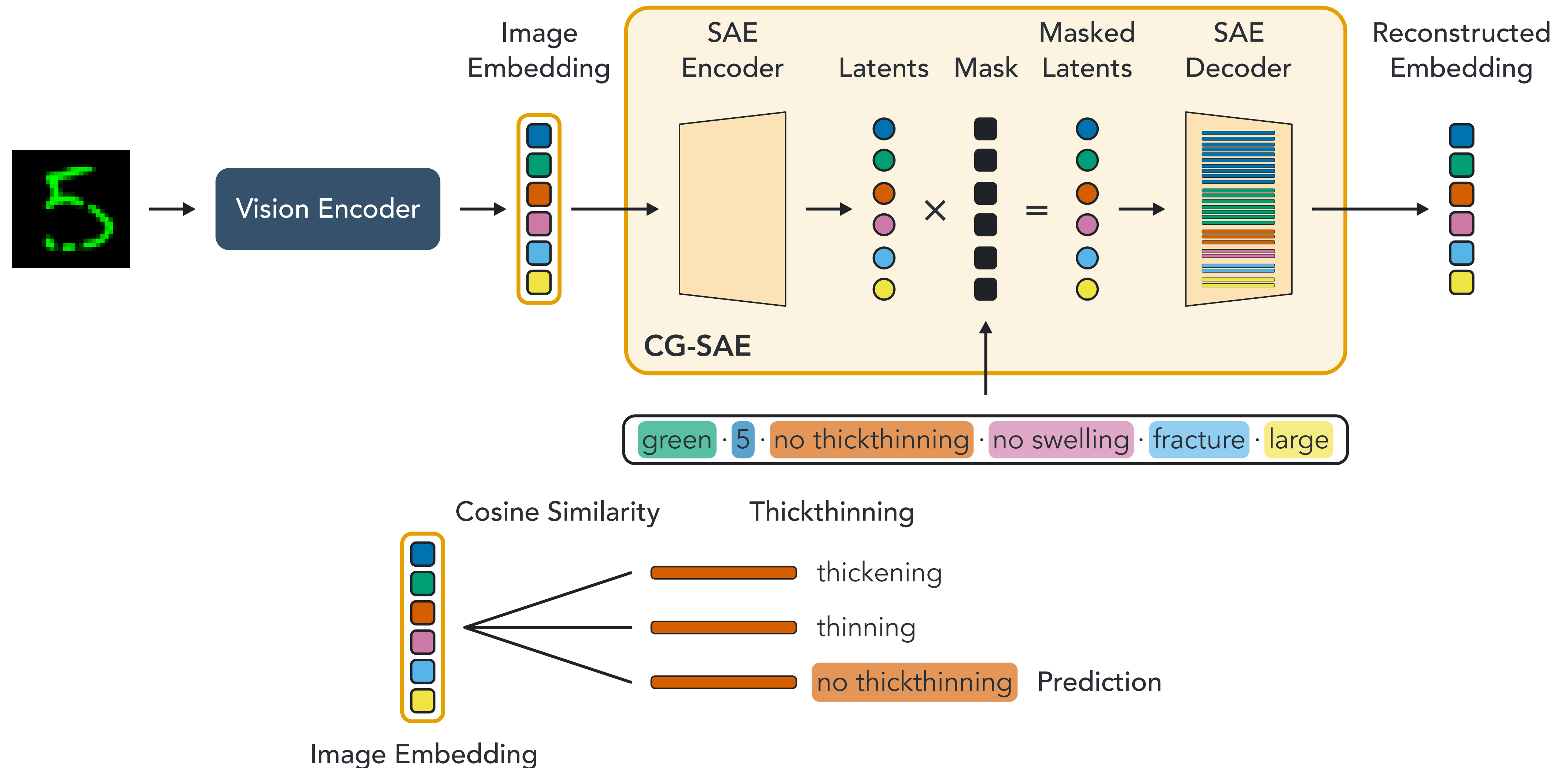
Each latent encodes the attribute assigned to it



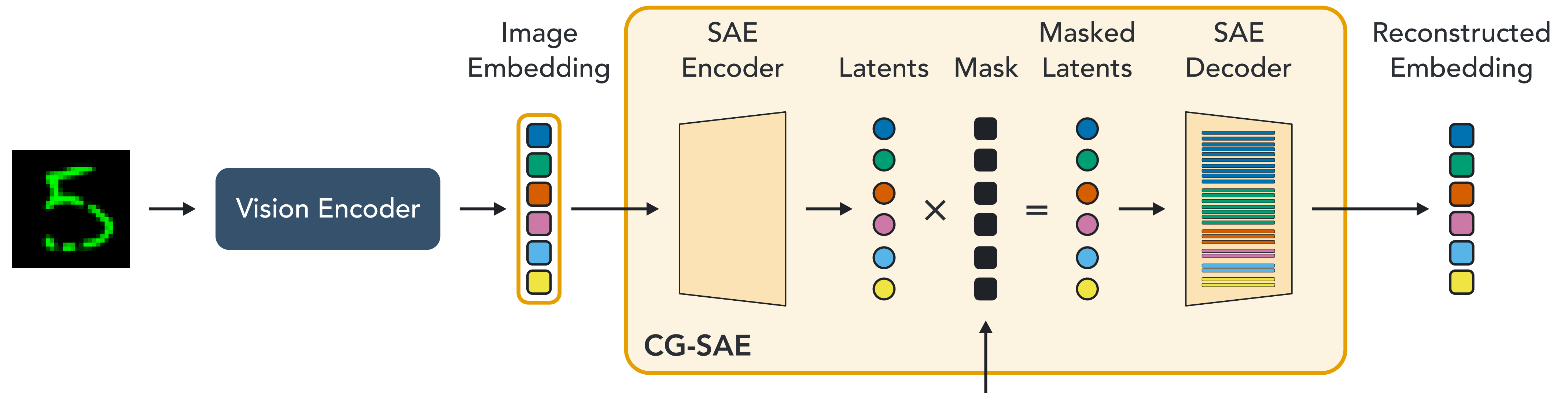
Each attribute is classified zero-shot from the embedding



Masking latents removes only their attribute



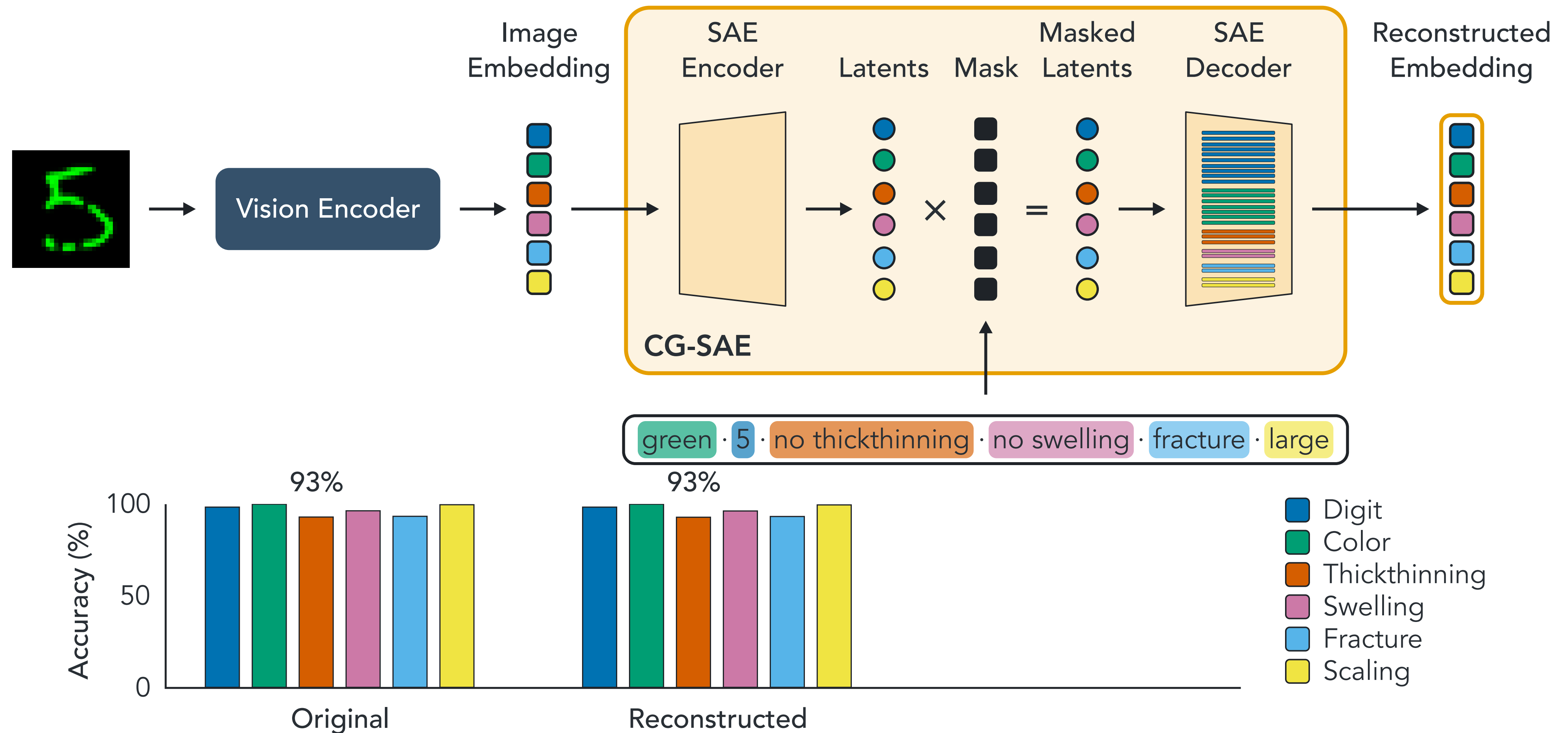
Masking latents removes only their attribute



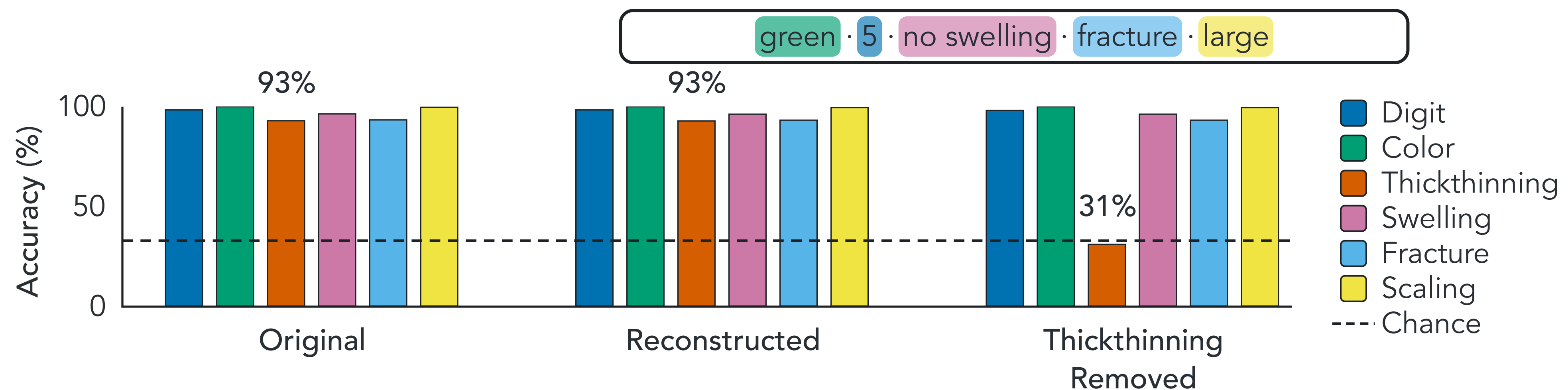
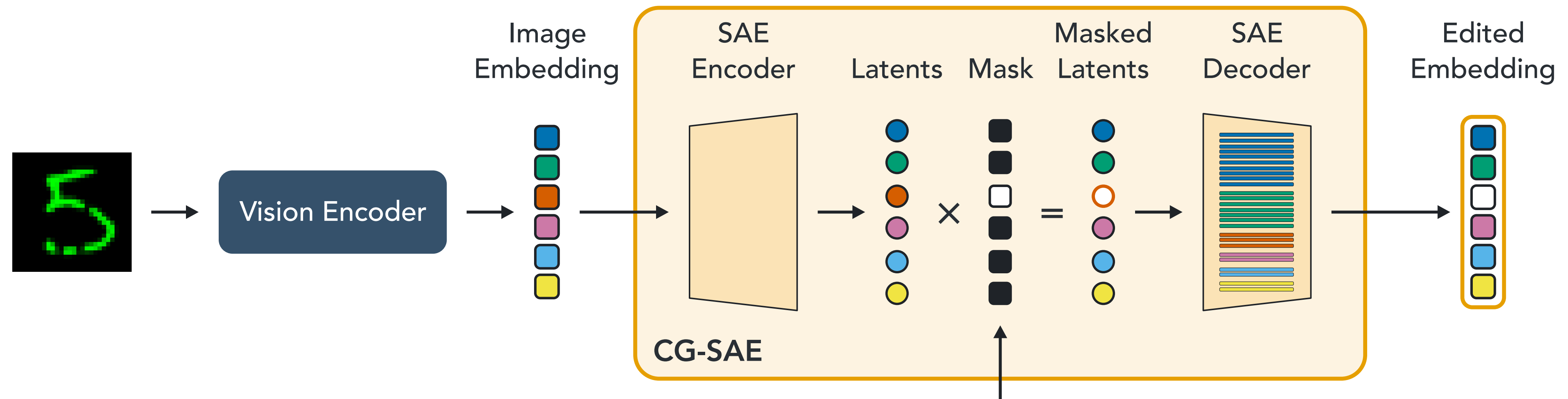
green · 5 · no thickthinning · no swelling · fracture · large



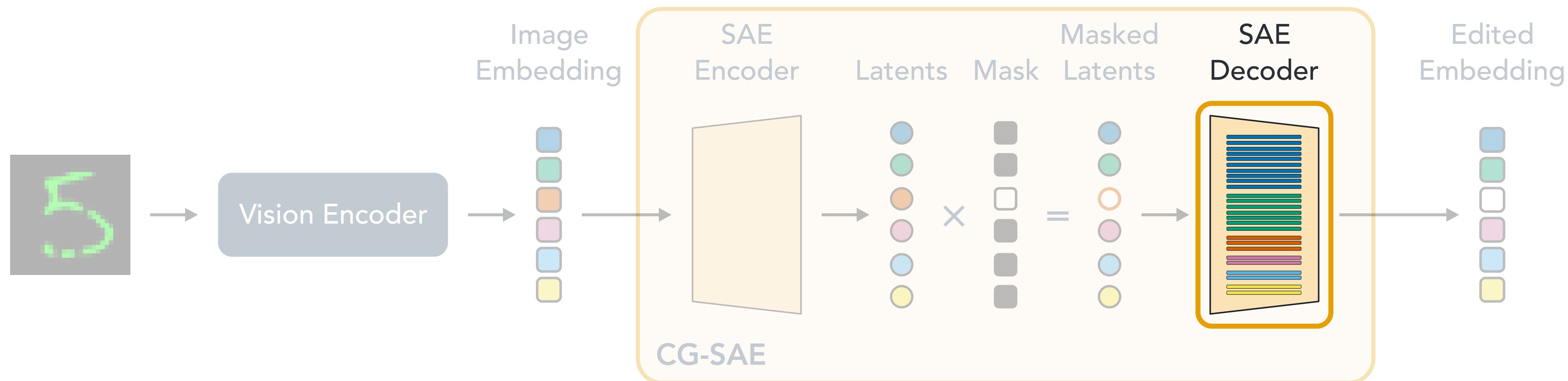
Masking latents removes only their attribute



Masking latents removes only their attribute



CG-SAEs edit embeddings, but need predefined concepts



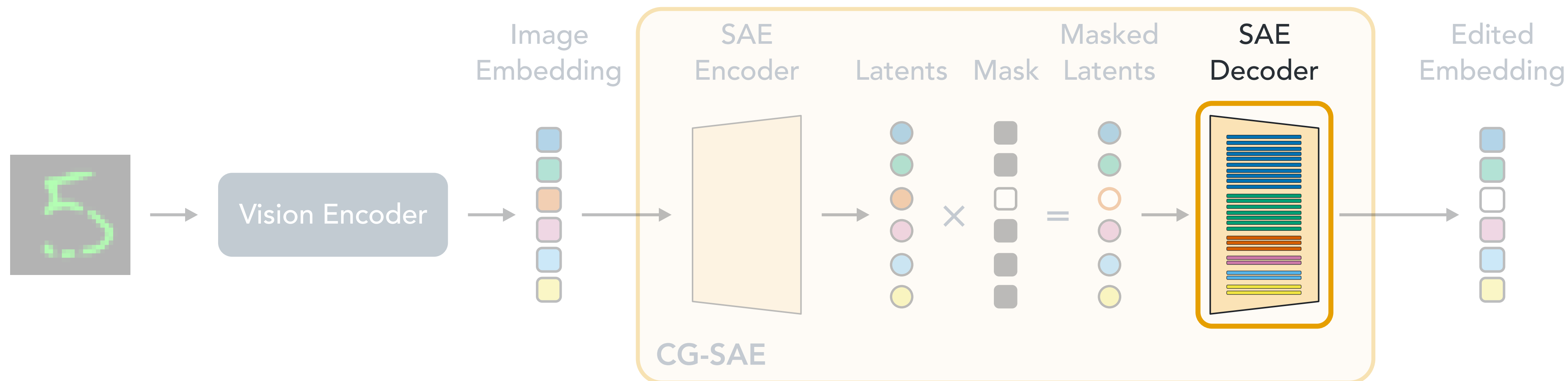
✓ The latents disentangle the attributes

✗ The decoder is fixed to known concepts

✓ Masking removes only that attribute

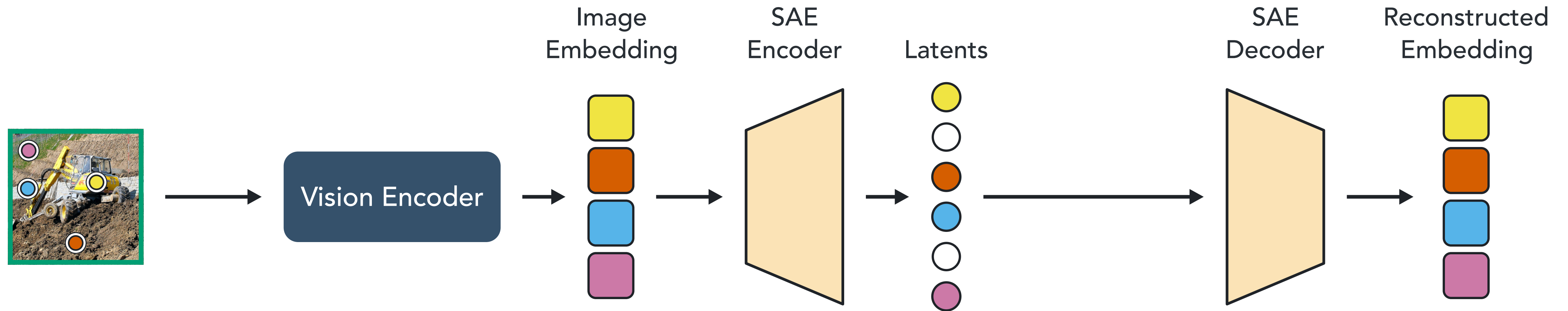
✗ CLIP models have no such fixed set

TEVI relaxes the assumption, learns what to mask

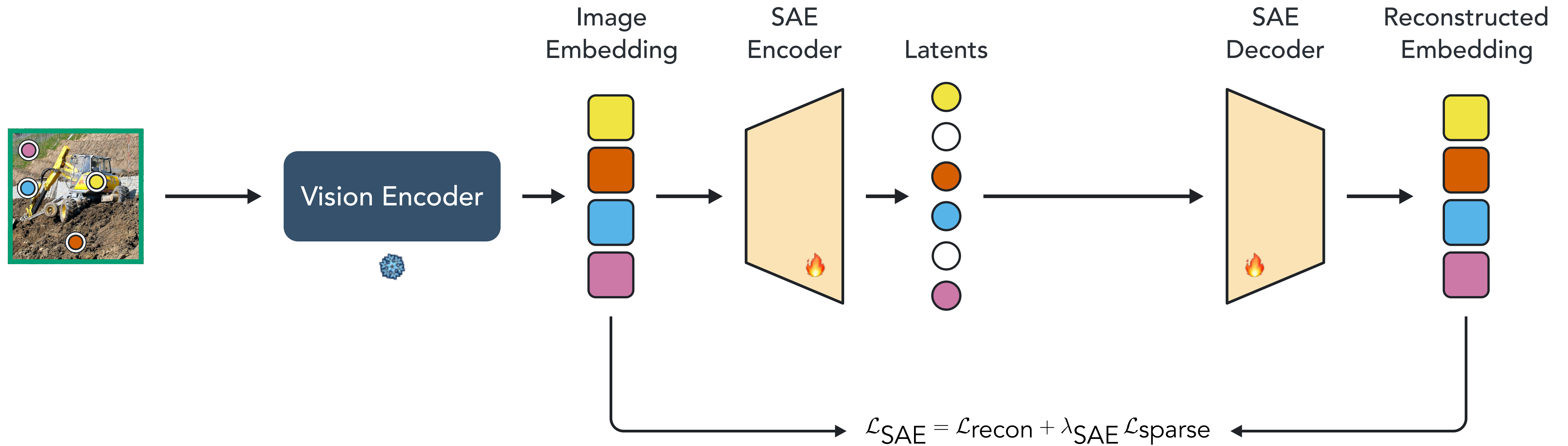


- ✓ The latents disentangle the attributes
- ✓ Masking removes only that attribute
- TEVI learns what to mask given the caption
- ✗ The decoder is fixed to known concepts
- ✗ CLIP models have no such fixed set

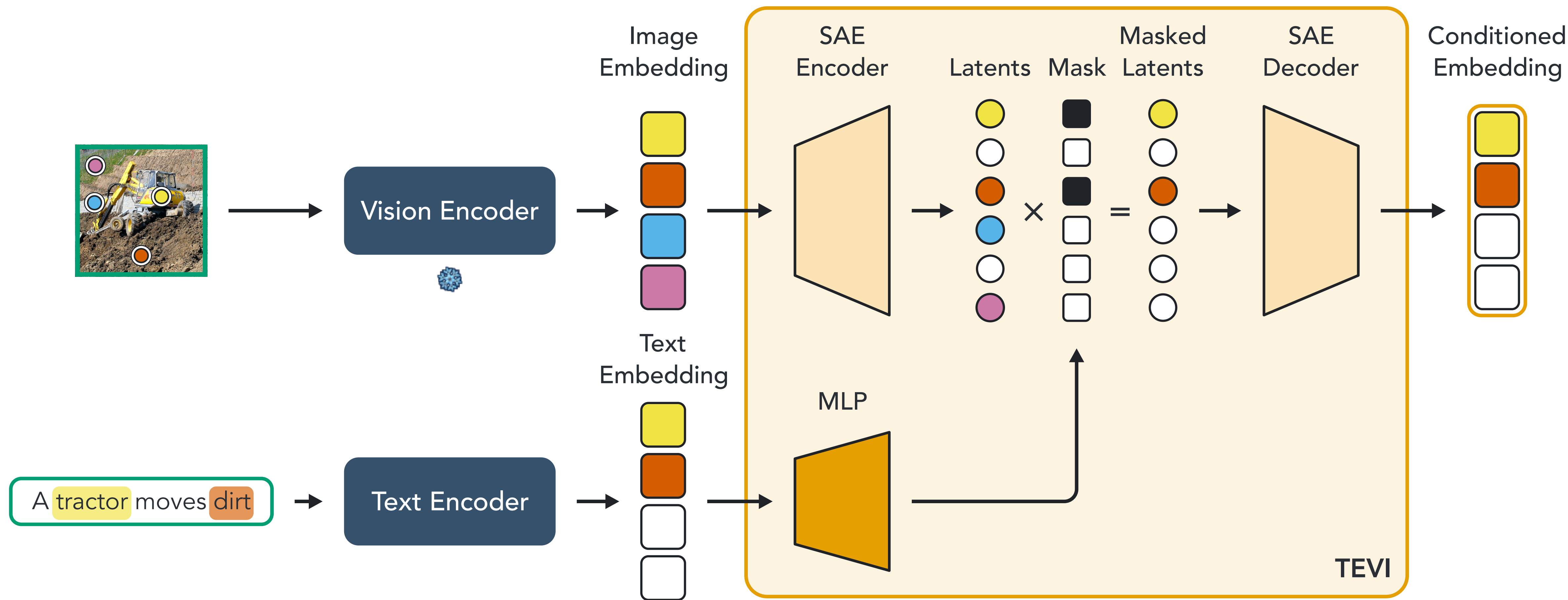
TEVI first trains a conventional SAE on CLIP image embeddings



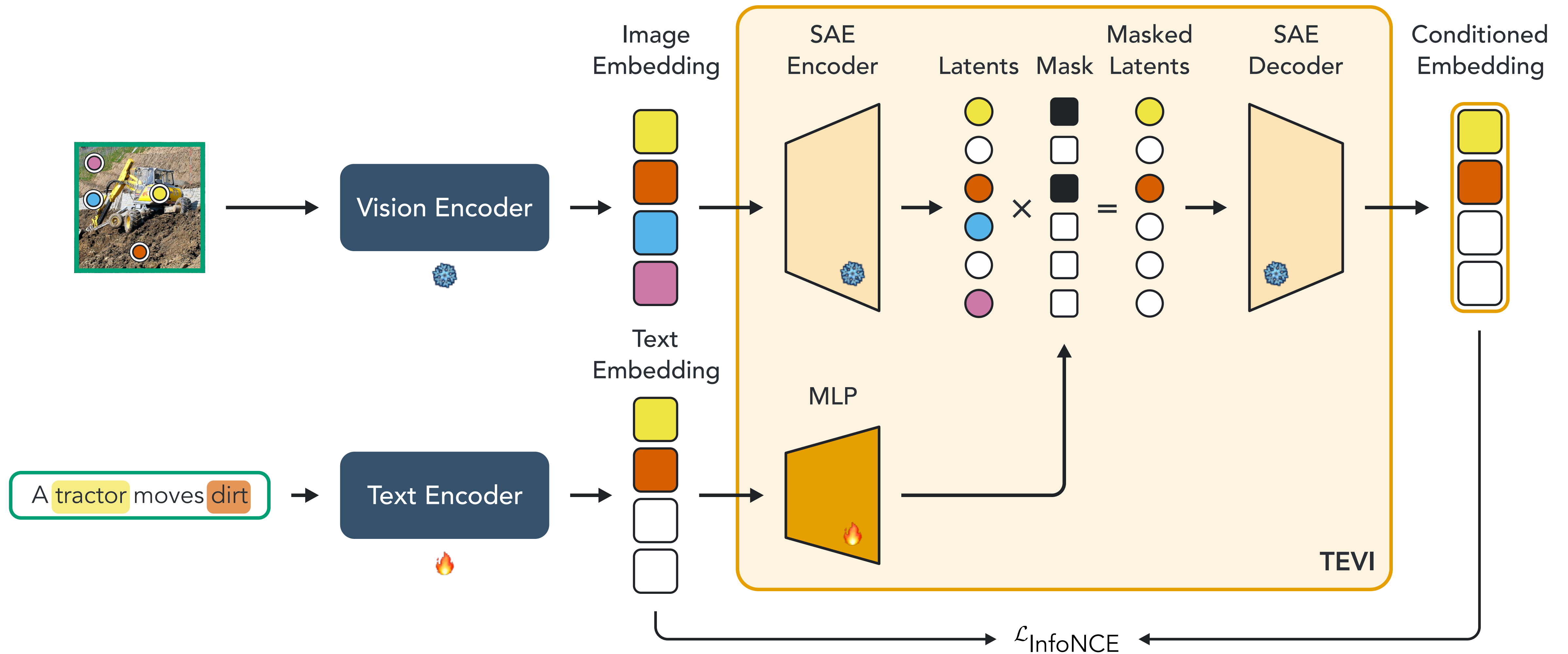
TEVI first trains a conventional SAE on CLIP image embeddings



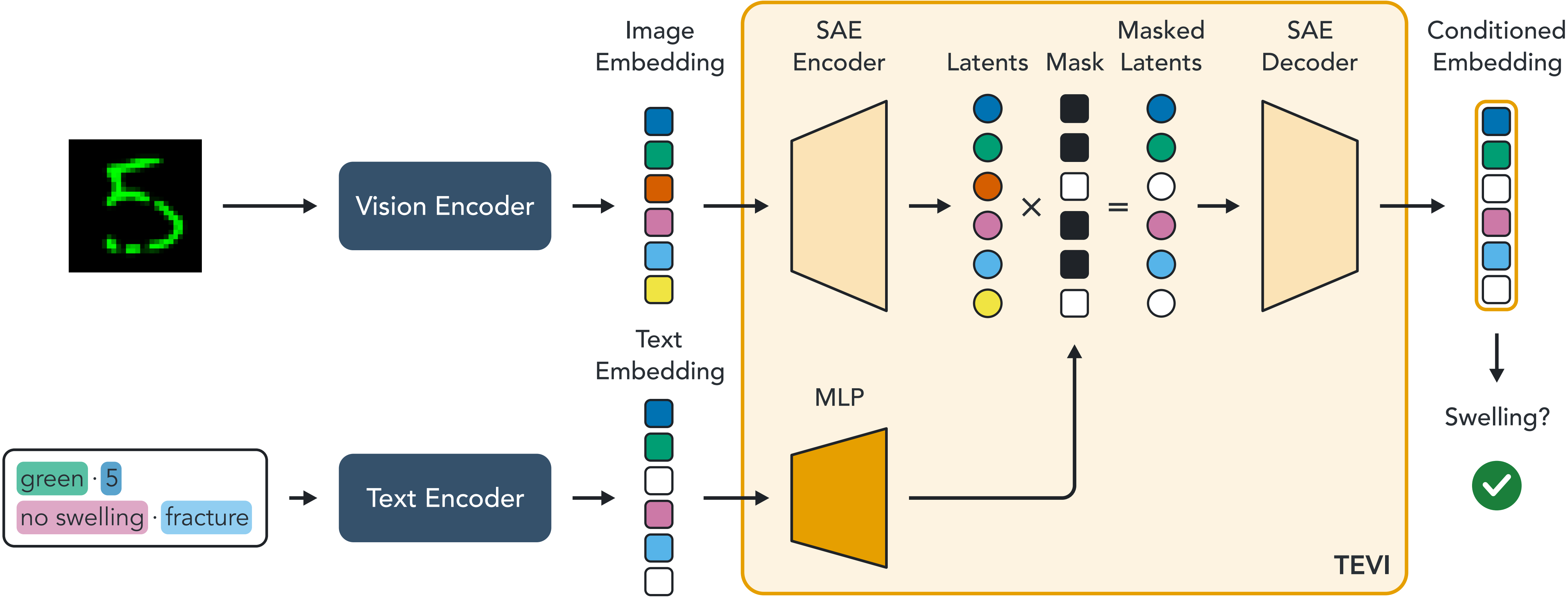
TEVI then learns a mask over the SAE latents based on the caption



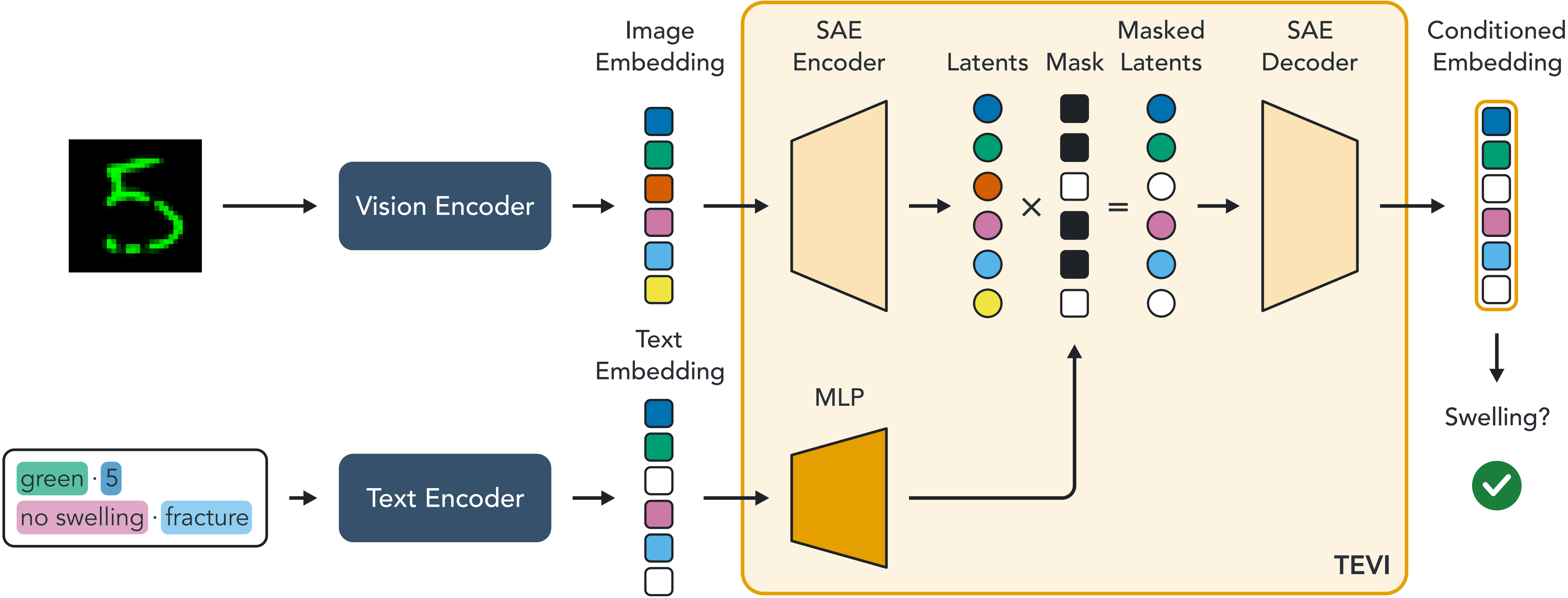
TEVI then learns a mask over the SAE latents based on the caption



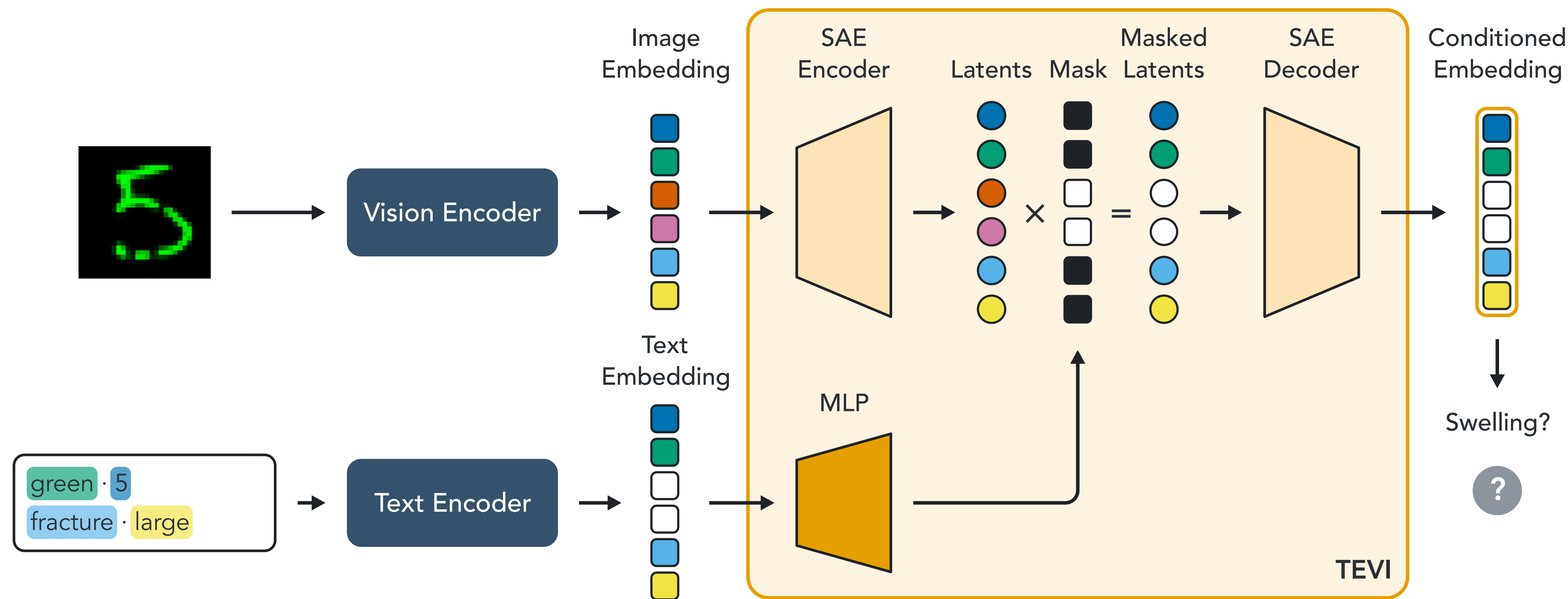
Attributes in the caption should be kept



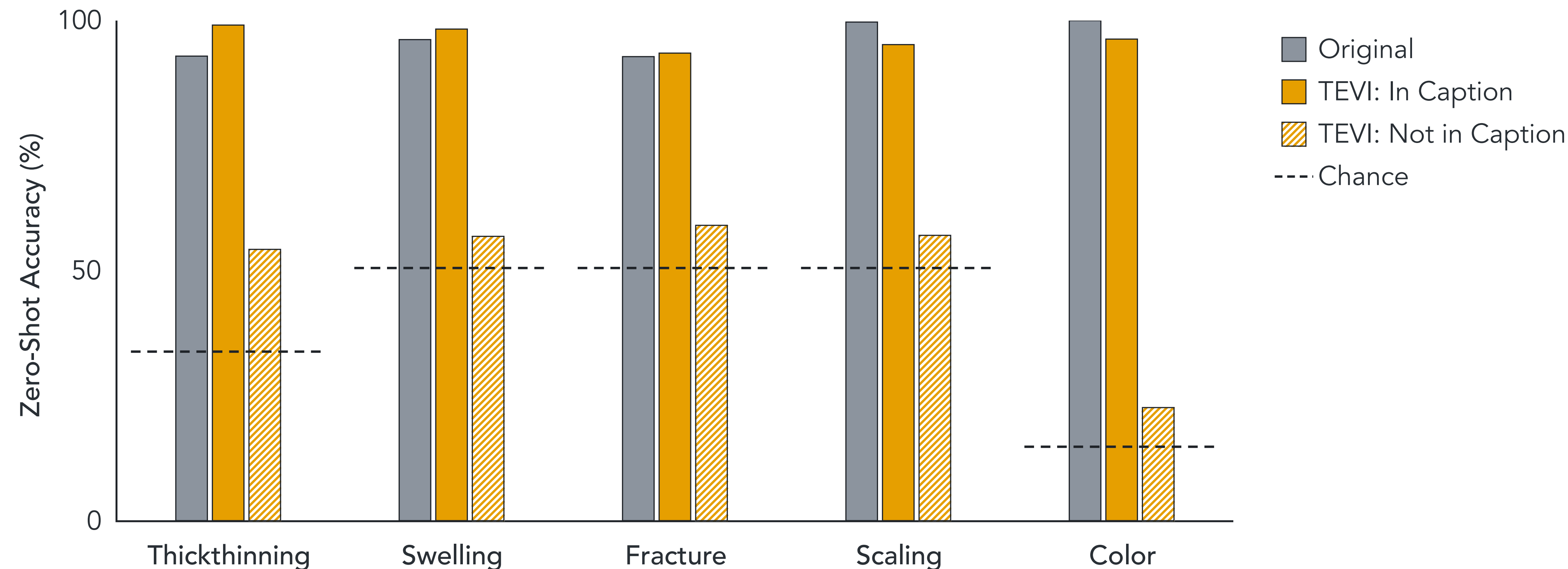
Attributes not in the caption should drop to chance



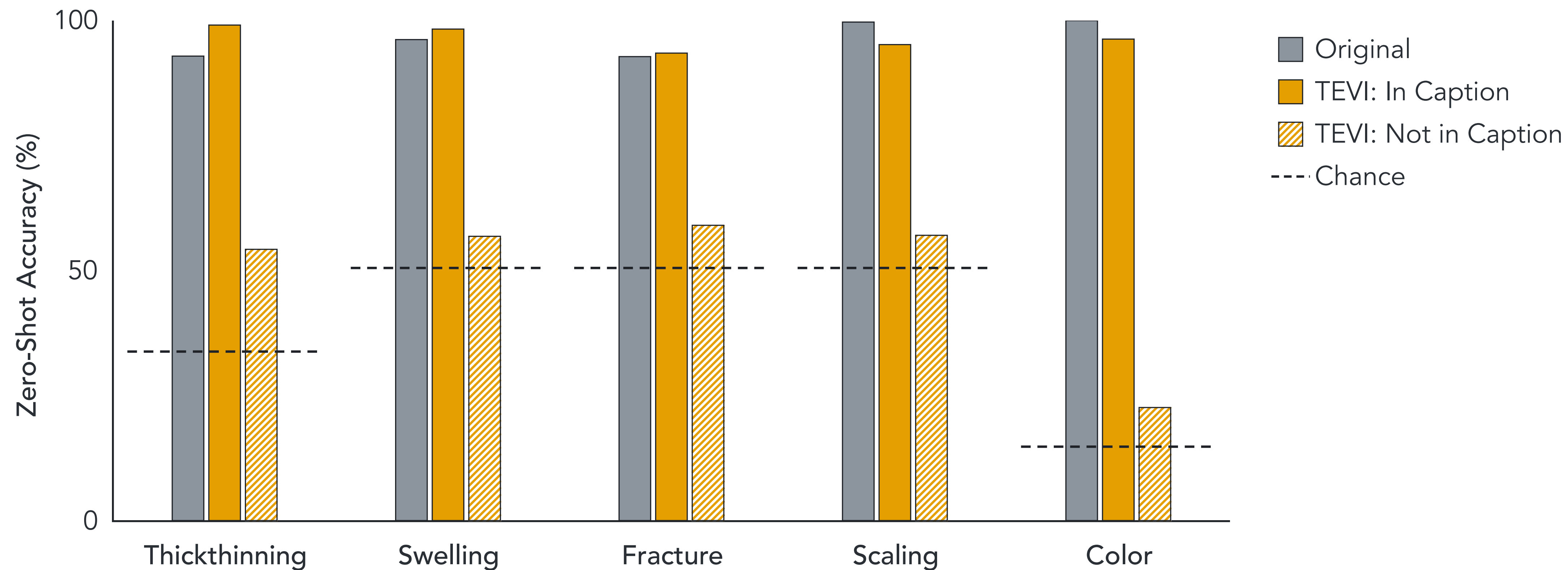
Attributes not in the caption should drop to chance



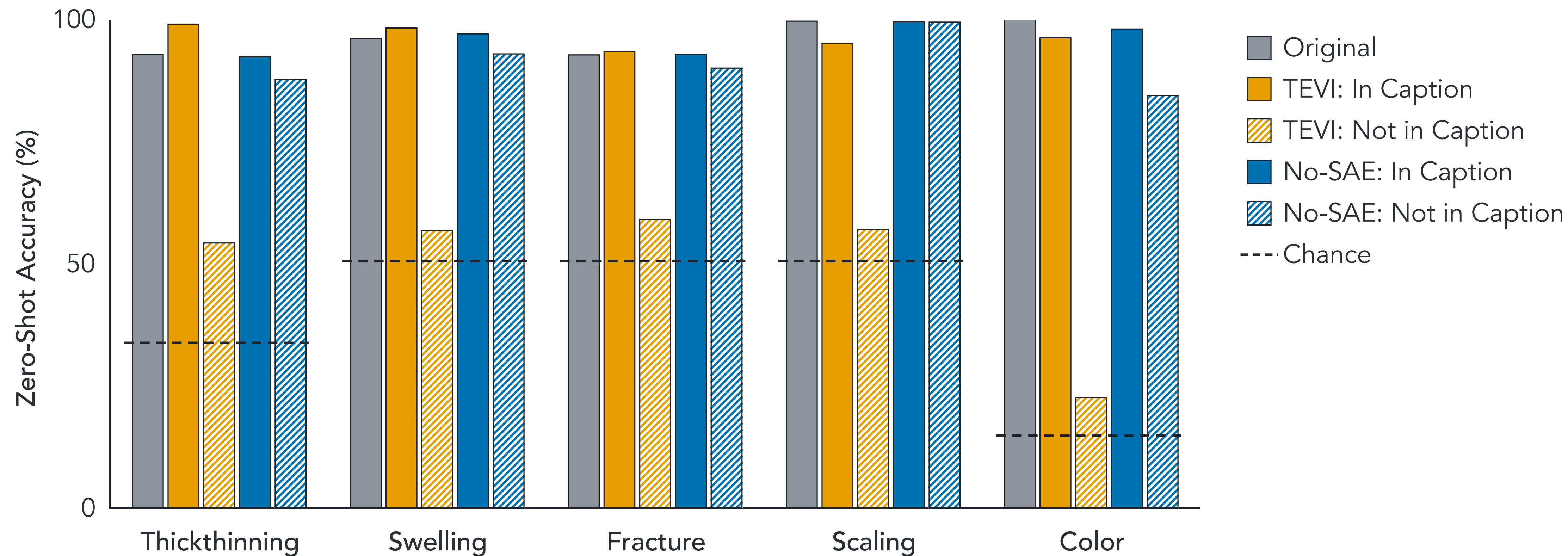
TEVI keeps attributes in the caption and removes the rest



Without the SAE, the mask has almost no effect

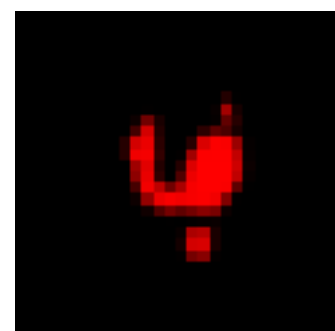


Without the SAE, the mask has almost no effect



TEVI reorders retrieval based on the conditioning caption

Query Image



Vision Encoder



Image Embedding



Candidate Captions

red · fracture > red · no fracture ≈ blue · fracture > blue · no fracture

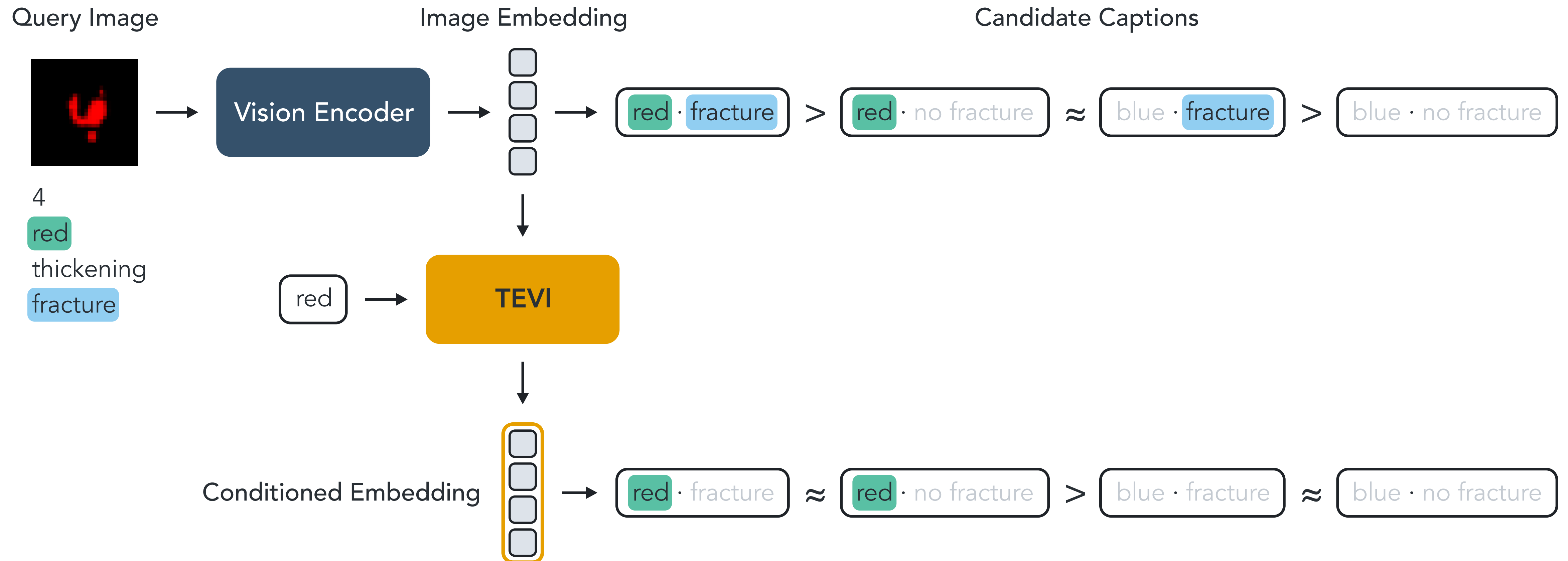
4

red

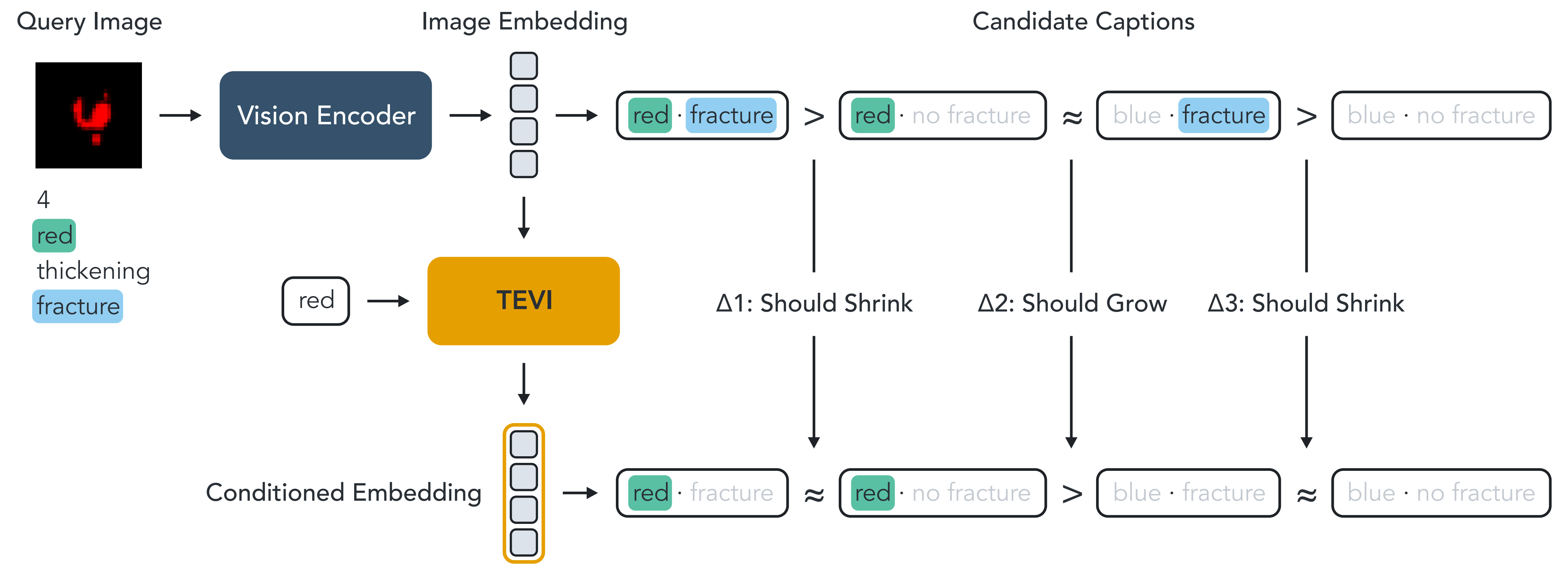
thickening

fracture

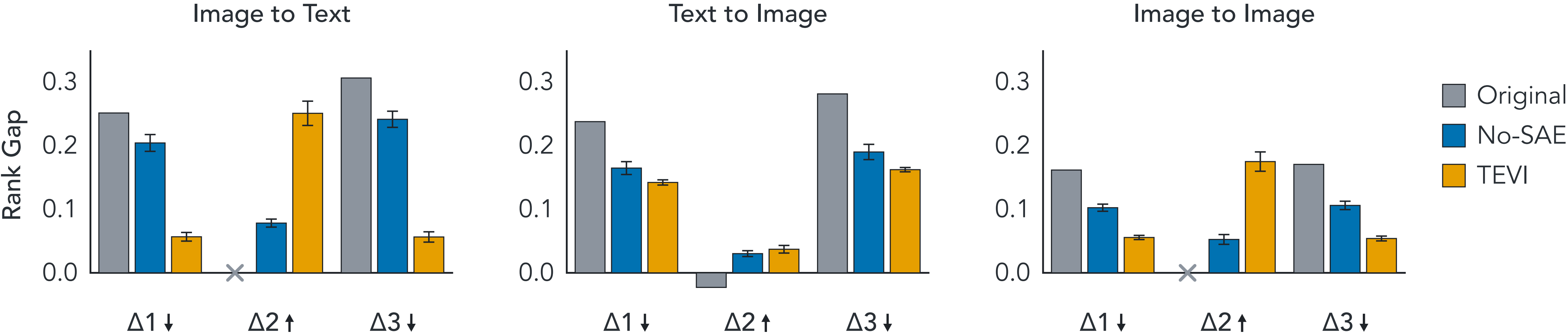
After conditioning on red, only color should count



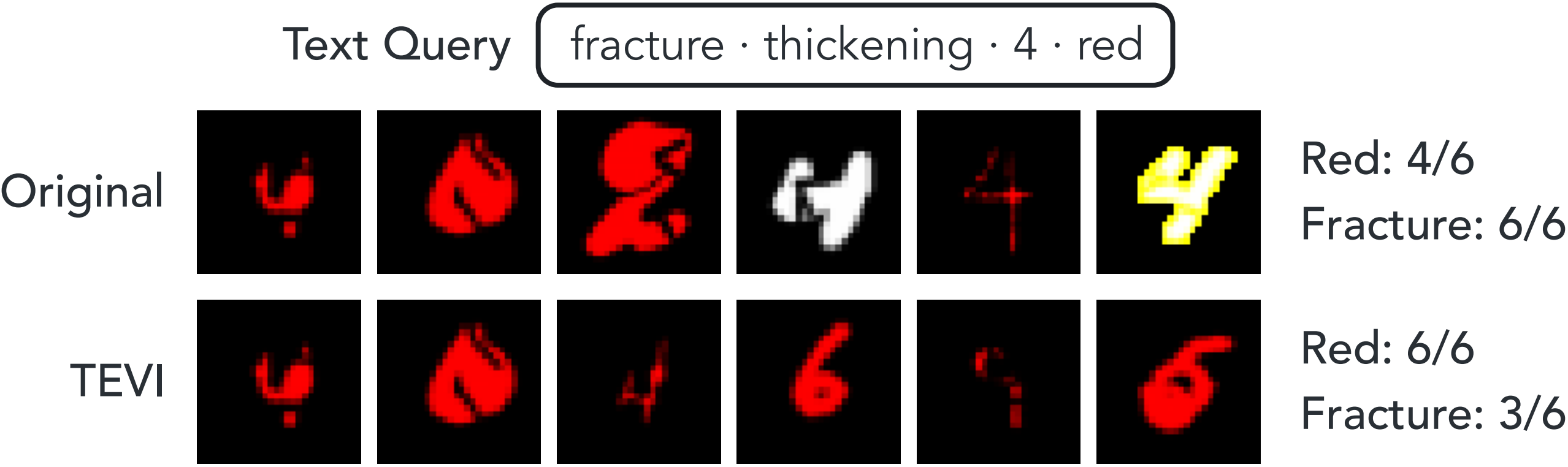
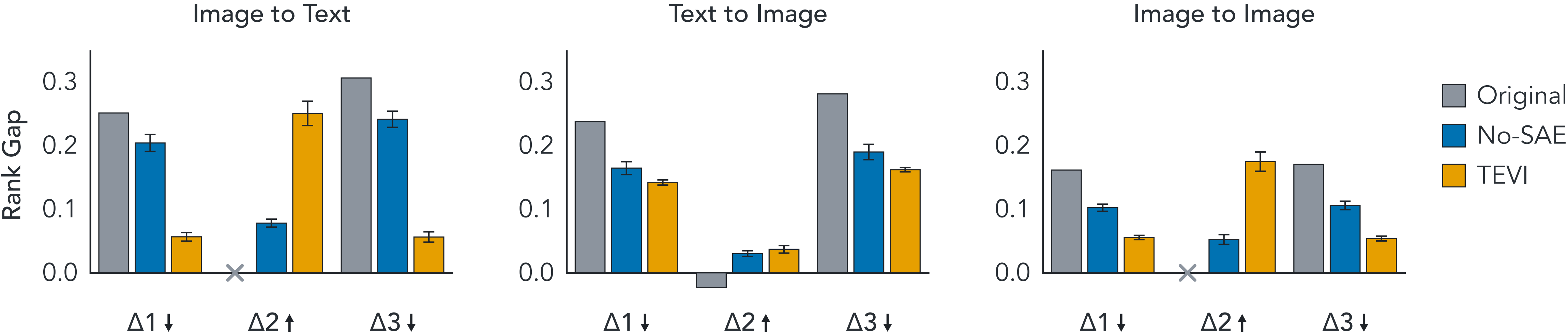
After conditioning on red, only color should count



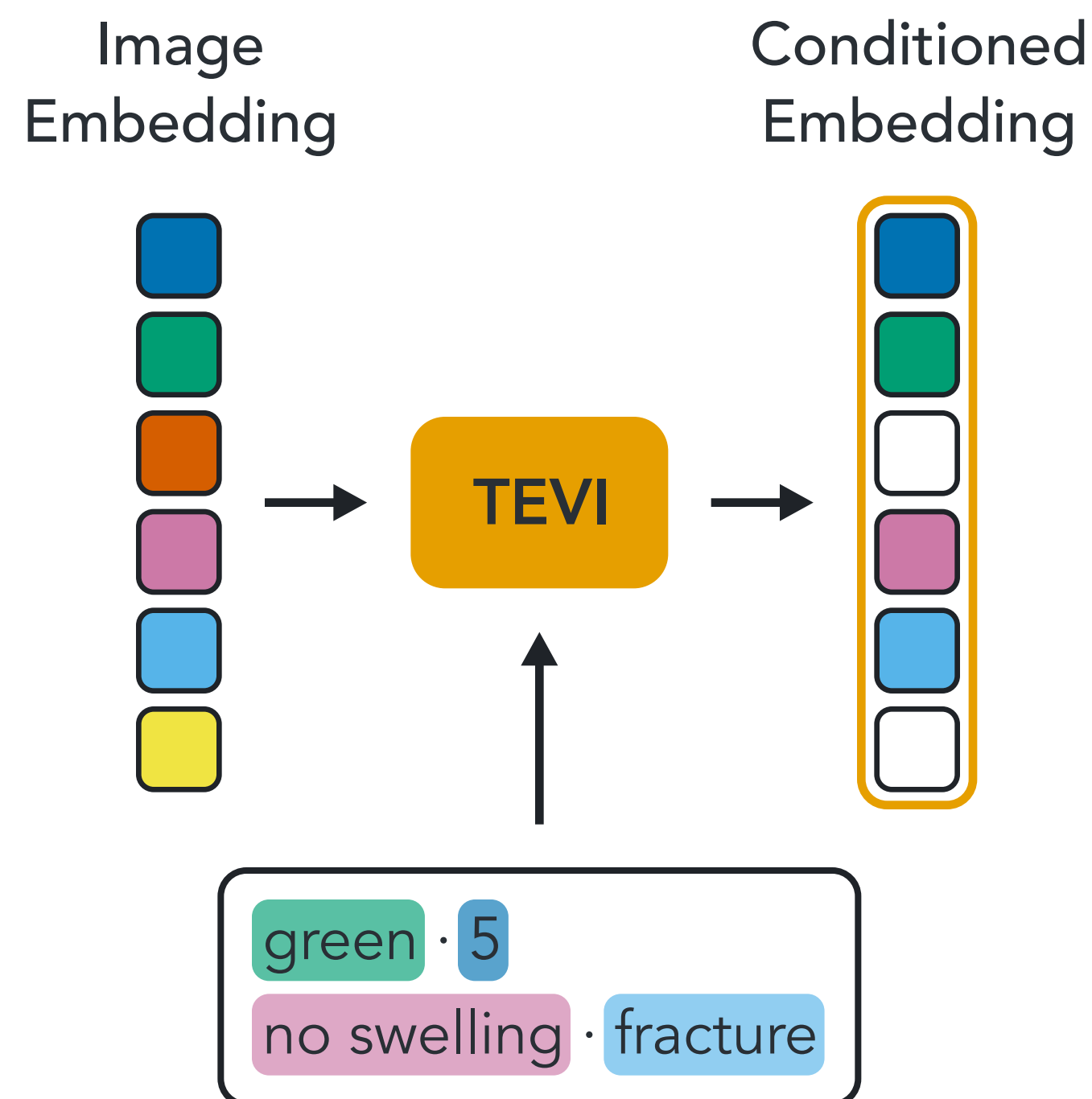
TEVI moves all three gaps in the expected direction



TEVI moves all three gaps in the expected direction



TEVI preserves caption attributes and discards the rest



- ✓ Attributes in the caption are kept, the rest drop to chance
- ✓ Conditioning reorders retrieval in the expected direction
- ✓ Both effects are much weaker without the SAE
- ✗ MAD captions are short lists of known attributes
- ✗ Real captions are free-form and noisy

We then extend TEVI to CLIP models on natural images

Controlled Setup



Synthetic Digits with Known Attributes



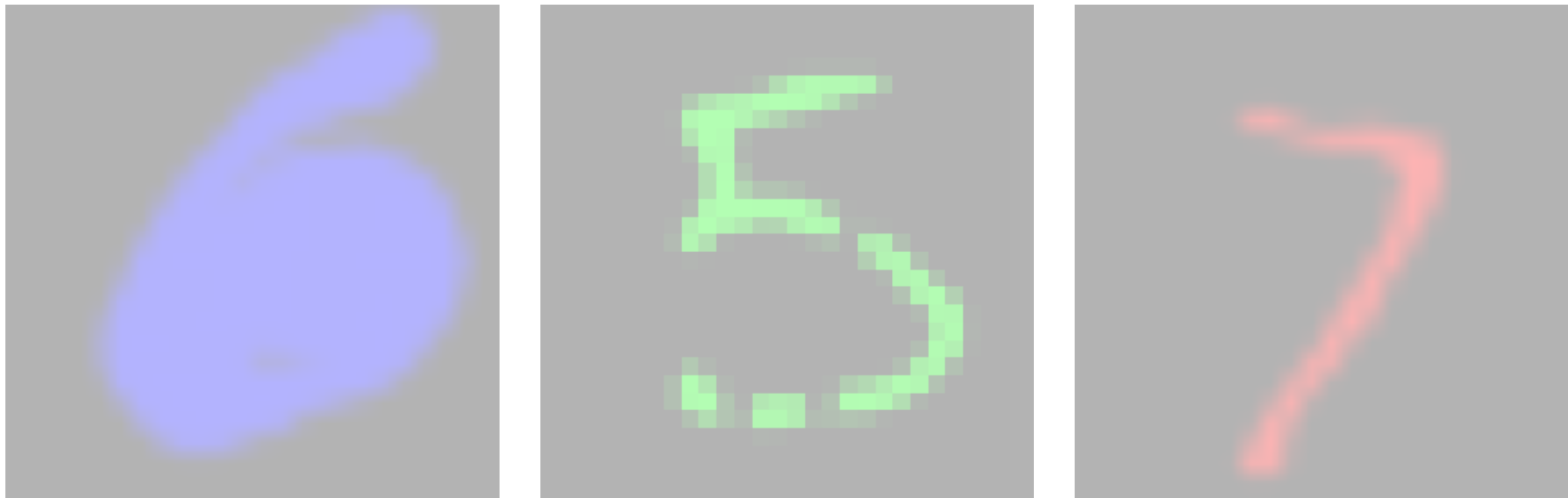
Natural Images



CLIP Models Trained on CC12M

CLIP ranks candidate captions by cosine similarity

Controlled Setup



Synthetic Digits with Known Attributes

Natural Images



CLIP Models Trained on CC12M

CLIP ranks candidate captions by cosine similarity

Query Image



Vision Encoder



Image Embedding



CLIP ranks candidate captions by cosine similarity

Query Image



Vision Encoder



Image Embedding



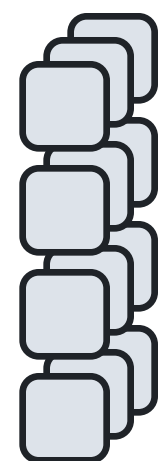
Candidate Captions



Text Encoder



Text Embeddings



CLIP ranks candidate captions by cosine similarity

Query Image



Vision Encoder

Image Embedding



Cosine Similarity

Top 10 by CLIP

A low-angle view of a yellow rhombus-shaped road sign with a black illustration of the silhouette of a tortoise walking toward the left side of the image attached to the top of a gray metal pole

⋮

An outdoor medium shot, looking up at a yellow, weathered, diamond-shaped caution sign with a black horse and buggy facing the right

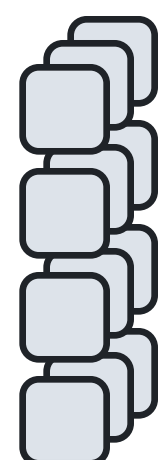
⋮

Candidate Captions

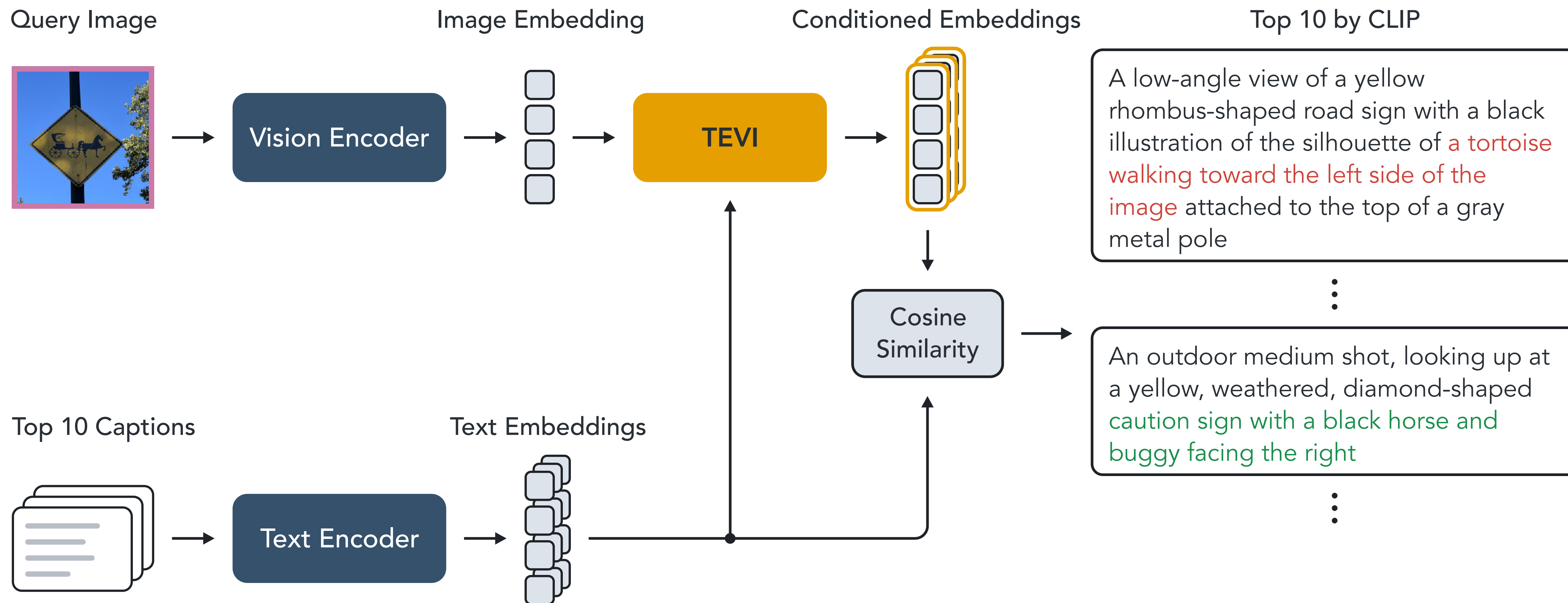


Text Encoder

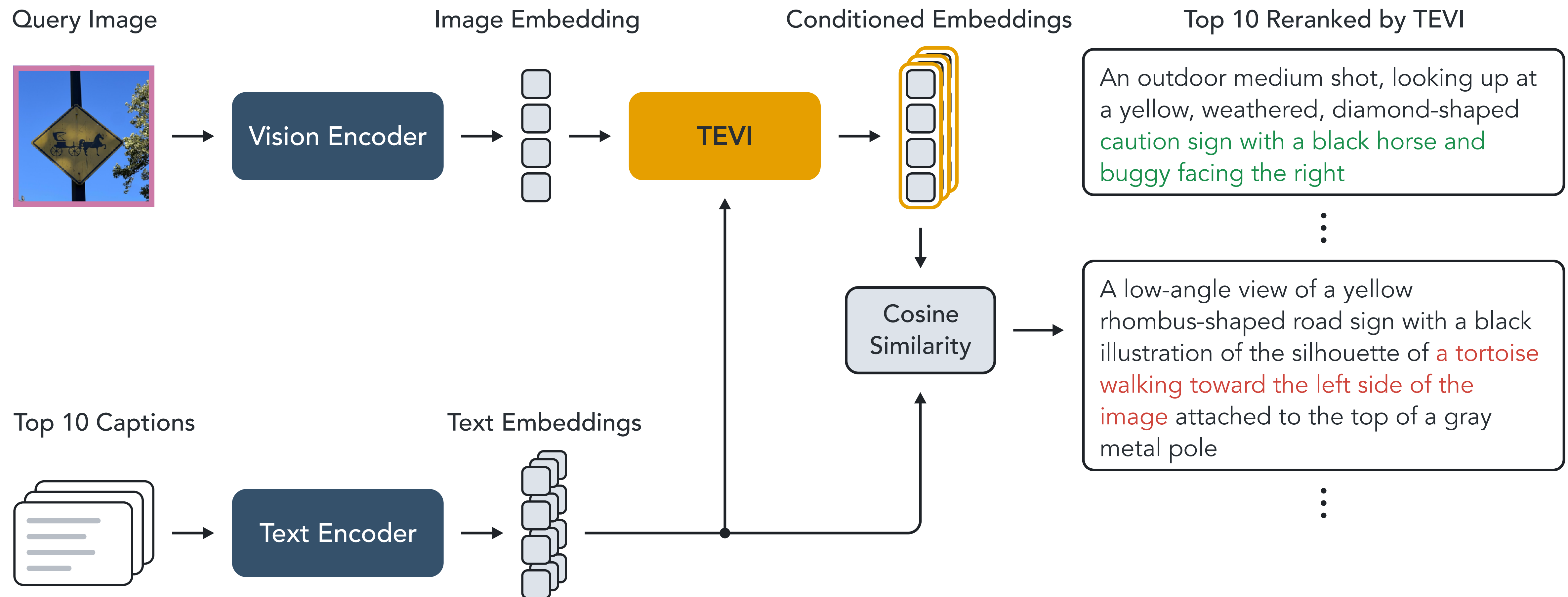
Text Embeddings



TEVI reranks the top-10 candidates of the base model



TEVI reranks the top-10 candidates of the base model



TEVI improves retrieval across models and datasets

Model	Image to Text, R@1				Text to Image, R@1			
	DOCCI	IIW	MS COCO	Flickr30k	DOCCI	IIW	MS COCO	Flickr30k
CLIP ViT-B/16	20.38	50.98	32.98	59.66	7.16	16.88	21.38	42.46
+TEVI	24.52 ± 0.15	55.49 ± 0.36	36.25 ± 0.23	64.22 ± 0.37	8.30 ± 0.04	18.77 ± 0.08	23.48 ± 0.04	44.79 ± 0.13

DOCCI: Onoe et al., ECCV 2024. IIW: Garg et al., EMNLP 2024. MS COCO: Lin et al., ECCV 2014. Flickr30k: Plummer et al., ICCV 2015.

TEVI improves retrieval across models and datasets

Model	Image to Text, R@1				Text to Image, R@1			
	DOCCI	IIW	MS COCO	Flickr30k	DOCCI	IIW	MS COCO	Flickr30k
CLIP ViT-B/16	20.38	50.98	32.98	59.66	7.16	16.88	21.38	42.46
+TEVI	24.52 ± 0.15	55.49 ± 0.36	36.25 ± 0.23	64.22 ± 0.37	8.30 ± 0.04	18.77 ± 0.08	23.48 ± 0.04	44.79 ± 0.13
No-SAE	24.29 ± 0.22	56.86 ± 0.31	36.26 ± 0.37	63.06 ± 0.40	8.40 ± 0.08	18.76 ± 0.15	23.59 ± 0.12	45.00 ± 0.14

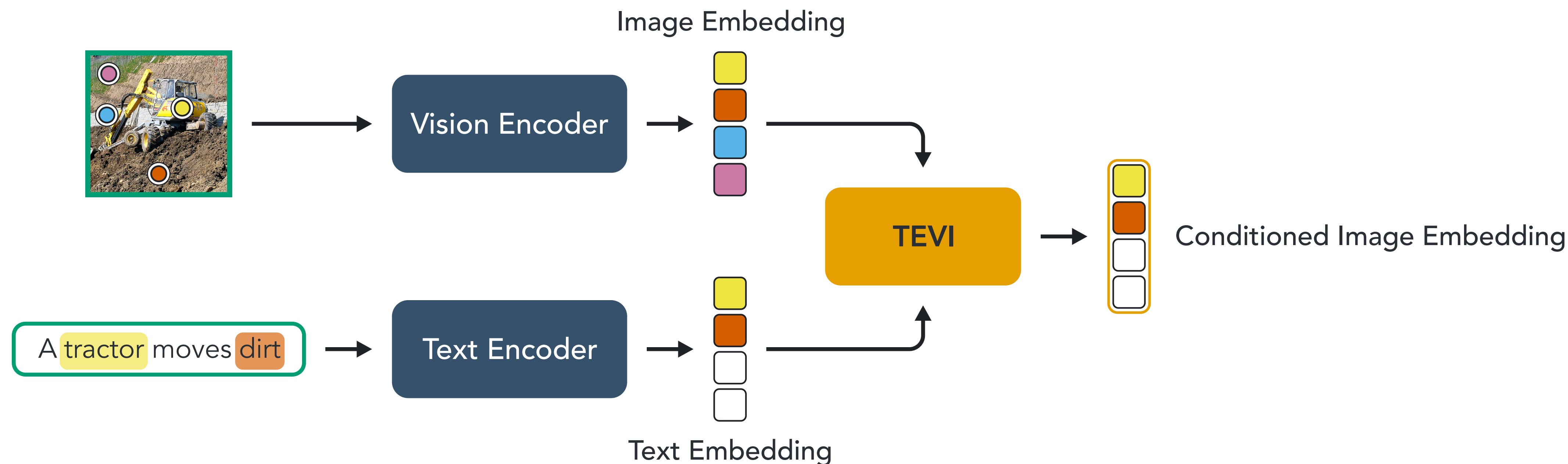
DOCCI: Onoe et al., ECCV 2024. IIW: Garg et al., EMNLP 2024. MS COCO: Lin et al., ECCV 2014. Flickr30k: Plummer et al., ICCV 2015.

TEVI improves retrieval across models and datasets

Model	Image to Text, R@1				Text to Image, R@1			
	DOCCI	IIW	MS COCO	Flickr30k	DOCCI	IIW	MS COCO	Flickr30k
CLIP ViT-B/16	20.38	50.98	32.98	59.66	7.16	16.88	21.38	42.46
+TEVI	24.52 ± 0.15	55.49 ± 0.36	36.25 ± 0.23	64.22 ± 0.37	8.30 ± 0.04	18.77 ± 0.08	23.48 ± 0.04	44.79 ± 0.13
No-SAE	24.29 ± 0.22	56.86 ± 0.31	36.26 ± 0.37	63.06 ± 0.40	8.40 ± 0.08	18.76 ± 0.15	23.59 ± 0.12	45.00 ± 0.14
CLIP ViT-L/14	23.60	53.92	36.52	63.71	8.21	18.05	23.85	46.51
+TEVI	26.84 ± 0.47	60.78 ± 0.60	38.41 ± 0.24	65.92 ± 0.43	9.09 ± 0.06	19.72 ± 0.05	24.98 ± 0.10	47.92 ± 0.10
SigLIP ViT-B/16	20.68	48.20	33.88	62.23	7.41	16.94	22.06	44.69
+TEVI	25.29 ± 0.22	57.84 ± 0.95	37.22 ± 0.34	64.48 ± 0.40	8.32 ± 0.06	18.90 ± 0.11	22.53 ± 0.06	43.96 ± 0.23
SharedCLIP ViT-B/16	19.02	51.14	32.62	60.75	7.16	17.10	21.54	43.31
+TEVI	22.64 ± 0.27	55.59 ± 0.31	34.80 ± 0.13	60.87 ± 0.61	7.94 ± 0.06	17.91 ± 0.18	22.15 ± 0.11	44.17 ± 0.16
AlignCLIP ViT-B/16	20.32	53.92	32.70	57.50	7.39	17.46	21.79	41.91
+TEVI	23.65 ± 0.26	58.92 ± 0.65	35.53 ± 0.13	61.56 ± 0.55	8.13 ± 0.08	19.75 ± 0.20	21.63 ± 0.16	42.85 ± 0.44

DOCCI: Onoe et al., ECCV 2024. IIW: Garg et al., EMNLP 2024. MS COCO: Lin et al., ECCV 2014. Flickr30k: Plummer et al., ICCV 2015.

TEVI makes the information imbalance controllable



- ✓ Keeps caption attributes, drops the rest
- ✓ Learns meaningful, sparse masks
- ✓ Improves retrieval
- ✗ Scaling to billion-scale data is a challenge
- ✗ Limited by how well the SAE disentangles
- ✗ Concepts are per image, not per region

Thank you!

Project Page



Poster Session

Date: Sunday, 25 October 2026

Time: 11:00–12:30

Location: Hall H

Session: 2 (Poster Session A)

Paper ID: 8041-FIND