Composed Image Retrieval for Training-FREE DOMain Conversion

Nikos Efthymiadis[1], Bill Psomas[1,2], Zakaria Laskar[1], Konstantinos Karantzalos[2] , Yannis Avrithis[3], Ondrej Chum[1], Giorgos Tolias[1]

[1]Visual Recognition Group, Czech Technical University in Prague
[2]National Technical University of Athens
[3] Institute of Advanced Research in Artificial Intelligence (IARAI)

Composed Image Retrieval

Visual Query

Is playing with a red ball

Textual Query

Composed Image Retrieval

Visual Query

Is playing with a red ball

Textual Query

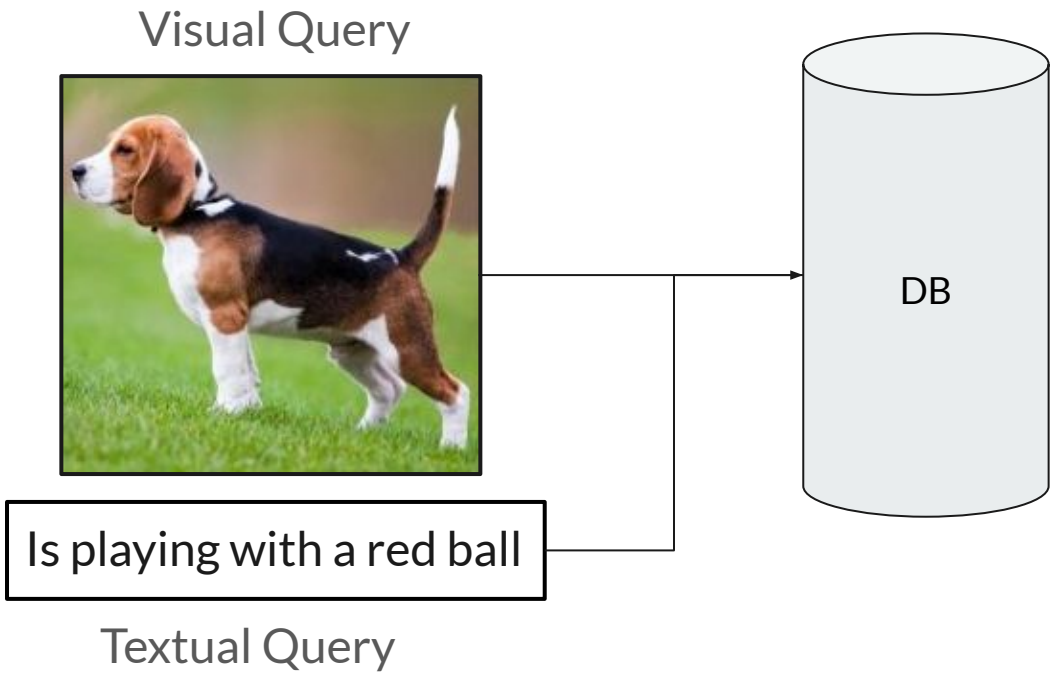
Composed Image Retrieval

Visual Query

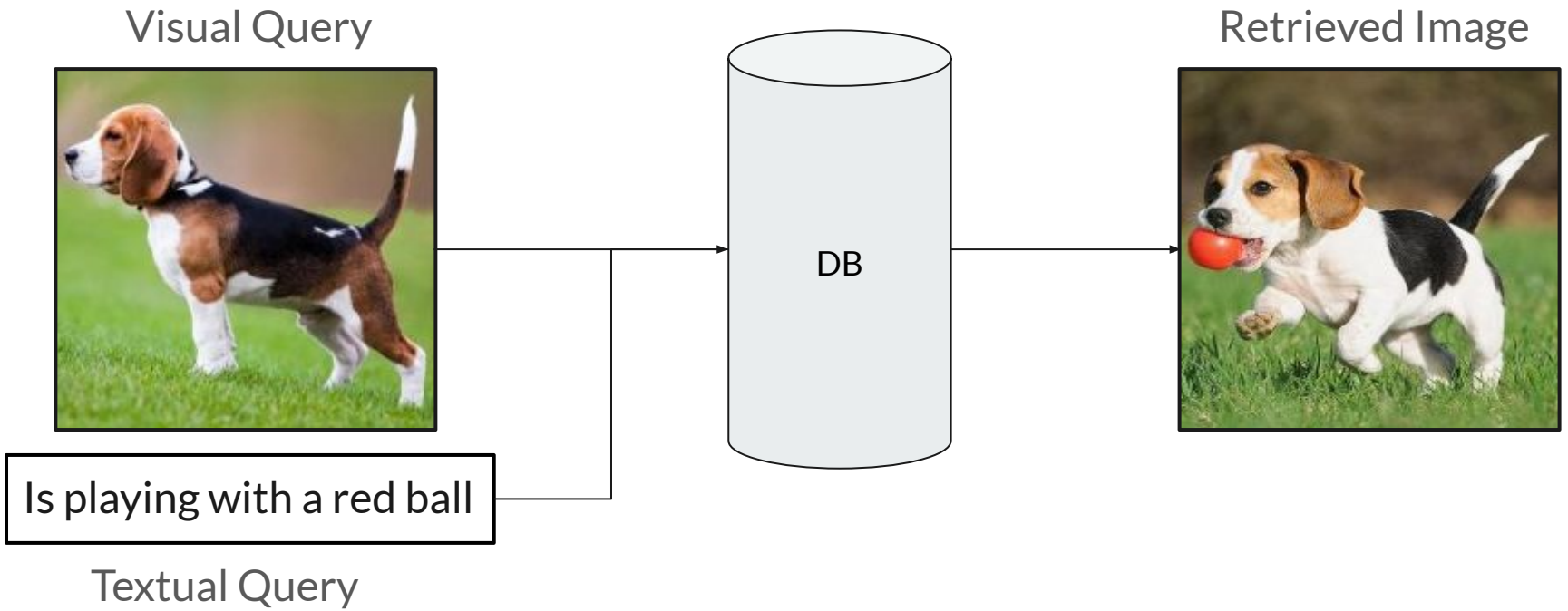
Is playing with a red ball

Textual Query

Composed Image Retrieval



Composed Image Retrieval



Composed Image Retrieval for Remote Sensing

Query Image

B. Psomas, I. Kakogeorgiou, N. Efthymiadis, G. Tolias, O. Chum, Y. Avrithis, K. Karantzalos, Composed Image Retrieval for Remote Sensing, IGARSS, 2024

Composed Image Retrieval for Remote Sensing

Query Image

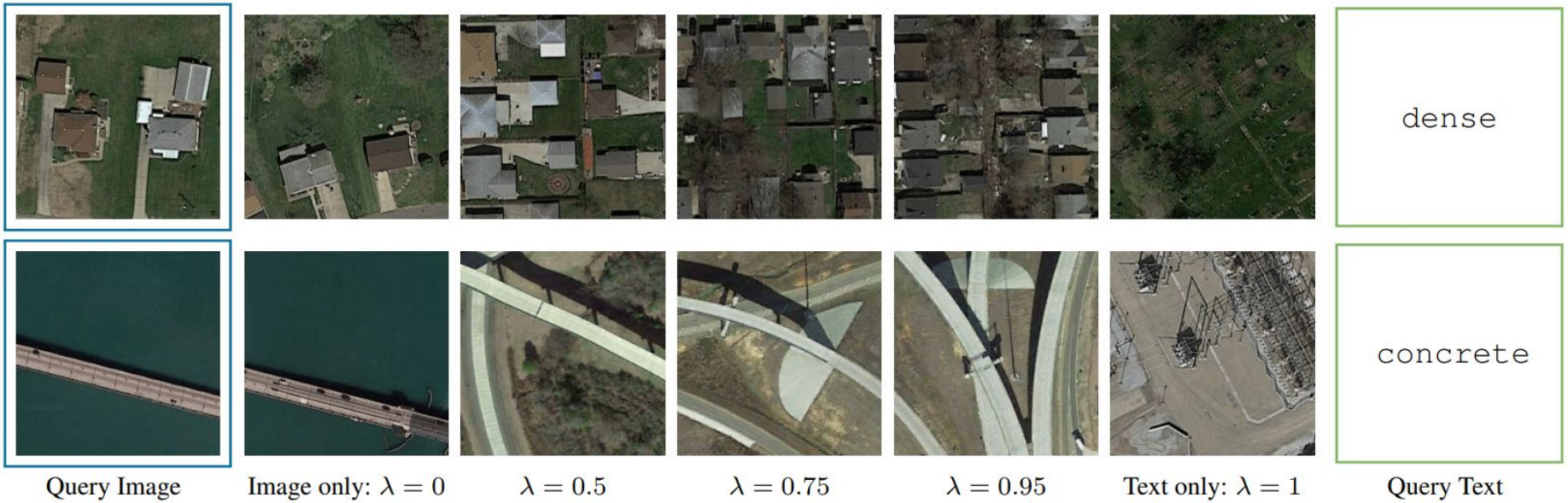
dense

concrete

Query Text

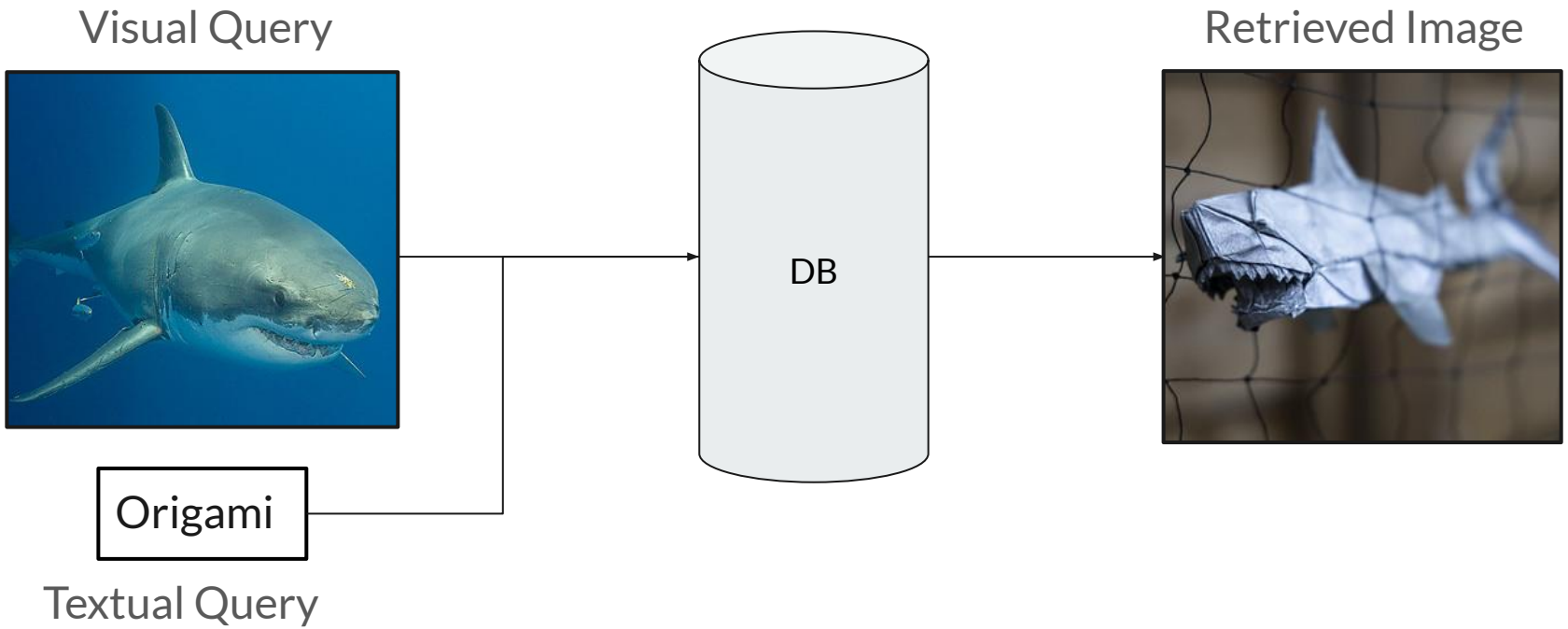
B. Psomas, I. Kakogeorgiou, N. Efthymiadis, G. Tolias, O. Chum, Y. Avrithis, K. Karantzalos, Composed Image Retrieval for Remote Sensing, IGARSS, 2024

Composed Image Retrieval for Remote Sensing



B. Psomas, I. Kakogeorgiou, N. Efthymiadis, G. Tolias, O. Chum, Y. Avrithis, K. Karantzalos, Composed Image Retrieval for Remote Sensing, IGARSS, 2024

Composed Image Retrieval for Domain Conversion



Contributions

This paper presents the following key contributions:

Contributions

This paper presents the following key contributions:

- Proposes a strong **testbed** for **domain conversion** that is composed of **four datasets**.

Contributions

This paper presents the following key contributions:

- Proposes a strong **testbed** for **domain conversion** that is composed of **four datasets**.
  - **Extends** ImageNet-R by using images from **every domain** as image queries

Contributions

This paper presents the following key contributions:

- Proposes a strong **testbed** for **domain conversion** that is composed of **four datasets**.
  - **Extends** ImageNet-R by using images from **every domain** as image queries
  - **Repurposes three** popular computer vision **datasets** for the domain conversion task

Contributions

This paper presents the following key contributions:

- Proposes a strong **testbed** for **domain conversion** that is composed of **four datasets**.
  - **Extends** ImageNet-R by using images from **every domain** as image queries
  - **Repurposes three** popular computer vision **datasets** for the domain conversion task
- Proposes **FreeDom**, a training-free domain conversion method that is the **new state-of-the-art** in all four domain conversion datasets.

Why do we need a strong testbed?

	ImageNet-R
METHOD	PHO
Text	0.68
Image	0.84
Text × Image	6.98
Text + Image	2.17
Pic2Word [1]	7.64
CompoDiff [2]	8.76
WeiCom [3]	10.06
SEARLE [4]	9.94
MagicLens [5]	11.02

[1] Kuniaki Saito et. al. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In CVPR, 2023
[2] Geonmo Gu et. al. Compodiff: Versatile composed image retrieval with latent diffusion. TMLR, 2024
[3] Bill Psomas et. al. Composed image retrieval for remote sensing. In IGARSS, 2024
[4] Alberto Baldrati et. al. Zero-shot composed image retrieval with textual inversion. In ICCV, 2023
[5] Kai Zhang et. al.. MagicLens: Self-supervised image retrieval with open-ended instructions. In ICML, 2024

Why do we need a strong testbed?

	ImageNet-R
METHOD	PHO
Text	0.68
Image	0.84
Text × Image	6.98
Text + Image	2.17
Pic2Word [1]	7.64
CompoDiff [2]	8.76
WeiCom [3]	10.06
SEARLE [4]	9.94
MagicLens [5]	11.02

[1] Kuniaki Saito et. al. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In CVPR, 2023
[2] Geonmo Gu et. al. Compodiff: Versatile composed image retrieval with latent diffusion. TMLR, 2024
[3] Bill Psomas et. al. Composed image retrieval for remote sensing. In IGARSS, 2024
[4] Alberto Baldrati et. al. Zero-shot composed image retrieval with textual inversion. In ICCV, 2023
[5] Kai Zhang et. al.. MagicLens: Self-supervised image retrieval with open-ended instructions. In ICML, 2024

Why do we need a strong testbed?

ImageNet-R

METHOD	CAR	ORI	PHO	SCU	TOY	AVG
Text	0.82	0.63	0.68	0.78	0.78	0.74
Image	4.27	3.12	0.84	5.86	5.08	3.84
Text × Image	8.21	5.62	6.98	8.95	9.41	7.83
Text + Image	6.61	4.45	2.17	9.18	8.62	6.21
Pic2Word [1]	7.60	5.53	7.64	9.39	9.27	7.88
CompoDiff [2]	13.71	10.61	8.76	15.17	16.17	12.88
WeiCom [3]	10.07	7.61	10.06	11.26	13.38	10.47
SEARLE [4]	18.11	9.02	9.94	17.26	15.83	14.04
MagicLens [5]	7.79	6.33	11.02	9.94	10.57	9.13

[1] Kuniaki Saito et. al. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In CVPR, 2023
[2] Geonmo Gu et. al. Compodiff: Versatile composed image retrieval with latent diffusion. TMLR, 2024
[3] Bill Psomas et. al. Composed image retrieval for remote sensing. In IGARSS, 2024
[4] Alberto Baldrati et. al. Zero-shot composed image retrieval with textual inversion. In ICCV, 2023
[5] Kai Zhang et. al.. MagicLens: Self-supervised image retrieval with open-ended instructions. In ICML, 2024

Why do we need a strong testbed?

(a) ImageNet-R

METHOD	CAR	ORI	PHO	SCU	TOY	AVG
Text	0.82	0.63	0.68	0.78	0.78	0.74
Image	4.27	3.12	0.84	5.86	5.08	3.84
Text × Image	8.21	5.62	6.98	8.95	9.41	7.83
Text + Image	6.61	4.45	2.17	9.18	8.62	6.21
Pic2Word [1]	7.60	5.53	7.64	9.39	9.27	7.88
CompoDiff [2]	13.71	10.61	8.76	15.17	16.17	12.88
WeiCom [3]	10.07	7.61	10.06	11.26	13.38	10.47
SEARLE [4]	18.11	9.02	9.94	17.26	15.83	14.04
MagicLens [5]	7.79	6.33	11.02	9.94	10.57	9.13

(b) MiniDomainNet

METHOD	CLIP	PAINT	PHO	SKE	AVG
Text	0.63	0.52	0.63	0.51	0.57
Image	7.15	7.31	4.38	7.78	6.66
Text × Image	9.01	8.66	15.87	5.90	9.86
Text + Image	9.59	9.97	9.22	8.53	9.33
Pic2Word	13.39	8.63	17.96	8.03	12.00
CompoDiff	19.06	24.27	23.41	25.05	22.95
WeiCom	7.52	7.04	15.13	4.40	8.52
SEARLE	25.04	18.72	23.75	19.61	21.78
MagicLens	24.40	17.54	28.59	9.71	20.06

[1] Kuniaki Saito et. al. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In CVPR, 2023
[2] Geonmo Gu et. al. Compodiff: Versatile composed image retrieval with latent diffusion. TMLR, 2024
[3] Bill Psomas et. al. Composed image retrieval for remote sensing. In IGARSS, 2024
[4] Alberto Baldrati et. al. Zero-shot composed image retrieval with textual inversion. In ICCV, 2023
[5] Kai Zhang et. al.. MagicLens: Self-supervised image retrieval with open-ended instructions. In ICML, 2024

Why do we need a strong testbed?

(a) ImageNet-R

METHOD	CAR	ORI	PHO	SCU	TOY	AVG
Text	0.82	0.63	0.68	0.78	0.78	0.74
Image	4.27	3.12	0.84	5.86	5.08	3.84
Text × Image	8.21	5.62	6.98	8.95	9.41	7.83
Text + Image	6.61	4.45	2.17	9.18	8.62	6.21
Pic2Word	7.60	5.53	7.64	9.39	9.27	7.88
CompoDiff	13.71	10.61	8.76	15.17	16.17	12.88
WeiCom	10.07	7.61	10.06	11.26	13.38	10.47
SEARLE	18.11	9.02	9.94	17.26	15.83	14.04
MagicLens	7.79	6.33	11.02	9.94	10.57	9.13

(b) MiniDomainNet

METHOD	CLIP	PAINT	PHO	SKE	AVG
Text	0.63	0.52	0.63	0.51	0.57
Image	7.15	7.31	4.38	7.78	6.66
Text × Image	9.01	8.66	15.87	5.90	9.86
Text + Image	9.59	9.97	9.22	8.53	9.33
Pic2Word	13.39	8.63	17.96	8.03	12.00
CompoDiff	19.06	24.27	23.41	25.05	22.95
WeiCom	7.52	7.04	15.13	4.40	8.52
SEARLE	25.04	18.72	23.75	19.61	21.78
MagicLens	24.40	17.54	28.59	9.71	20.06

(c) NICO++

METHOD	AUT	DIM	GRA	OUT	ROC	WAT	AVG
Text	1.00	0.99	1.15	1.23	1.10	1.05	1.09
Image	6.45	4.85	5.67	7.67	7.65	5.65	6.32
Text × Image	8.24	6.36	12.11	12.71	10.46	8.84	9.79
Text + Image	8.46	6.58	9.22	11.91	11.20	8.41	9.30
Pic2Word	9.79	8.09	11.24	11.27	11.01	7.16	9.76
CompoDiff	10.07	7.83	10.53	11.41	11.93	10.15	10.32
WeiCom	8.58	7.39	13.04	13.17	11.32	9.73	10.54
SEARLE	13.49	13.73	17.91	17.99	15.79	11.84	15.13
MagicLens	18.76	15.17	22.14	23.61	21.99	16.30	19.66

(d) LTLL

METHOD	TODAY	ARCHIVE	AVG
Text	5.28	6.16	5.72
Image	8.47	24.51	16.49
Text × Image	16.42	29.90	23.16
Text + Image	9.60	26.13	17.86
Pic2Word	17.86	24.67	21.27
CompoDiff	15.45	27.76	21.61
WeiCom	24.56	28.63	26.60
SEARLE	20.82	30.10	25.46
MagicLens	33.77	14.65	24.21

Why do we need a strong testbed?

(a) ImageNet-R

METHOD	CAR	ORI	PHO	SCU	TOY	AVG
Text	0.82	0.63	0.68	0.78	0.78	0.74
Image	4.27	3.12	0.84	5.86	5.08	3.84
Text × Image	8.21	5.62	6.98	8.95	9.41	7.83
Text + Image	6.61	4.45	2.17	9.18	8.62	6.21
Pic2Word	7.60	5.53	7.64	9.39	9.27	7.88
CompoDiff	13.71	10.61	8.76	15.17	16.17	12.88
WeiCom	10.07	7.61	10.06	11.26	13.38	10.47
SEARLE	18.11	9.02	9.94	17.26	15.83	14.04
MagicLens	7.79	6.33	11.02	9.94	10.57	9.13

(b) MiniDomainNet

METHOD	CLIP	PAINT	PHO	SKE	AVG
Text	0.63	0.52	0.63	0.51	0.57
Image	7.15	7.31	4.38	7.78	6.66
Text × Image	9.01	8.66	15.87	5.90	9.86
Text + Image	9.59	9.97	9.22	8.53	9.33
Pic2Word	13.39	8.63	17.96	8.03	12.06
CompoDiff	19.06	24.27	23.41	25.05	22.95
WeiCom	7.52	7.04	15.13	4.40	8.52
SEARLE	25.04	18.72	23.75	19.61	21.78
MagicLens	24.40	17.54	28.59	9.71	20.06

(c) NICO++

METHOD	AUT	DIM	GRA	OUT	ROC	WAT	AVG
Text	1.00	0.99	1.15	1.23	1.10	1.05	1.09
Image	6.45	4.85	5.67	7.67	7.65	5.65	6.32
Text × Image	8.24	6.36	12.11	12.71	10.46	8.84	9.79
Text + Image	8.46	6.58	9.22	11.91	11.20	8.41	9.30
Pic2Word	9.79	8.09	11.24	11.27	11.01	7.16	9.76
CompoDiff	10.07	7.83	10.53	11.41	11.93	10.15	10.32
WeiCom	8.58	7.39	13.04	13.17	11.32	9.73	10.54
SEARLE	13.49	13.73	17.91	17.99	15.79	11.84	15.13
MagicLens	18.76	15.17	22.14	23.61	21.99	16.30	19.66

(d) LTLL

METHOD	TODAY	ARCHIVE	AVG
Text	5.28	6.16	5.72
Image	8.47	24.51	16.49
Text × Image	16.42	29.90	23.16
Text + Image	9.60	26.13	17.86
Pic2Word	17.86	24.67	21.27
CompoDiff	15.45	27.76	21.61
WeiCom	24.56	28.63	26.60
SEARLE	20.82	30.10	25.46
MagicLens	33.77	14.65	24.21

Question: Which method is the state-of-the-art?

Why do we need a strong testbed?

(a) ImageNet-R

METHOD	CAR	ORI	PHO	SCU	TOY	AVG
Text	0.82	0.63	0.68	0.78	0.78	0.74
Image	4.27	3.12	0.84	5.86	5.08	3.84
Text × Image	8.21	5.62	6.98	8.95	9.41	7.83
Text + Image	6.61	4.45	2.17	9.18	8.62	6.21
Pic2Word	7.60	5.53	7.64	9.39	9.27	7.88
CompoDiff	13.71	10.61	8.76	15.17	16.17	12.88
WeiCom	10.07	7.61	10.06	11.26	13.38	10.47
SEARLE	18.11	9.02	9.94	17.26	15.83	14.04
MagicLens	7.79	6.33	11.02	9.94	10.57	9.13

(b) MiniDomainNet

METHOD	CLIP	PAINT	PHO	SKE	AVG
Text	0.63	0.52	0.63	0.51	0.57
Image	7.15	7.31	4.38	7.78	6.66
Text × Image	9.01	8.66	15.87	5.90	9.86
Text + Image	9.59	9.97	9.22	8.53	9.33
Pic2Word	13.39	8.63	17.96	8.03	12.06
CompoDiff	19.06	24.27	23.41	25.05	22.95
WeiCom	7.52	7.04	15.13	4.40	8.52
SEARLE	25.04	18.72	23.75	19.61	21.78
MagicLens	24.40	17.54	28.59	9.71	20.06

(c) NICO++

METHOD	AUT	DIM	GRA	OUT	ROC	WAT	AVG
Text	1.00	0.99	1.15	1.23	1.10	1.05	1.09
Image	6.45	4.85	5.67	7.67	7.65	5.65	6.32
Text × Image	8.24	6.36	12.11	12.71	10.46	8.84	9.79
Text + Image	8.46	6.58	9.22	11.91	11.20	8.41	9.30
Pic2Word	9.79	8.09	11.24	11.27	11.01	7.16	9.76
CompoDiff	10.07	7.83	10.53	11.41	11.93	10.15	10.32
WeiCom	8.58	7.39	13.04	13.17	11.32	9.73	10.54
SEARLE	13.49	13.73	17.91	17.99	15.79	11.84	15.13
MagicLens	18.76	15.17	22.14	23.61	21.99	16.30	19.66

(d) LTLL

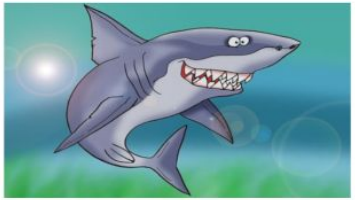
METHOD	TODAY	ARCHIVE	AVG
Text	5.28	6.16	5.72
Image	8.47	24.51	16.49
Text × Image	16.42	29.90	23.16
Text + Image	9.60	26.13	17.86
Pic2Word	17.86	24.67	21.27
CompoDiff	15.45	27.76	21.61
WeiCom	24.56	28.63	26.60
SEARLE	20.82	30.10	25.46
MagicLens	33.77	14.65	24.21

A **weak testbed** leads the community to propose literature **methods** that are **not** overall **strong**

Composed Image Retrieval for Domain Conversion

(a)
image: category
domain: style

image query



text query: “cartoon”

text query: “sculpture”

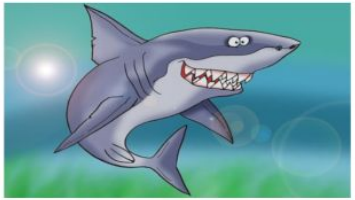
text query: “toy”

ImageNet-R and **Mini-DomainNet**

Composed Image Retrieval for Domain Conversion

(a)
image: category
domain: style

image query



text query: “cartoon”

text query: “sculpture”

text query: “toy”

(b)
image: category
domain: context

image query

text query: “autumn”

text query: “grass”

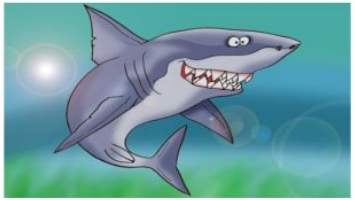
text query: “water”

NICO++

Composed Image Retrieval for Domain Conversion

(a)
image: category
domain: style

image query



text query: “cartoon”

text query: “sculpture”

text query: “toy”

(b)
image: category
domain: context

image query

text query: “autumn”

text query: “grass”

text query: “water”

(c)
image: instance
domain: style

image query

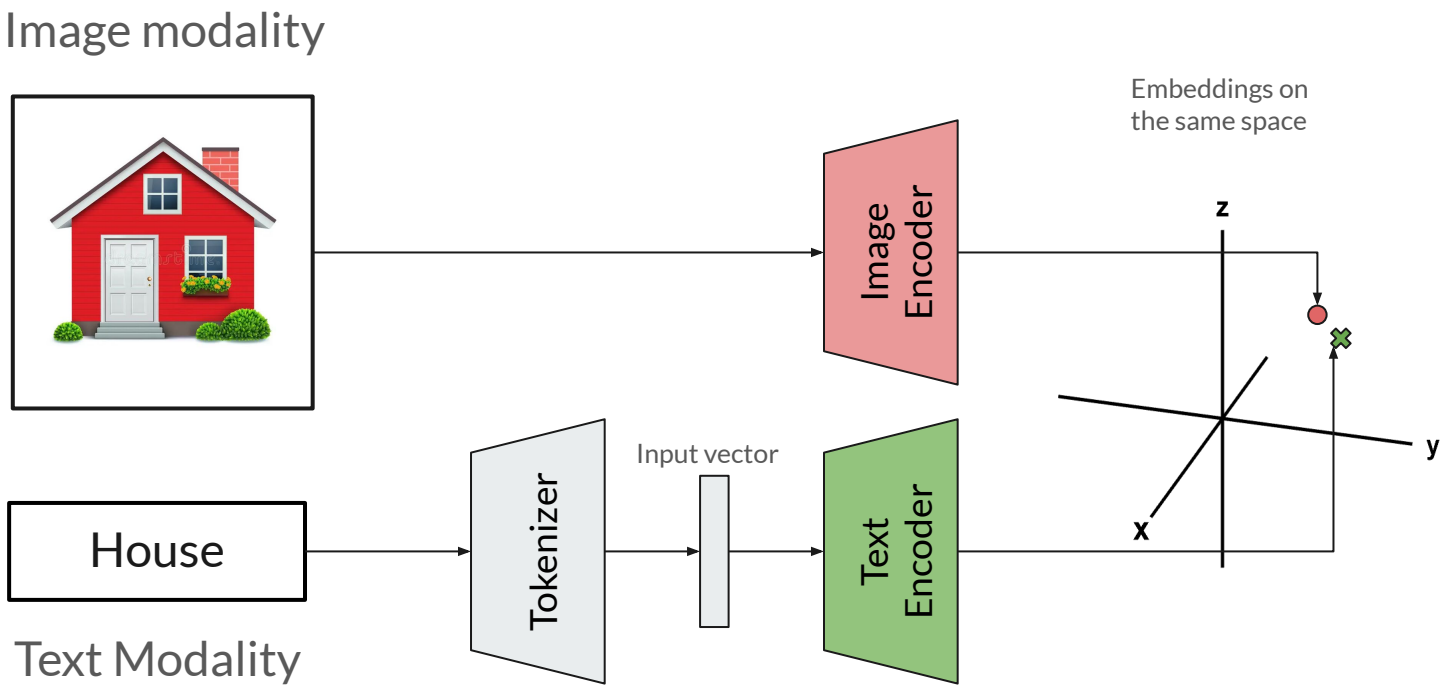
text query: “archive”

image query

text query: “today”

Large Time Lags Location (LTLL)

Visual-Language Models



Visual-Language Models and Combining Modalities

- Combining Embeddings
- Combination as Image
- Combination as Text

Combining Embeddings

Visual-Language Models (Combining Embeddings)

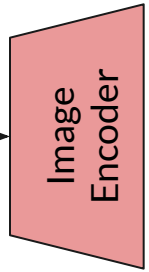
Visual Query

Origami

Textual Query

Visual-Language Models (Combining Embeddings)

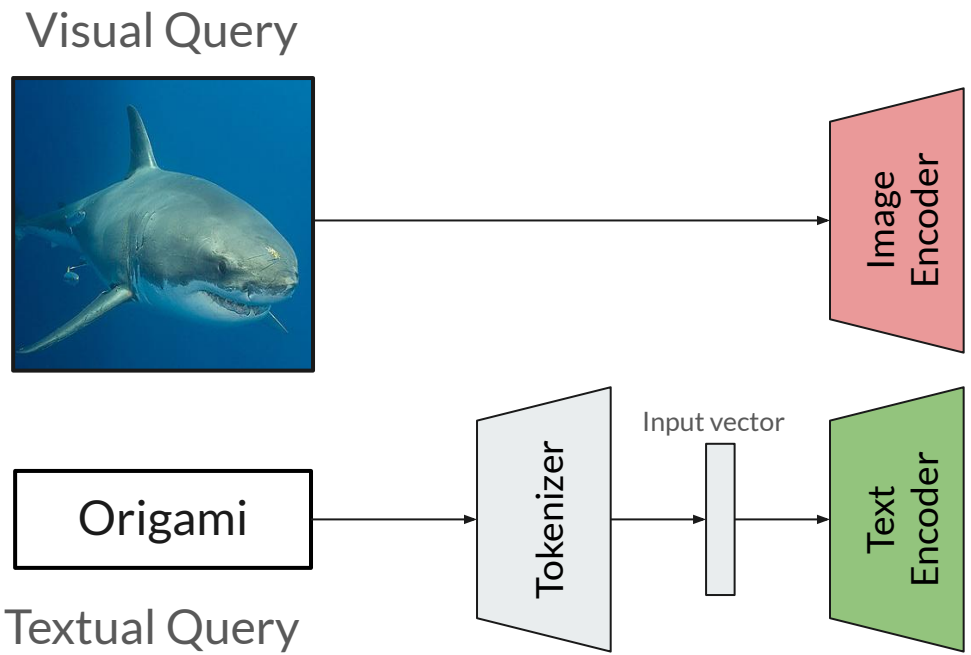
Visual Query



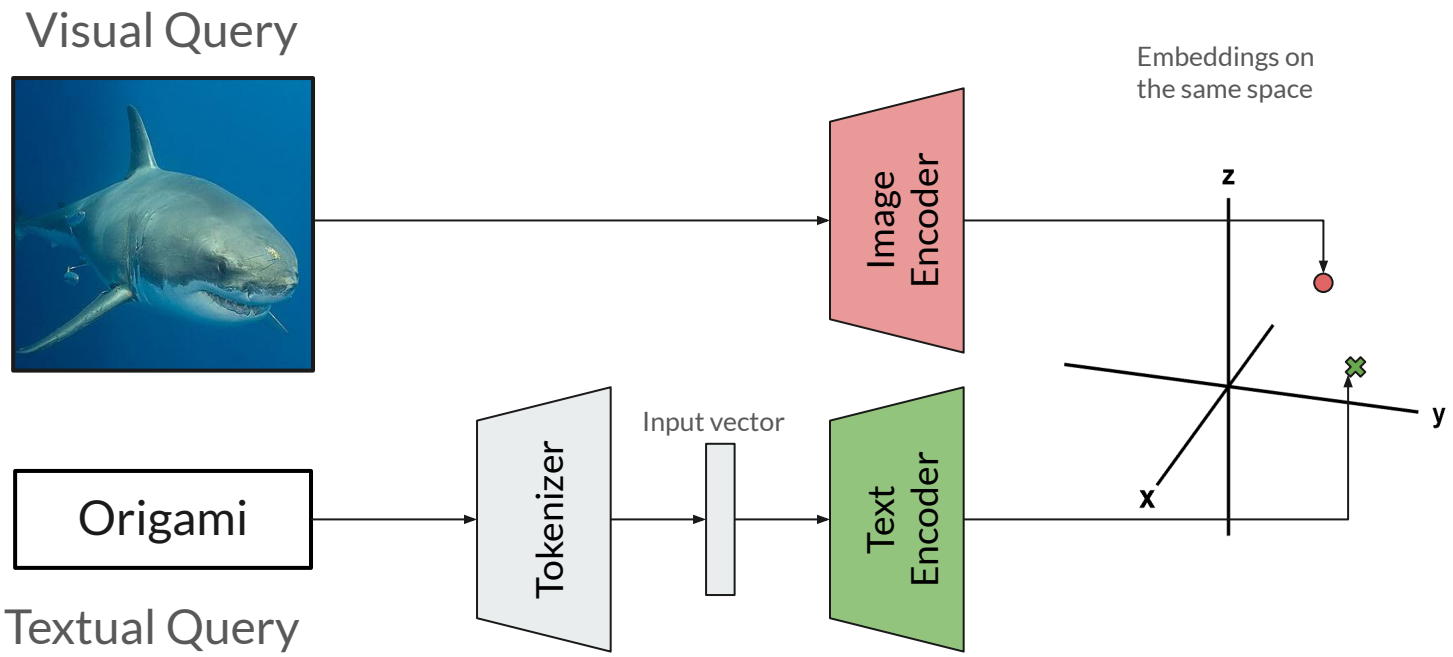
Origami

Textual Query

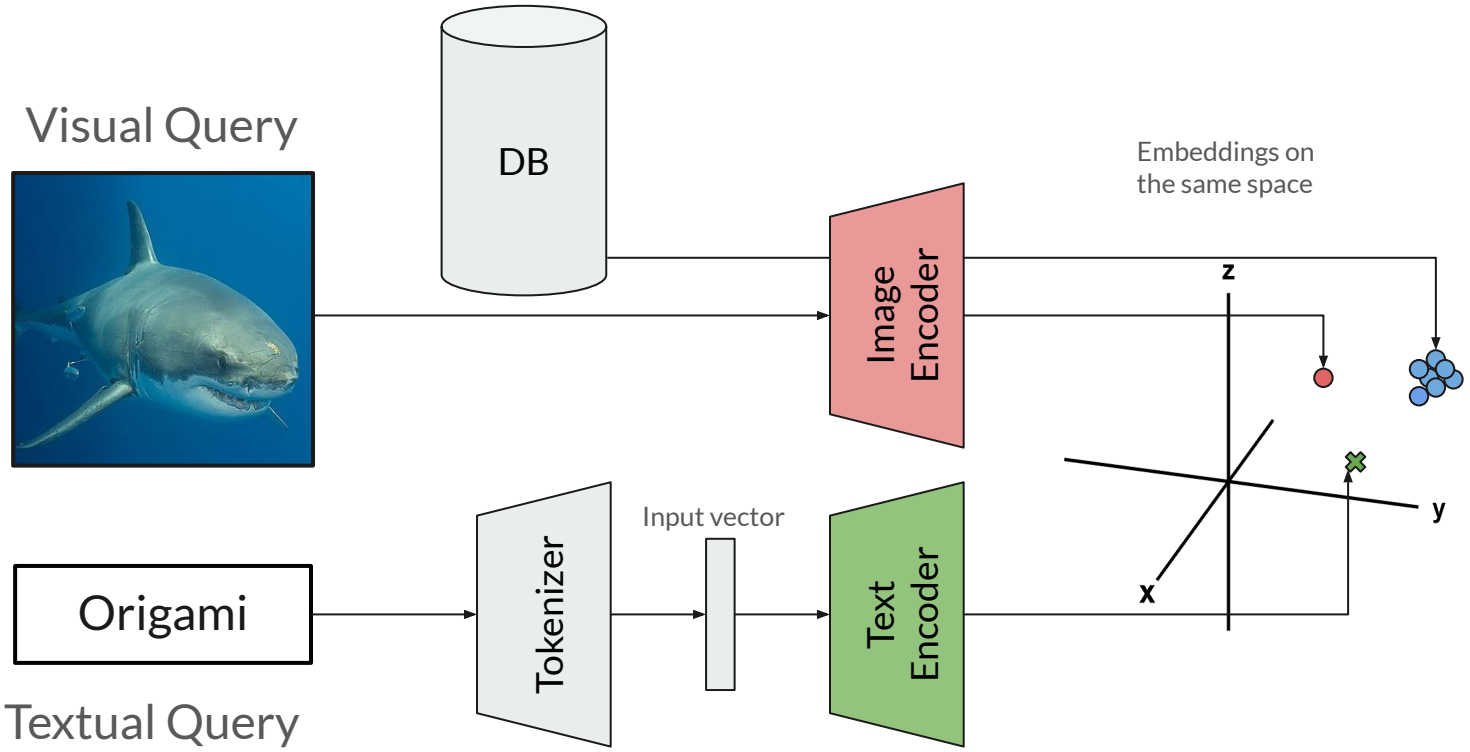
Visual-Language Models (Combining Embeddings)



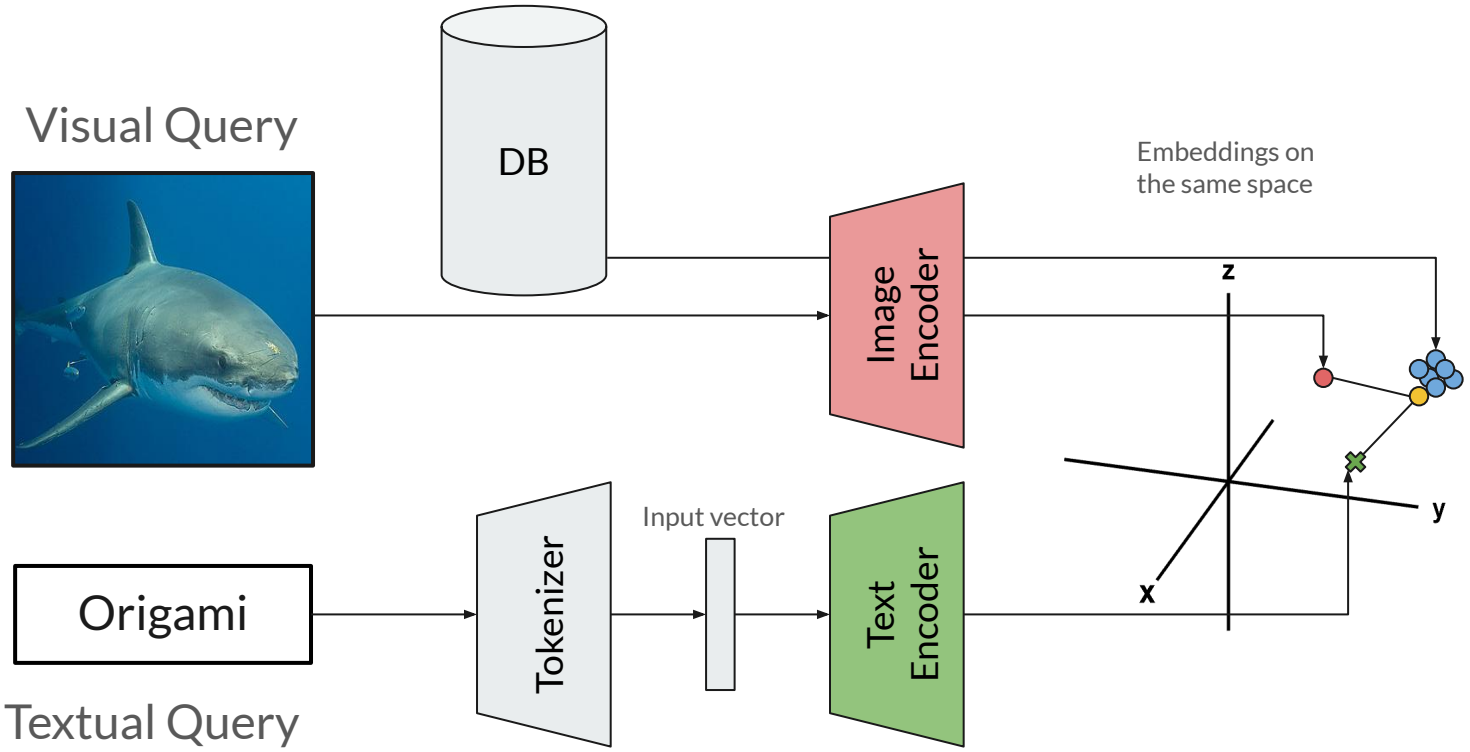
Visual-Language Models (Combining Embeddings)



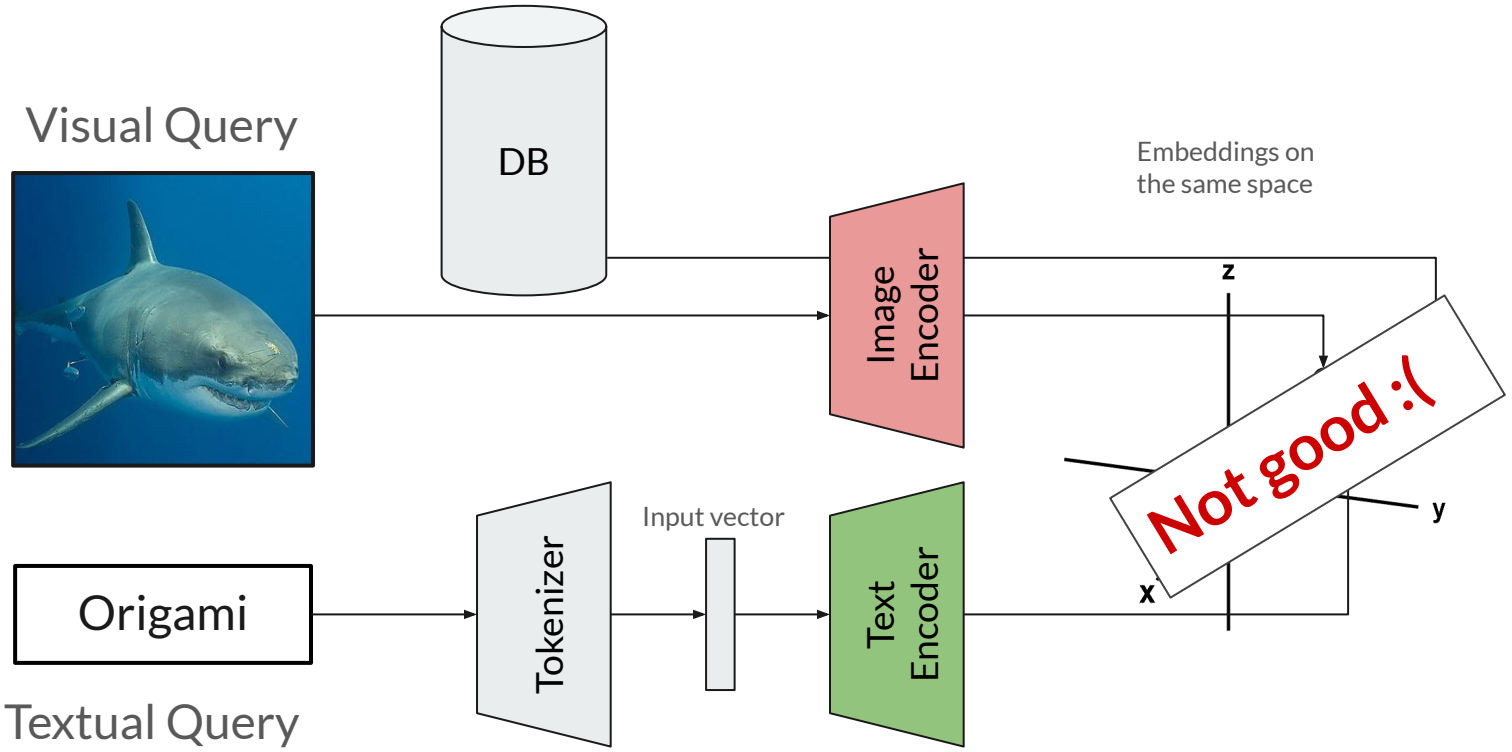
Visual-Language Models (Combining Embeddings)



Visual-Language Models (Combining Embeddings)



Visual-Language Models (Combining Embeddings)



Visual-Language Models and Combining Modalities

Combination as Image

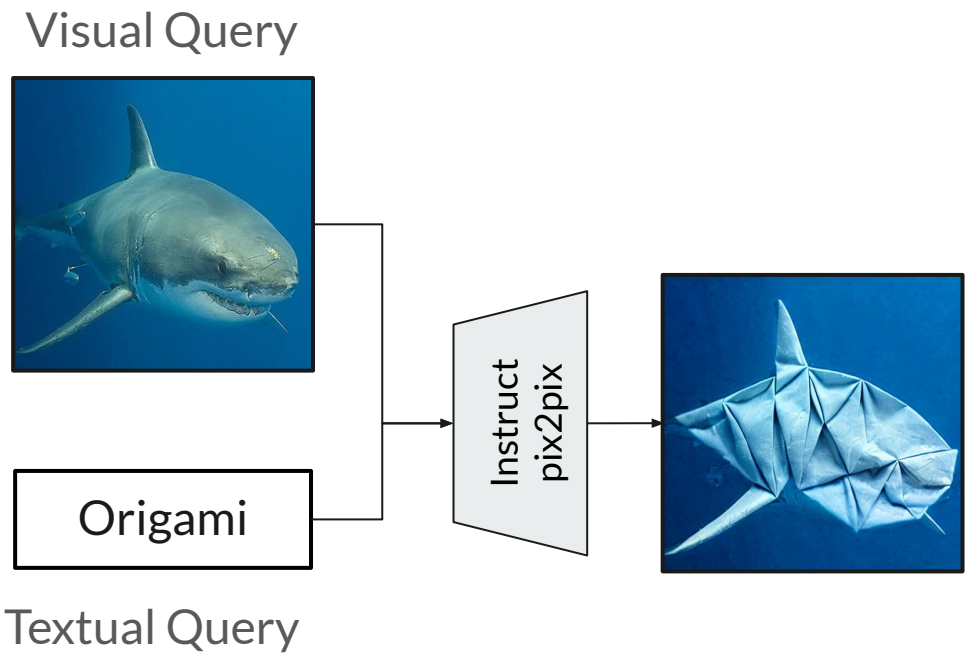
Language Guided Image Edit (Combination as Image)

Visual Query

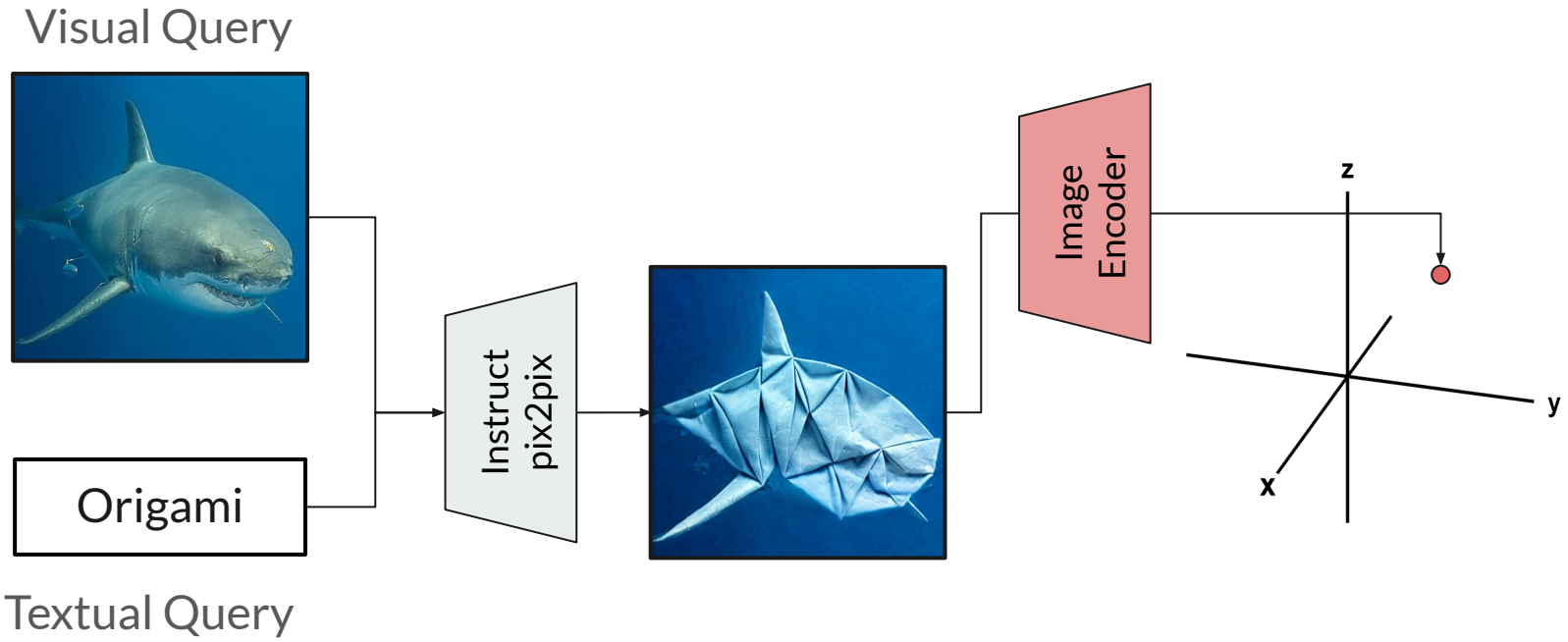
Origami

Textual Query

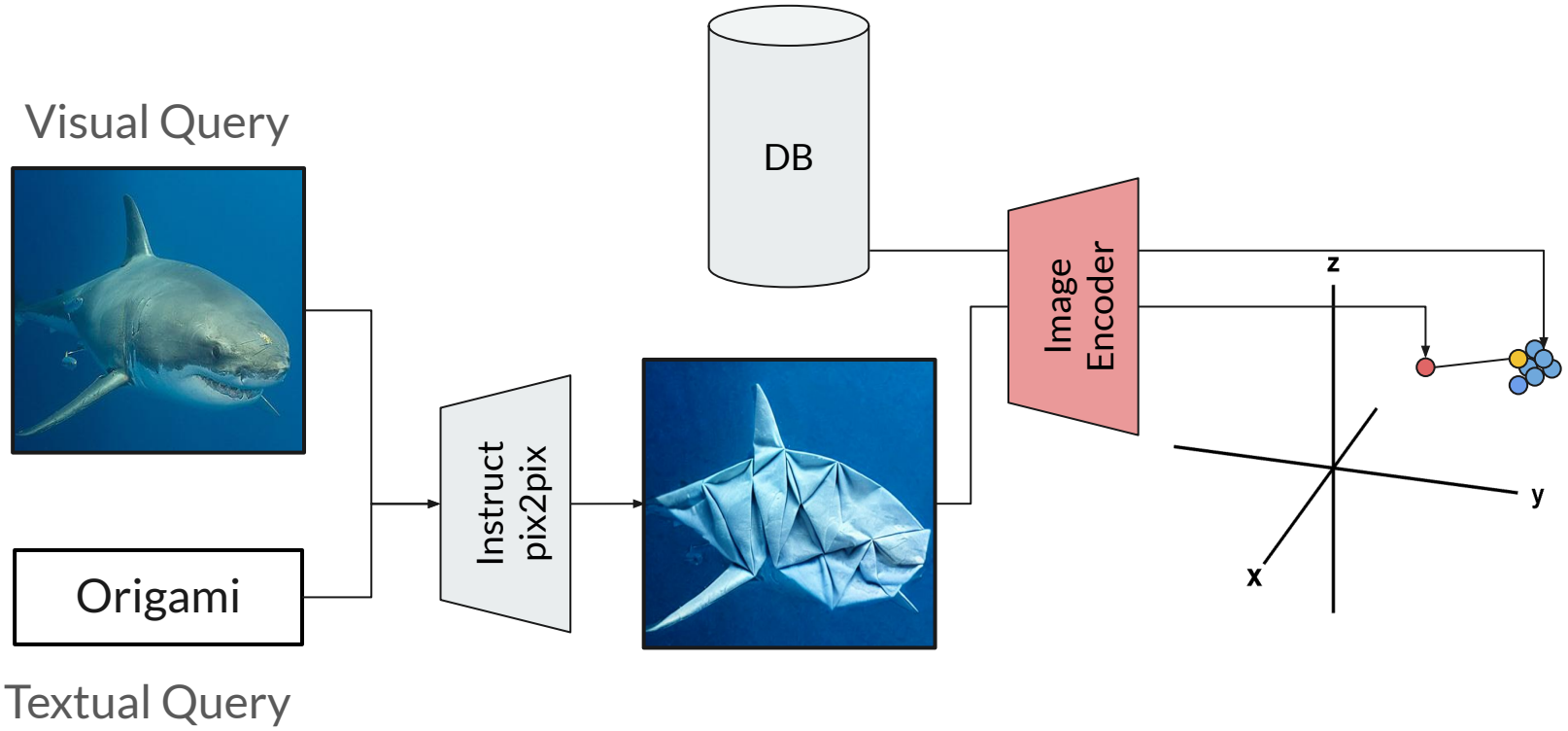
Language Guided Image Edit (Combination as Image)



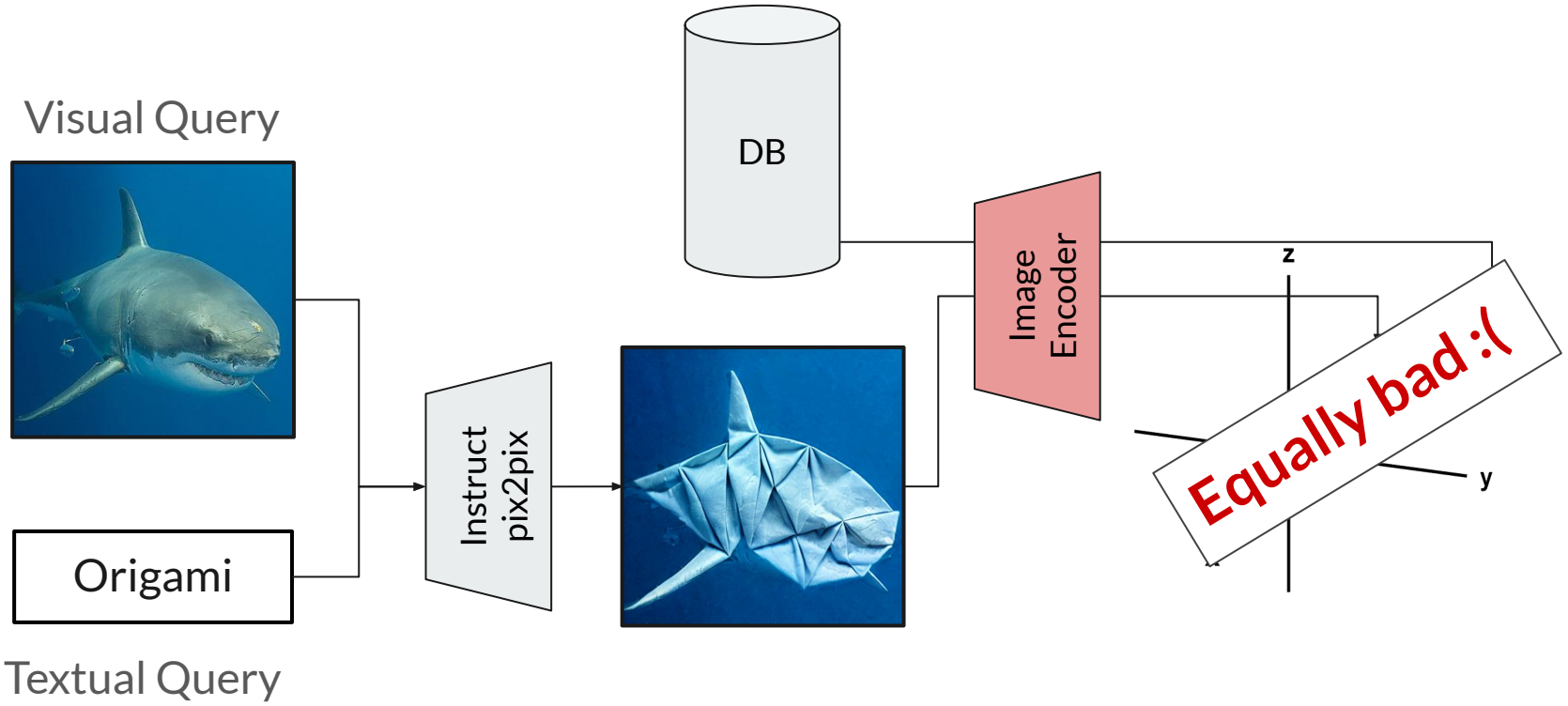
Language Guided Image Edit (Combination as Image)



Language Guided Image Edit (Combination as Image)



Language Guided Image Edit (Combination as Image)

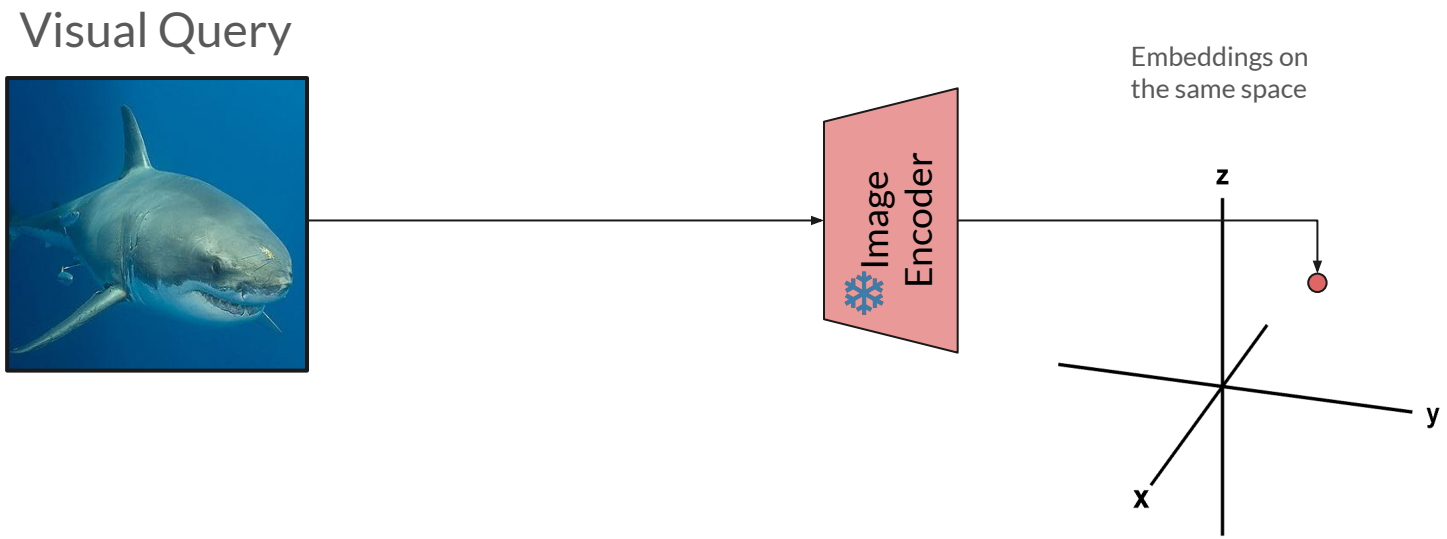


Combination as Text

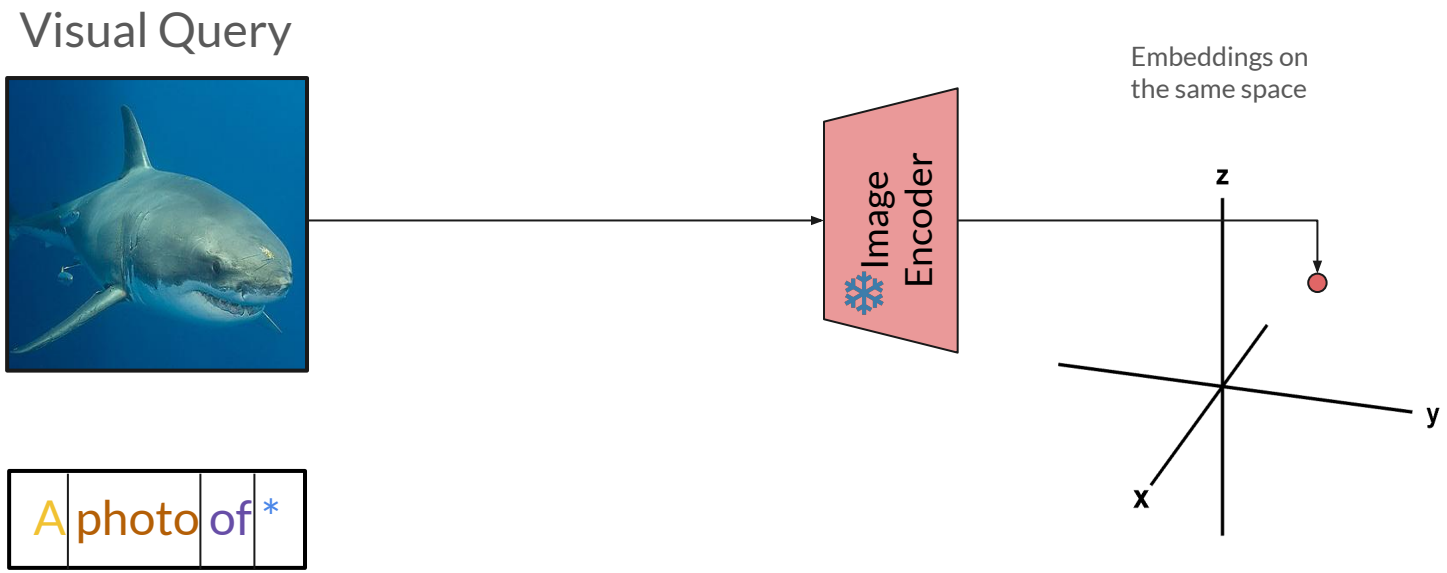
Textual Inversion in Continuous Token Space

Visual Query

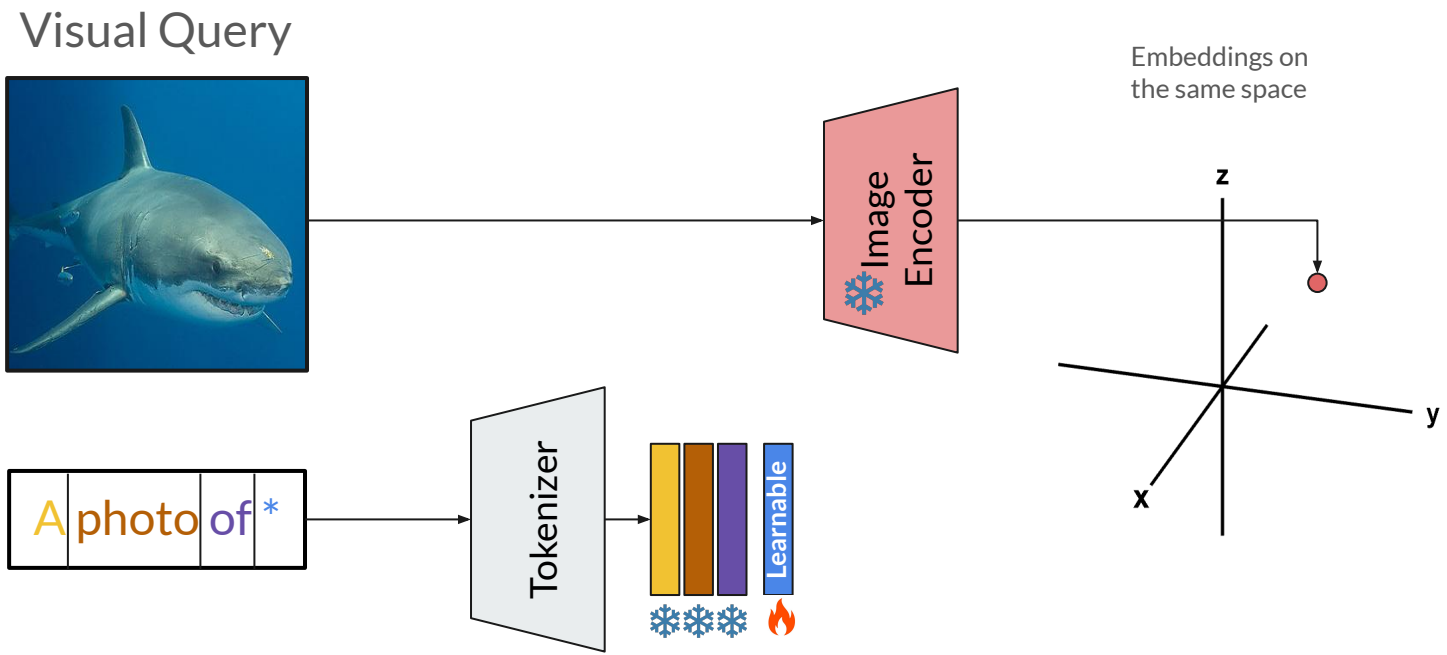
Textual Inversion in Continuous Token Space



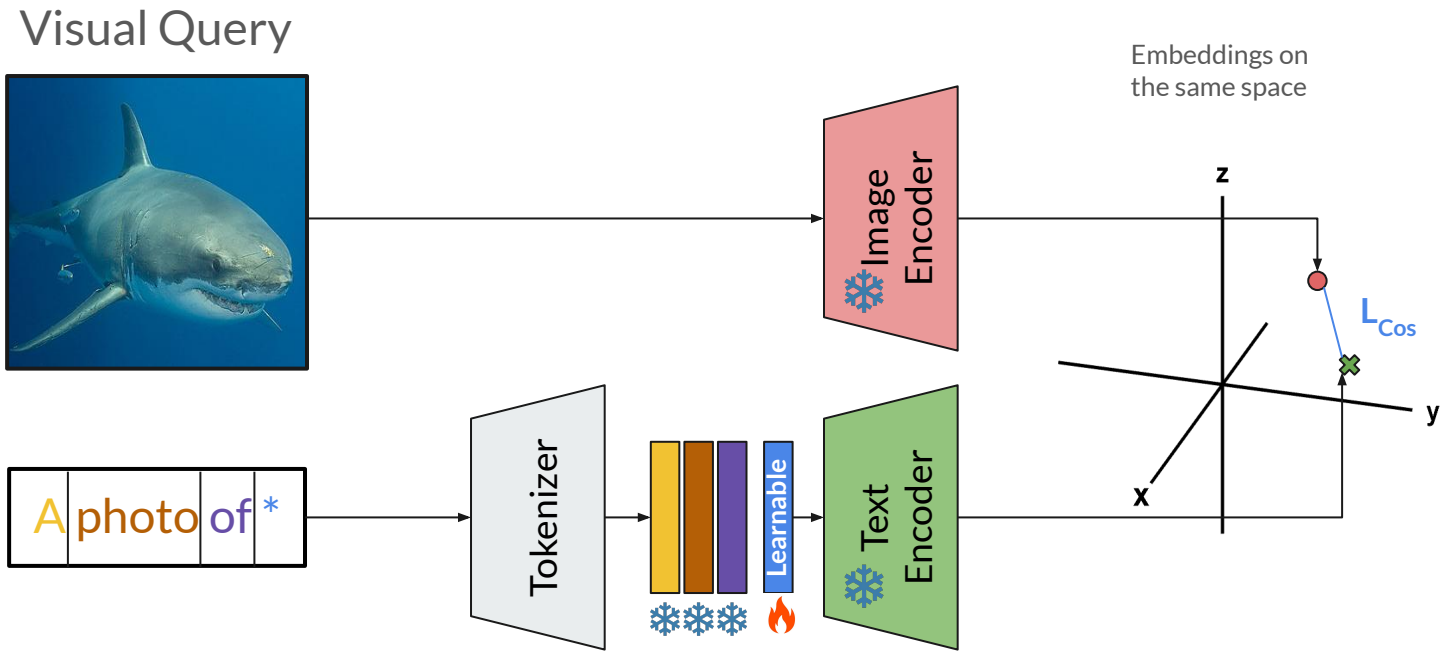
Textual Inversion in Continuous Token Space



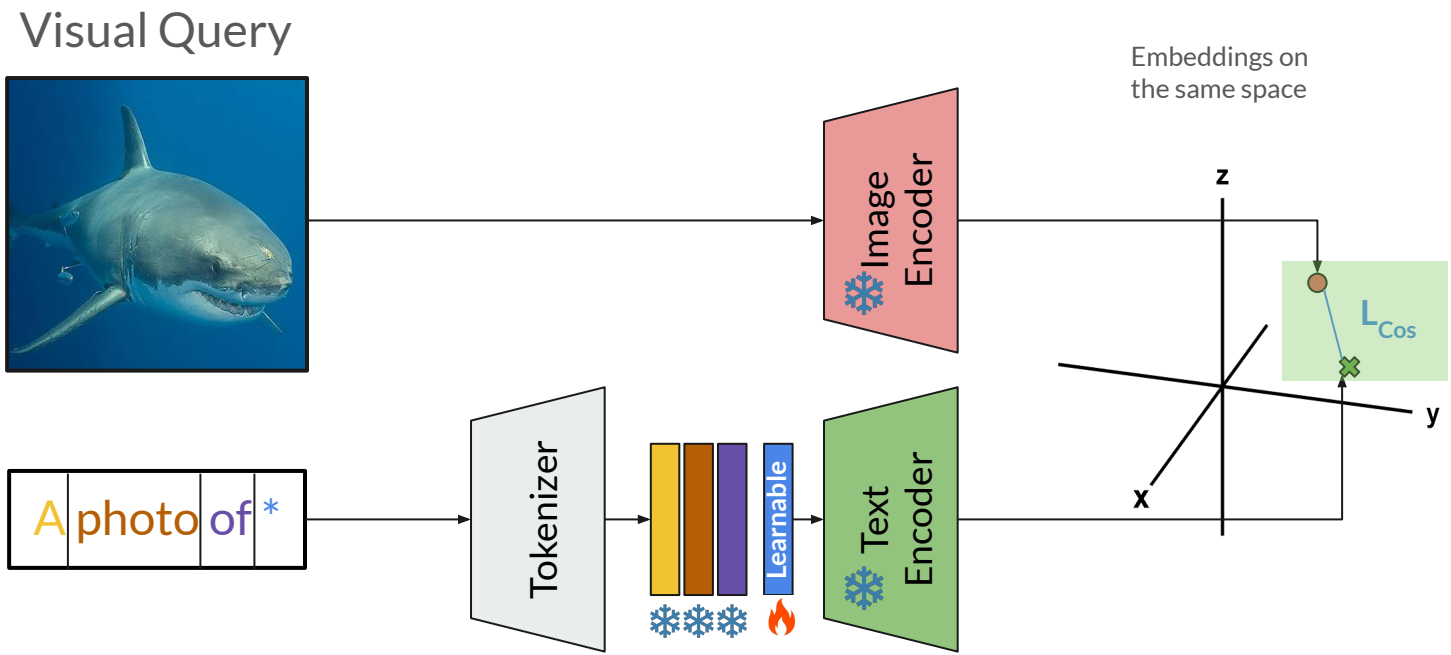
Textual Inversion in Continuous Token Space



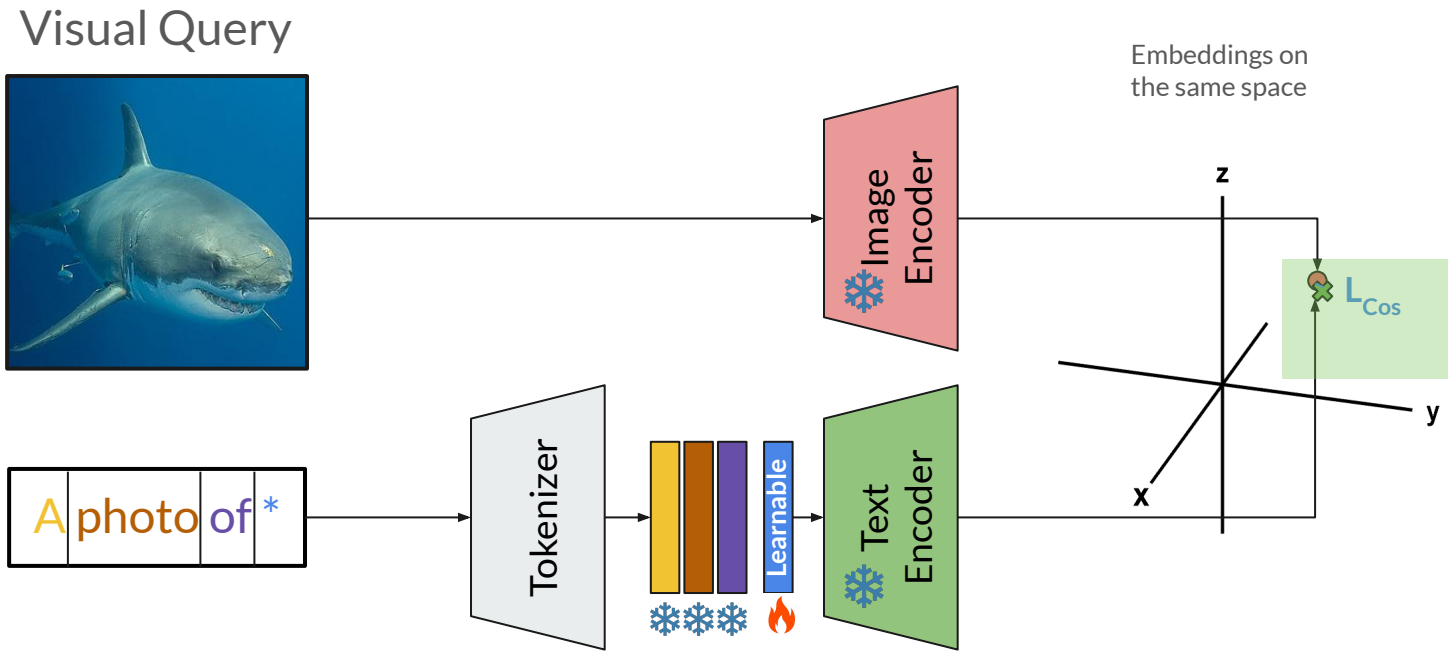
Textual Inversion in Continuous Token Space



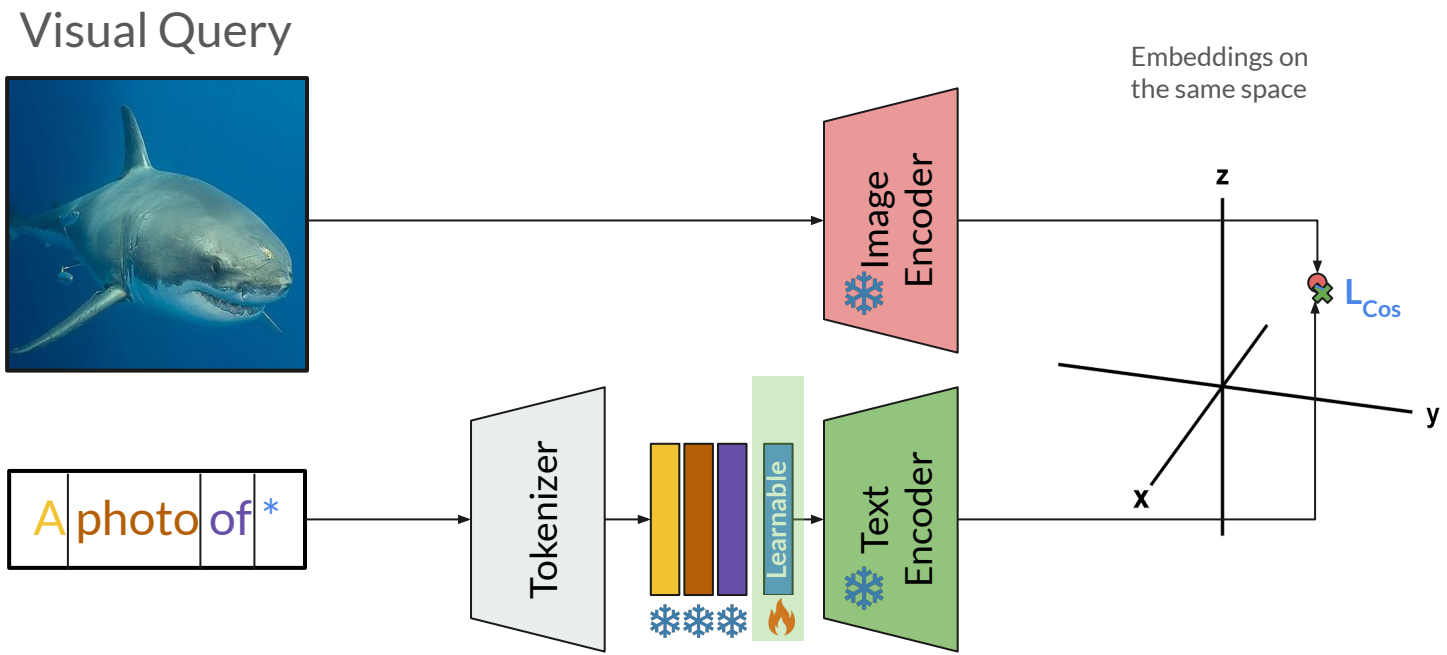
Textual Inversion in Continuous Token Space



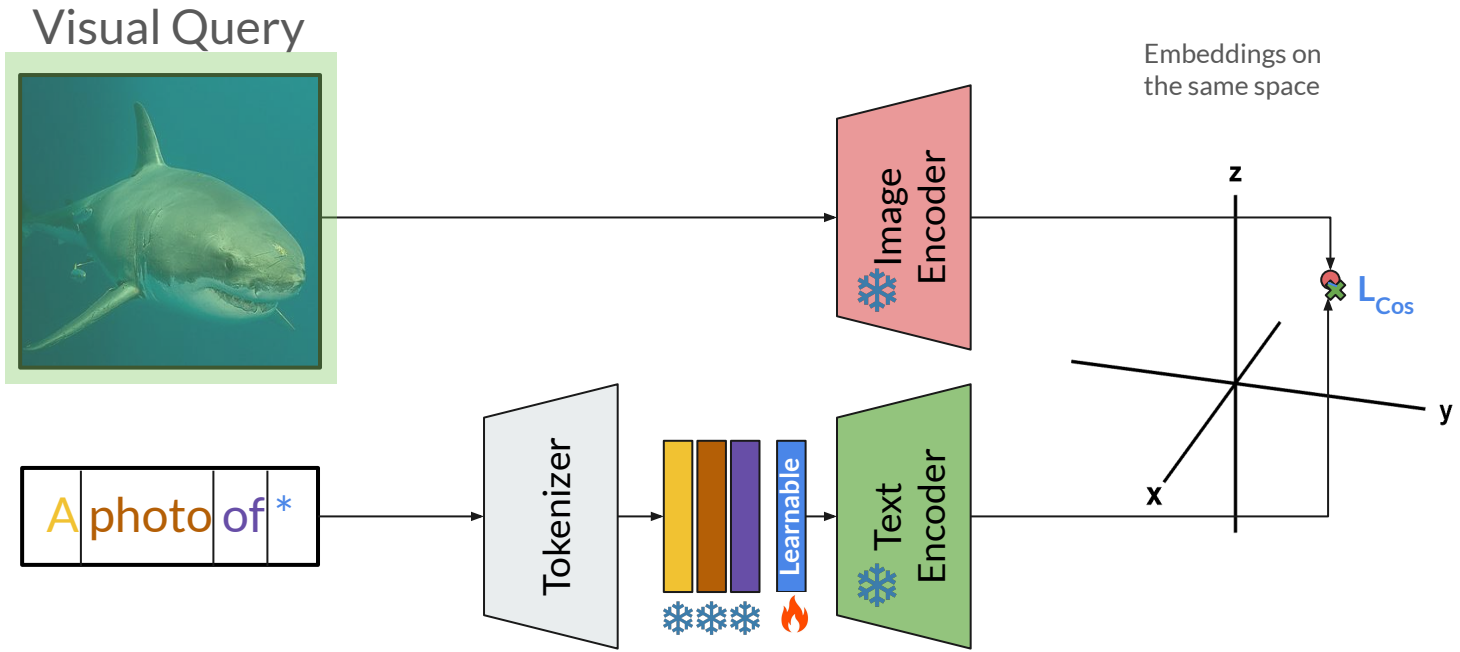
Textual Inversion in Continuous Token Space



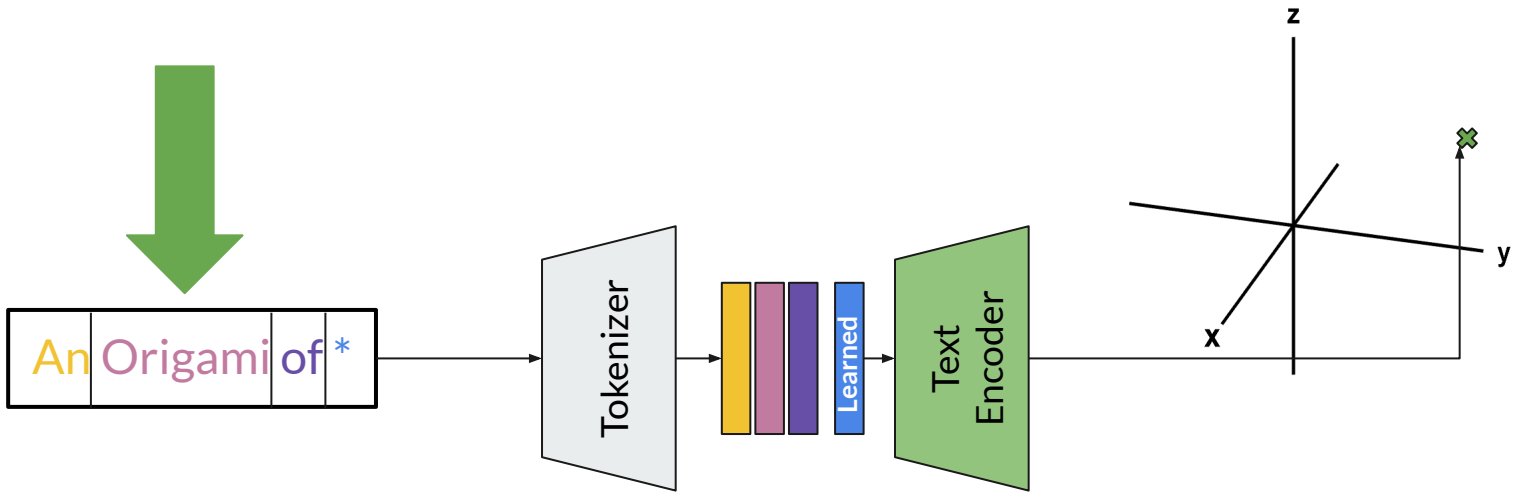
Textual Inversion in Continuous Token Space



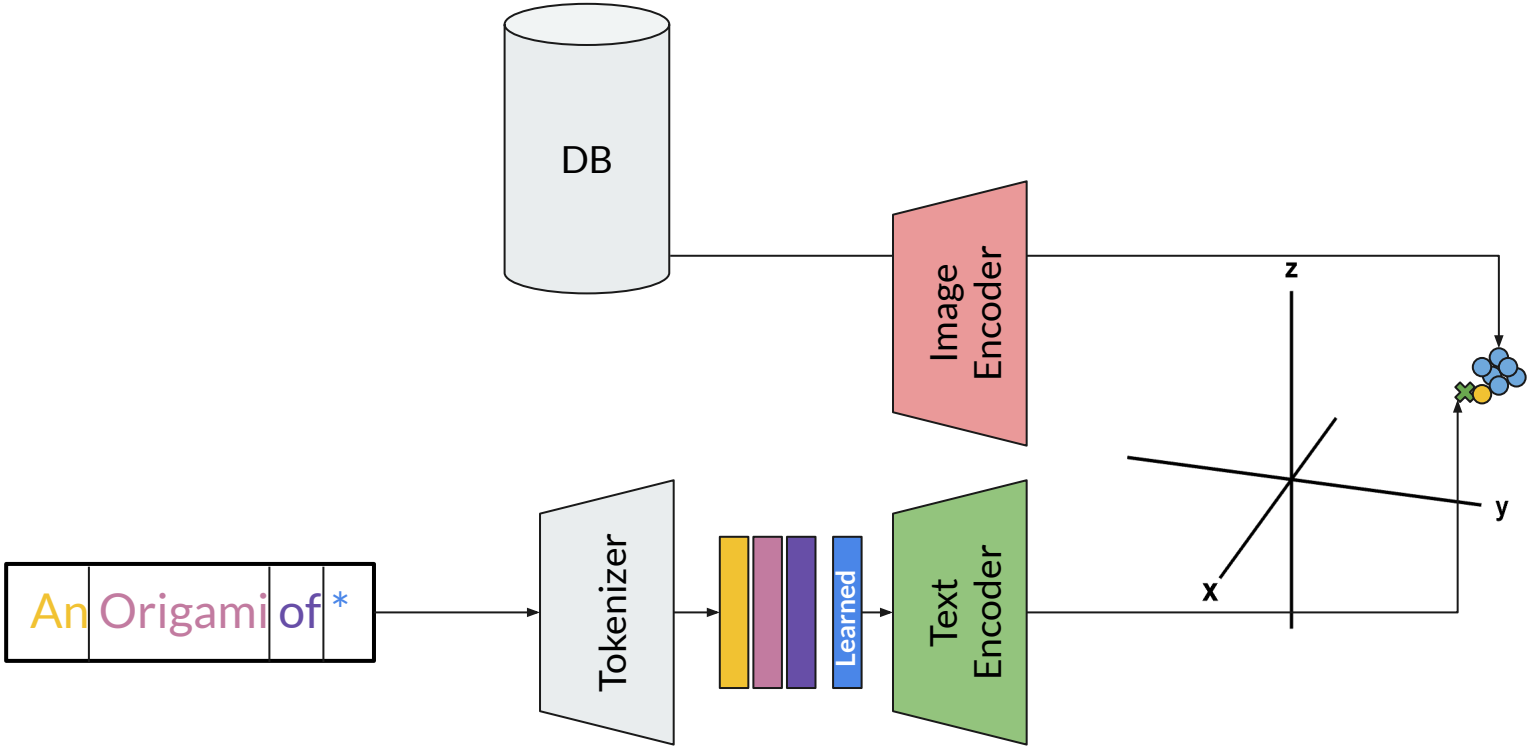
Textual Inversion in Continuous Token Space



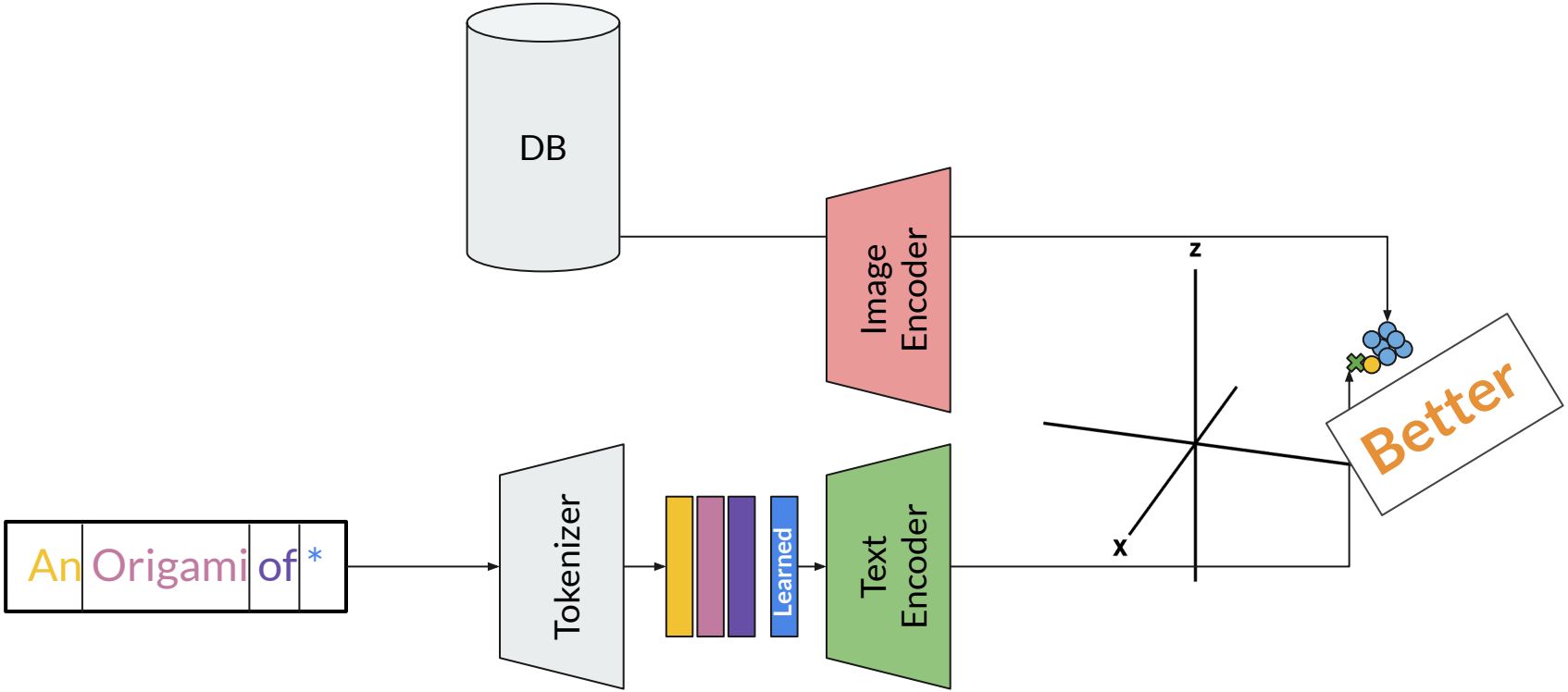
Textual Inversion in Continuous Token Space



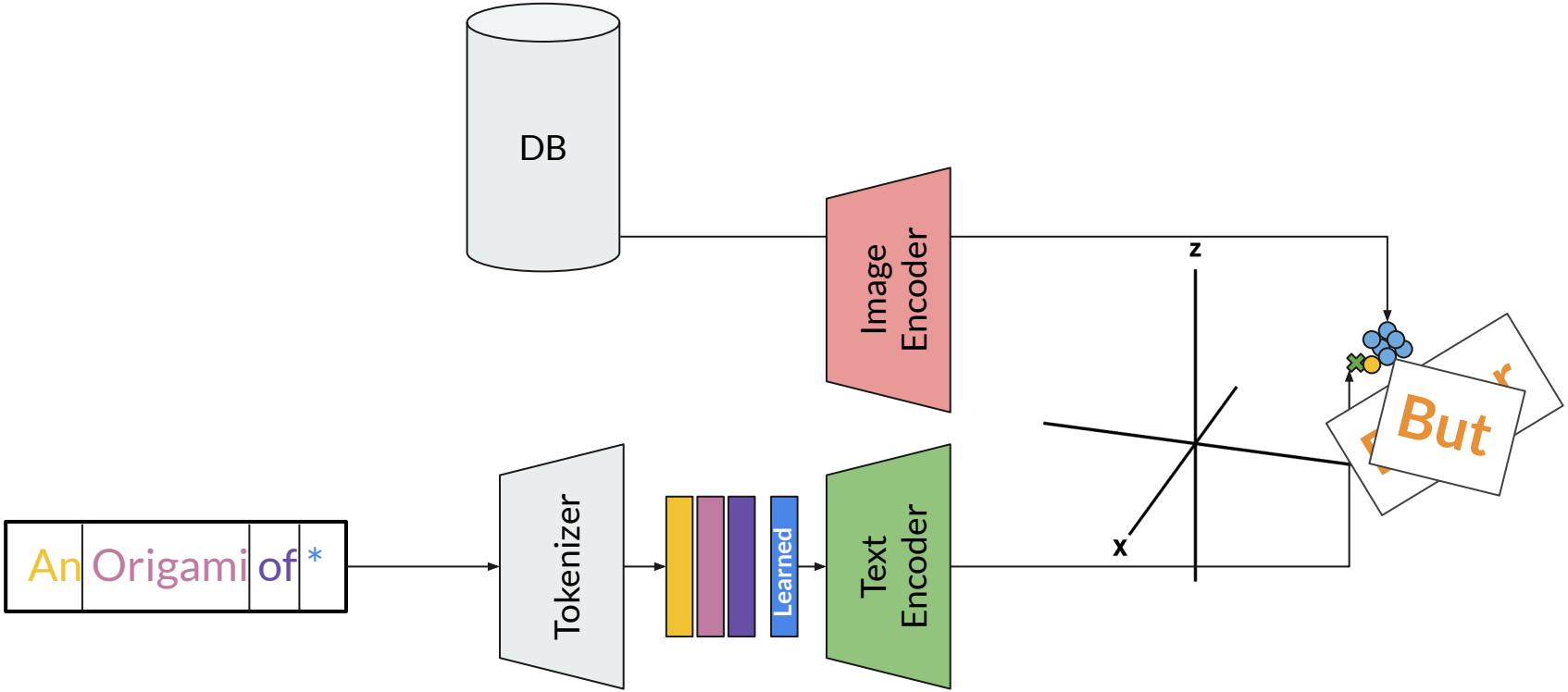
Textual Inversion in Continuous Token Space



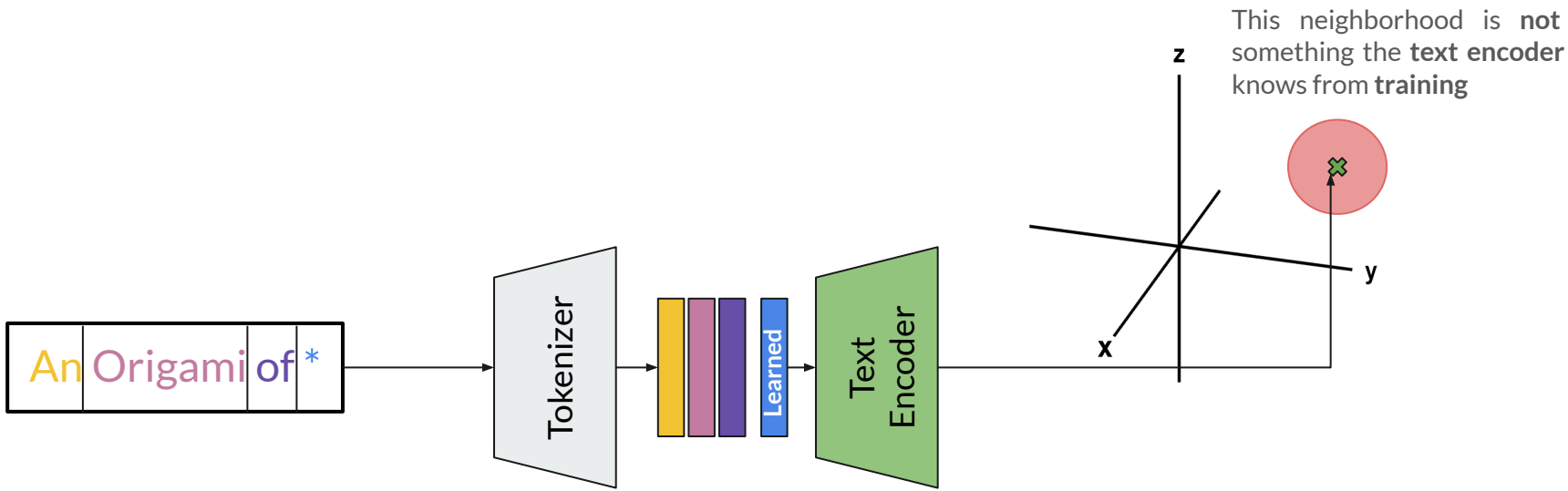
Textual Inversion in Continuous Token Space



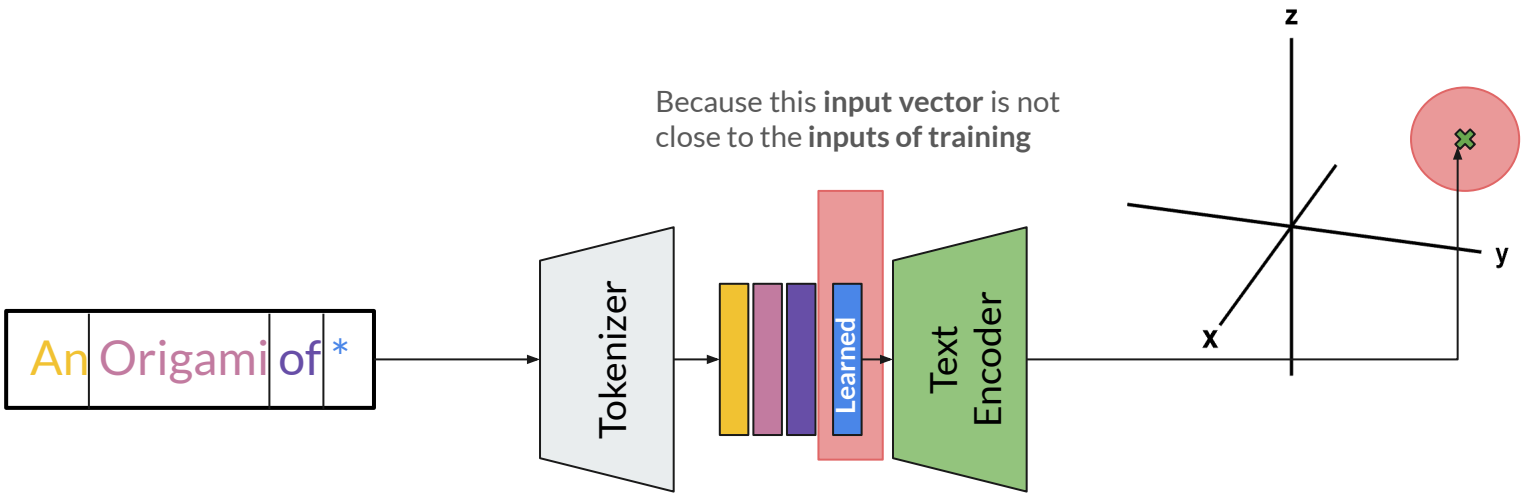
Textual Inversion in Continuous Token Space



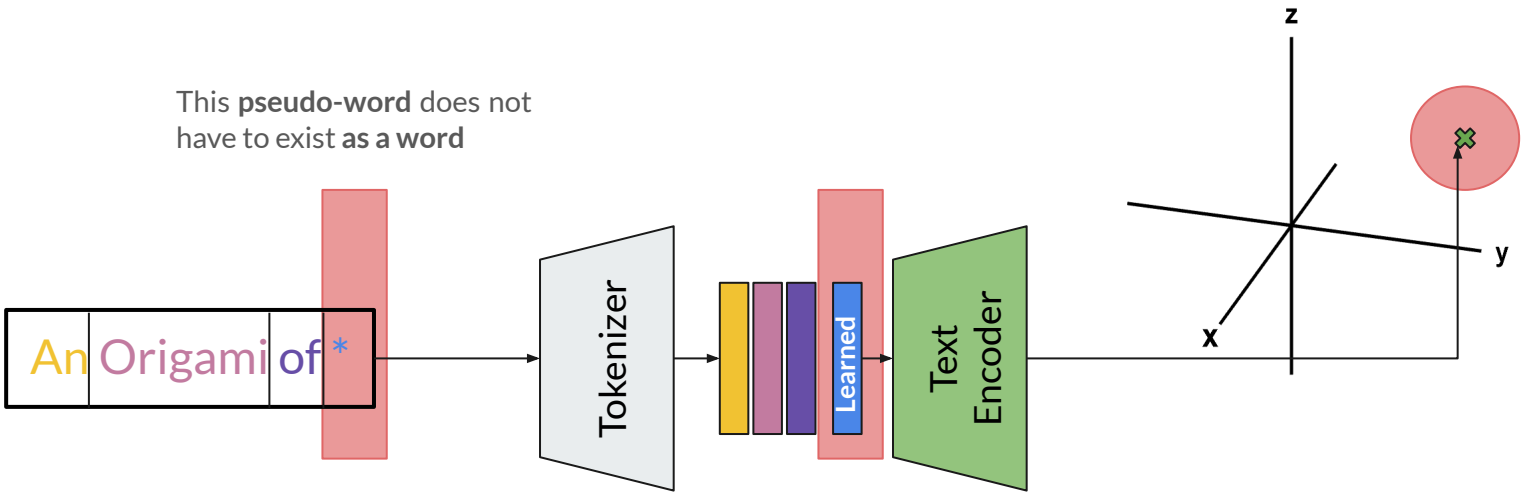
Textual Inversion in Continuous Token Space



Textual Inversion in Continuous Token Space

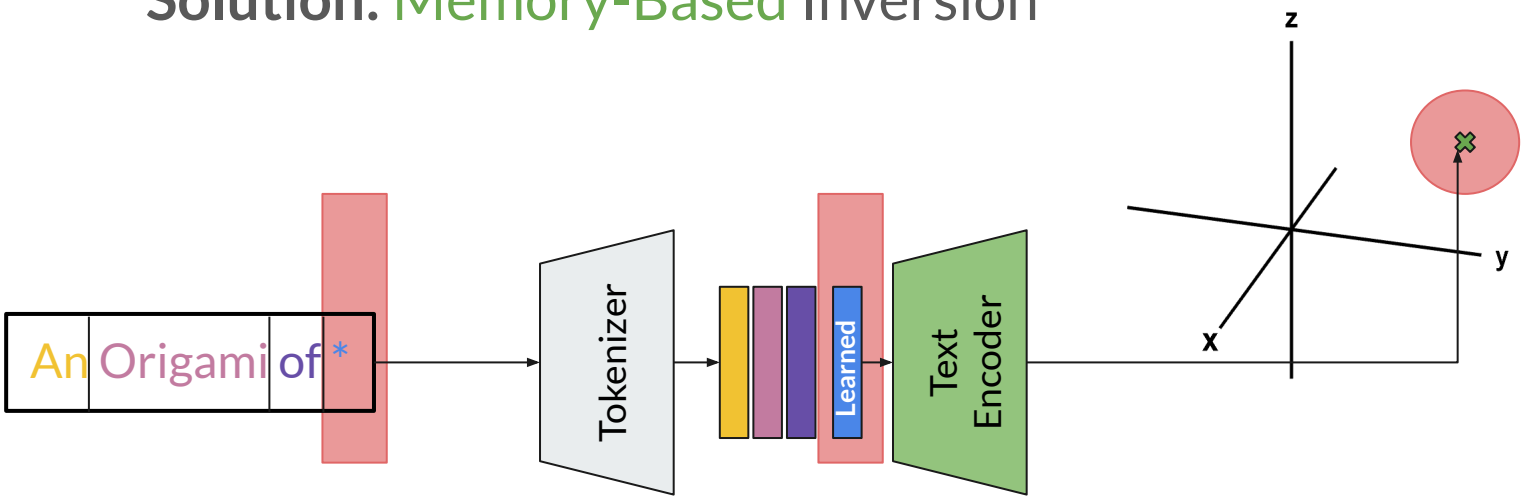


Textual Inversion in Continuous Token Space

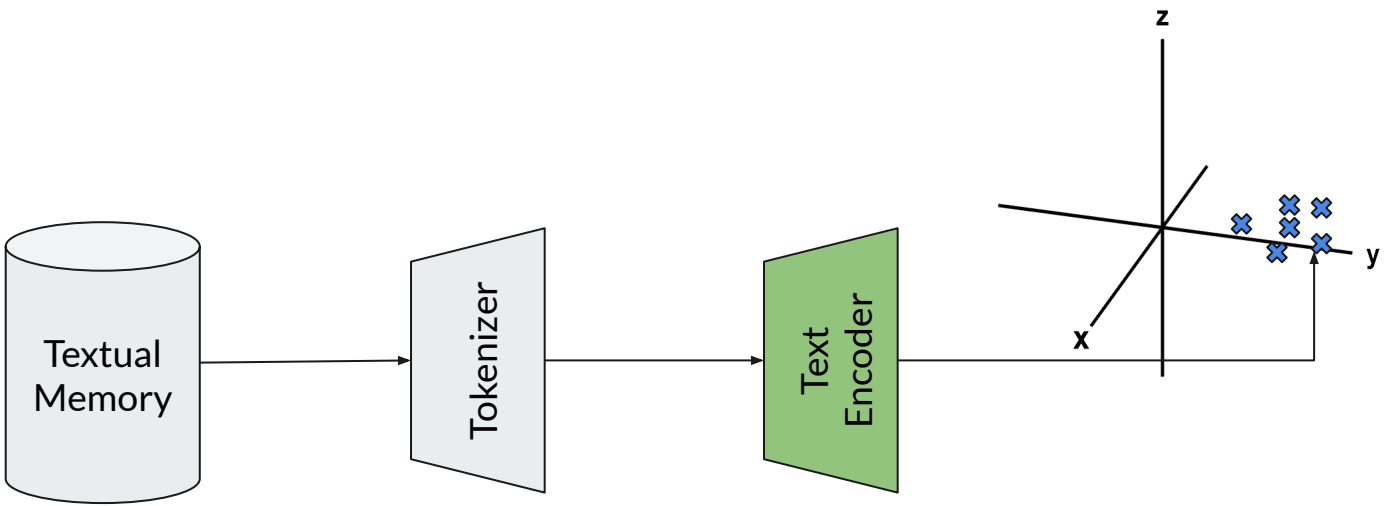


Textual Inversion in Continuous Token Space

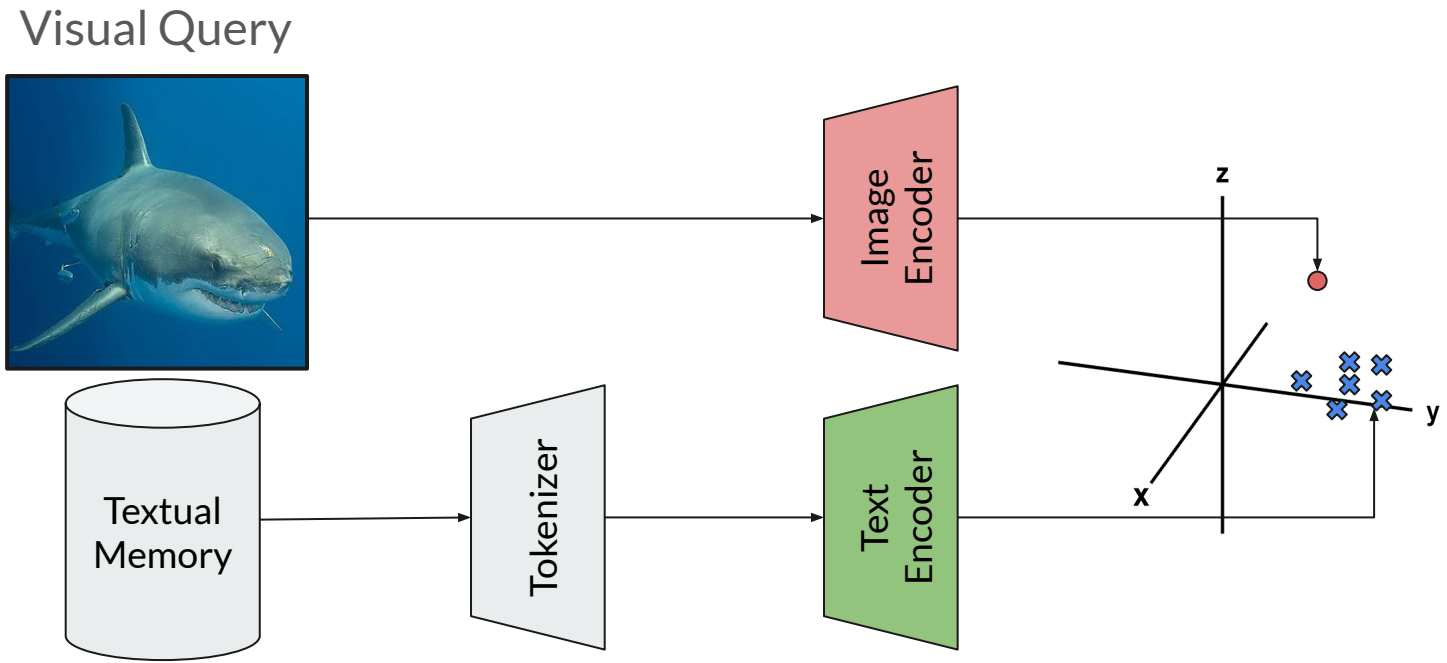
Solution: Memory-Based Inversion



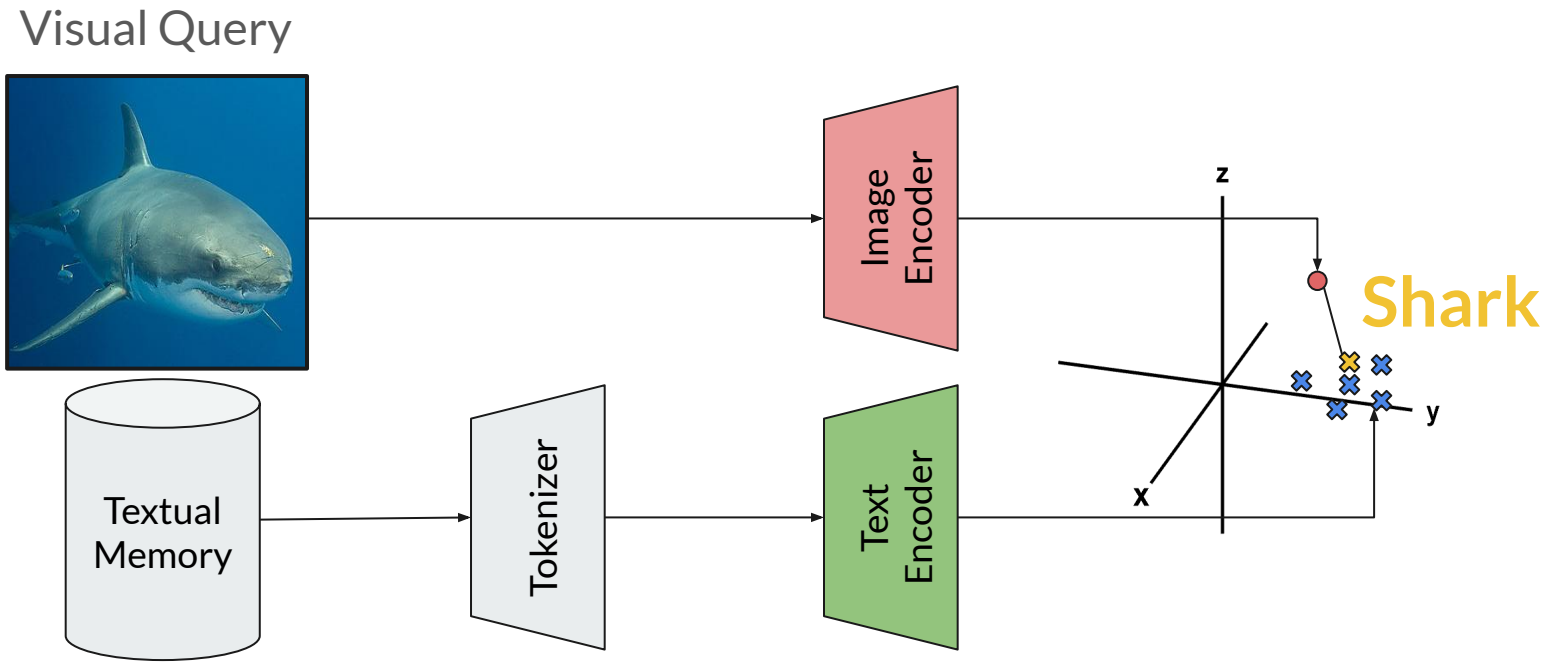
Memory-Based Textual Inversion in Text Space



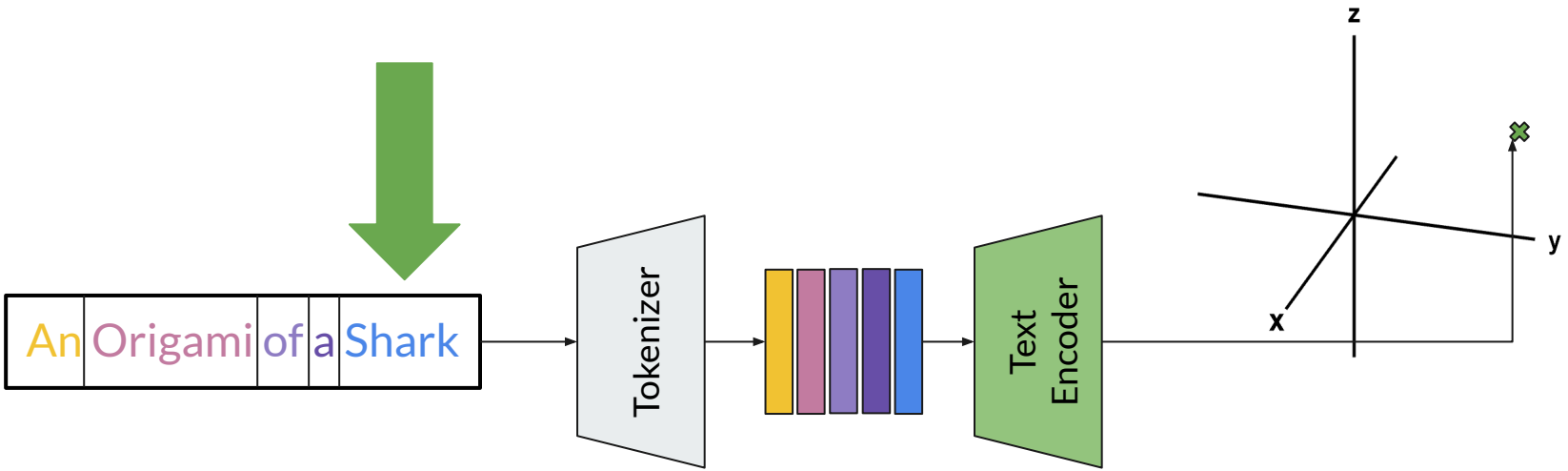
Memory-Based Textual Inversion



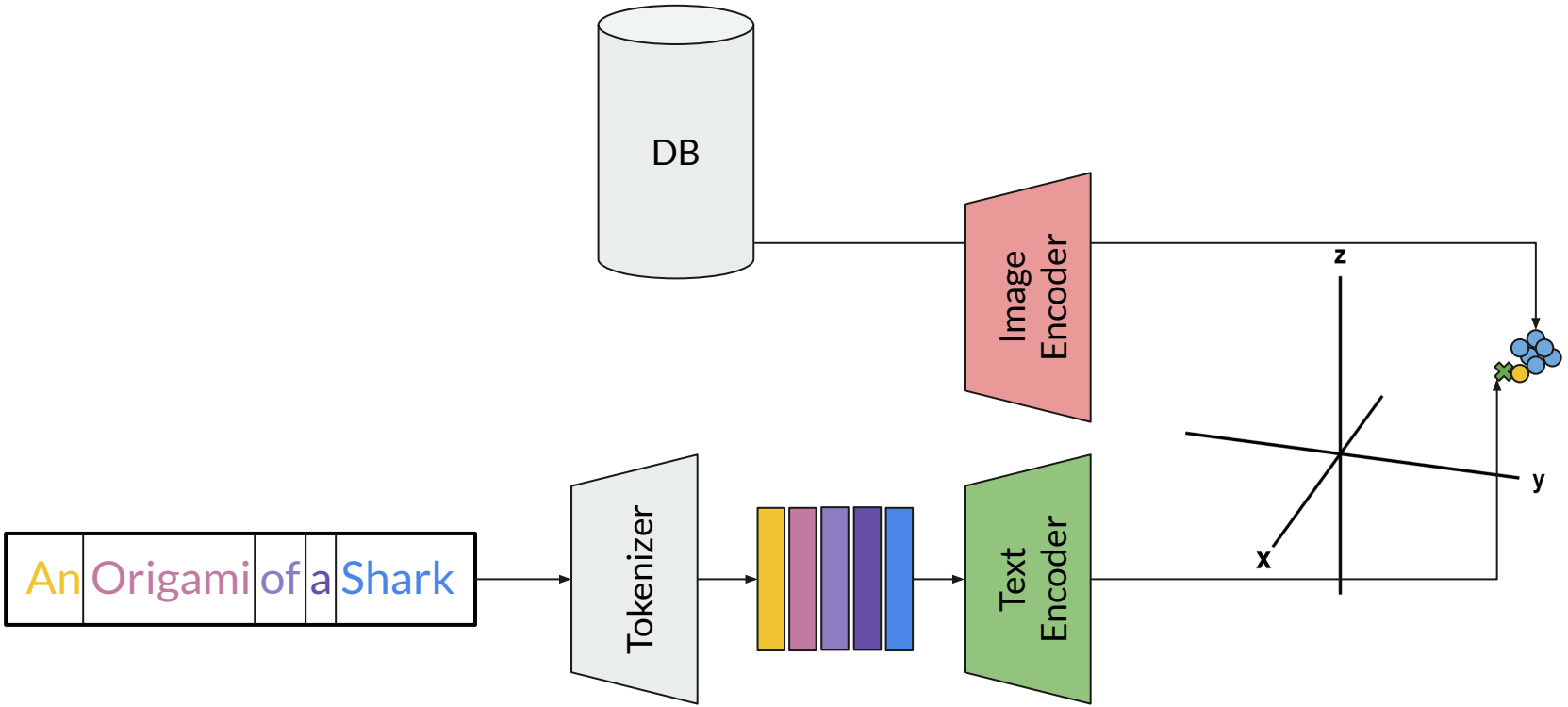
Memory-Based Textual Inversion



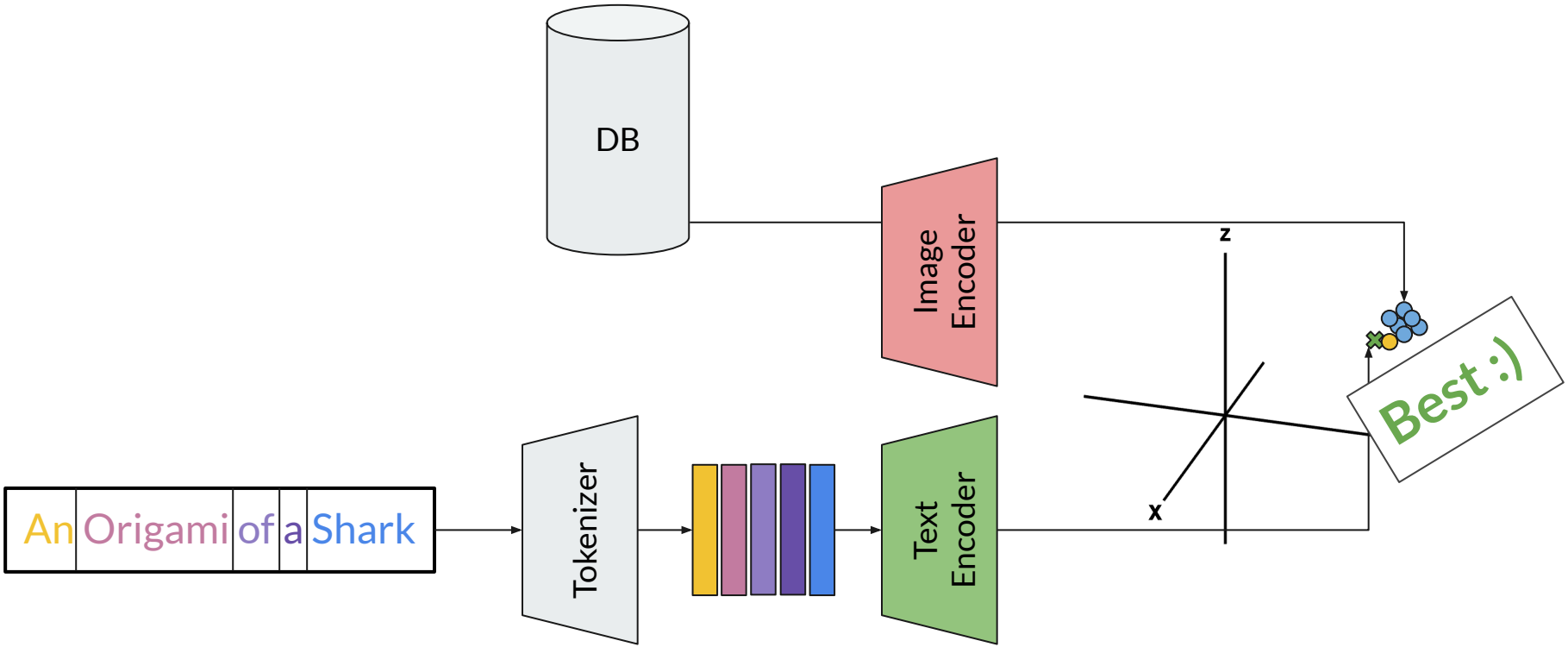
Memory-Based Textual Inversion



Memory-Based Textual Inversion



Memory-Based Textual Inversion



Memory-Based Textual Inversion

	Avg					ImageNet-R					MiniDn					NICO++					LTLL				
m	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7	9.3	8.9	8.5	8.4	10.2	24.3	24.2	22.7	21.9	20.8	15.9	15.9	16.0	16.0	13.7	28.4	28.2	27.5	26.5	26.2
W	26.9	30.6	31.6	31.6	31.1	28.5	30.2	29.9	29.4	28.4	34.9	37.7	37.3	36.8	36.2	22.3	25.6	26.1	26.1	25.9	22.0	29.1	33.2	34.0	33.8

Memory-Based Textual Inversion

Textual Inversion

	Avg					ImageNet-R					MiniDn					NICO++					LTLL				
m	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7	9.3	8.9	8.5	8.4	10.2	24.3	24.2	22.7	21.9	20.8	15.9	15.9	16.0	16.0	13.7	28.4	28.2	27.5	26.5	26.2
W	26.9	30.6	31.6	31.6	31.1	28.5	30.2	29.9	29.4	28.4	34.9	37.7	37.3	36.8	36.2	22.3	25.6	26.1	26.1	25.9	22.0	29.1	33.2	34.0	33.8

Memory-Based Textual Inversion

Memory-Based Textual Inversion

Problem on our Instance Level Dataset

	Avg					ImageNet-R					MiniDn					NICO++					LTLL				
m	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7	9.3	8.9	8.5	8.4	10.2	24.3	24.2	22.7	21.9	20.8	15.9	15.9	16.0	16.0	13.7	28.4	28.2	27.5	26.5	26.2
W	26.9	30.6	31.6	31.6	31.1	28.5	30.2	29.9	29.4	28.4	34.9	37.7	37.3	36.8	36.2	22.3	25.6	26.1	26.1	25.9	22.0	29.1	33.2	34.0	33.8

cartoon

photo

autumn

today

Memory-Based Textual Inversion

A single word cannot describe an instance well

	Avg					ImageNet-R					MiniDn					NICO++					LTLL				
m	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7	9.3	8.9	8.5	8.4	10.2	24.3	24.2	22.7	21.9	20.8	15.9	15.9	16.0	16.0	13.7	28.4	28.2	27.5	26.5	26.2
W	26.9	30.6	31.6	31.6	31.1	28.5	30.2	29.9	29.4	28.4	34.9	37.7	37.3	36.8	36.2	22.3	25.6	26.1	26.1	25.9	22.0	29.1	33.2	34.0	33.8

Memory-Based Textual Inversion

This can be fixed by using more words

	Avg					ImageNet-R					MiniDn					NICO++					LTLL				
m	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7	9.3	8.9	8.5	8.4	10.2	24.3	24.2	22.7	21.9	20.8	15.9	15.9	16.0	16.0	13.7	28.4	28.2	27.5	26.5	26.2
W	26.9	30.6	31.6	31.6	31.1	28.5	30.2	29.9	29.4	28.4	34.9	37.7	37.3	36.8	36.2	22.3	25.6	26.1	26.1	25.9	22.0	29.1	33.2	34.0	33.8

cartoon

photo

autumn

today

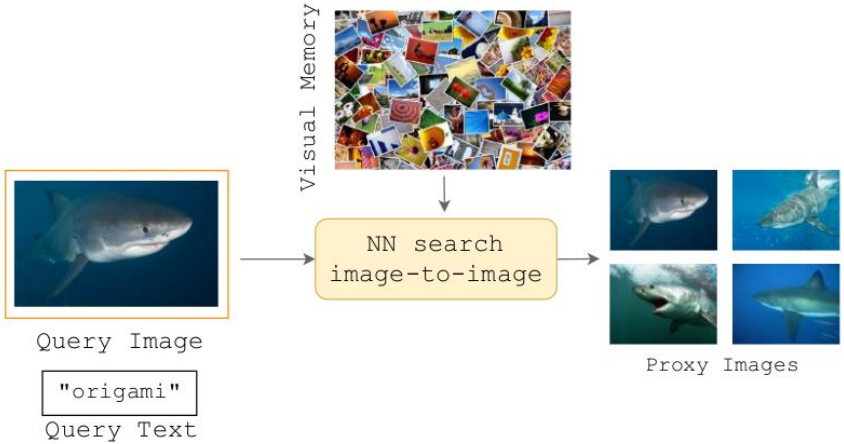
Our Full Workflow - FreeDom

Query Image

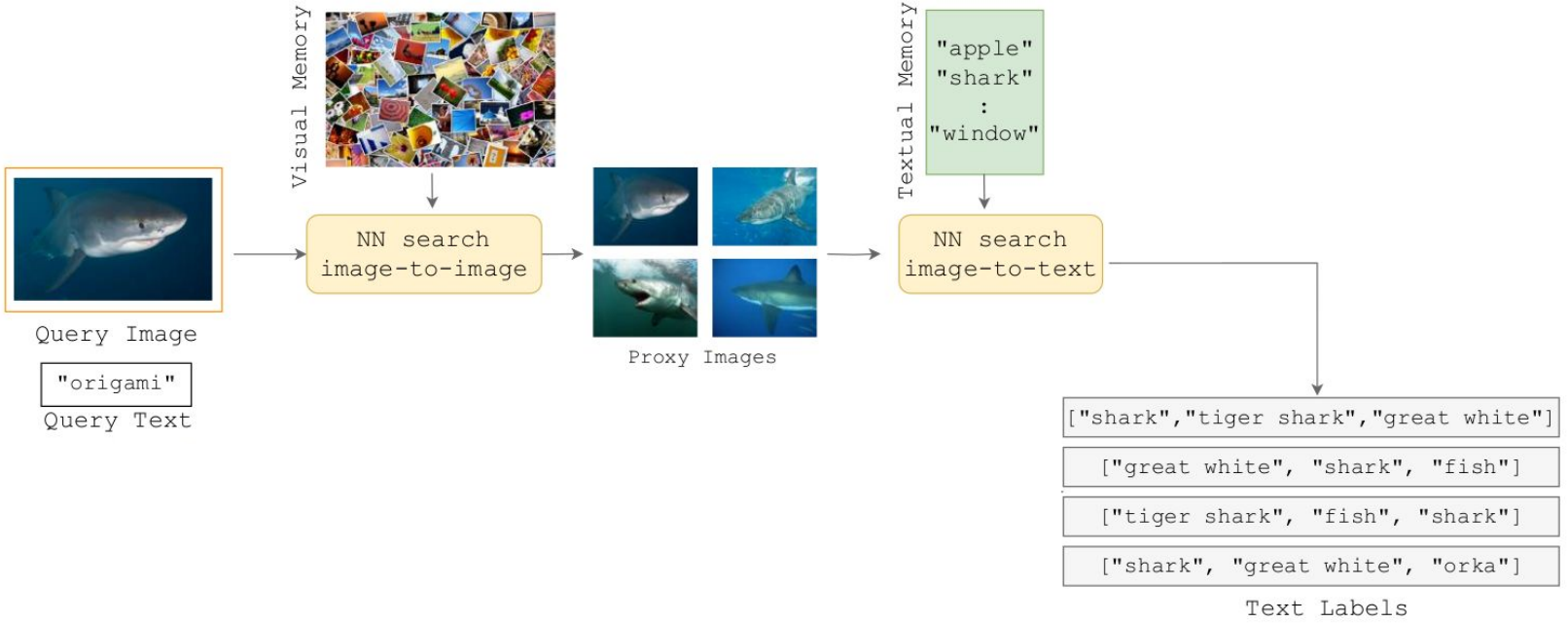
"origami"

Query Text

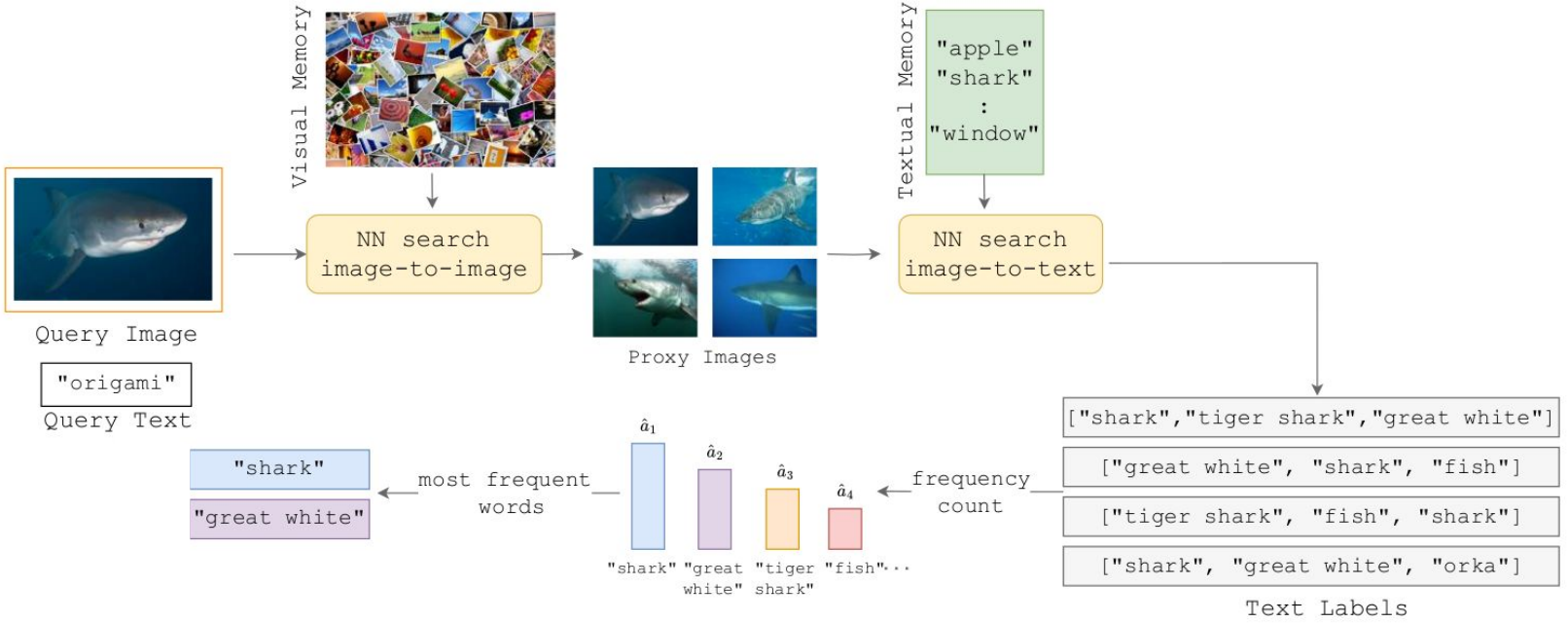
Our Full Workflow - FreeDom



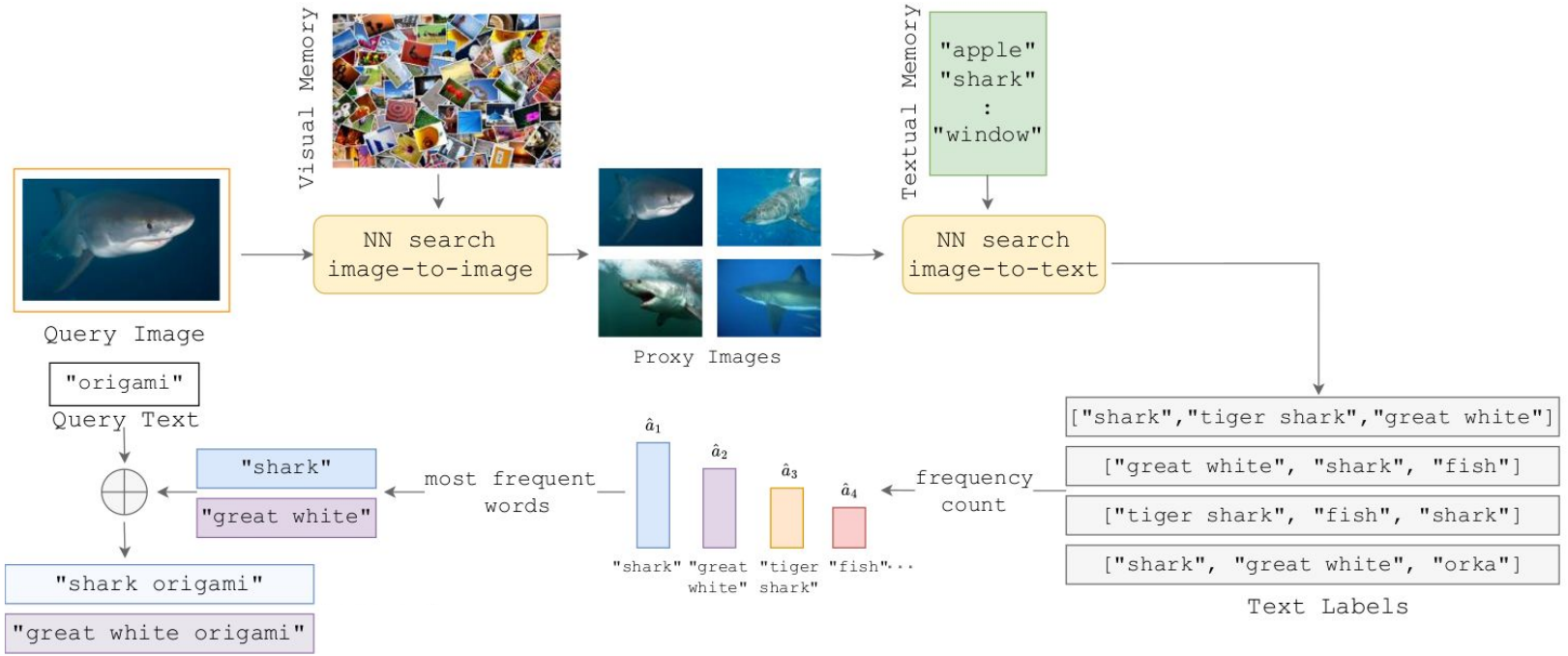
Our Full Workflow - FreeDom



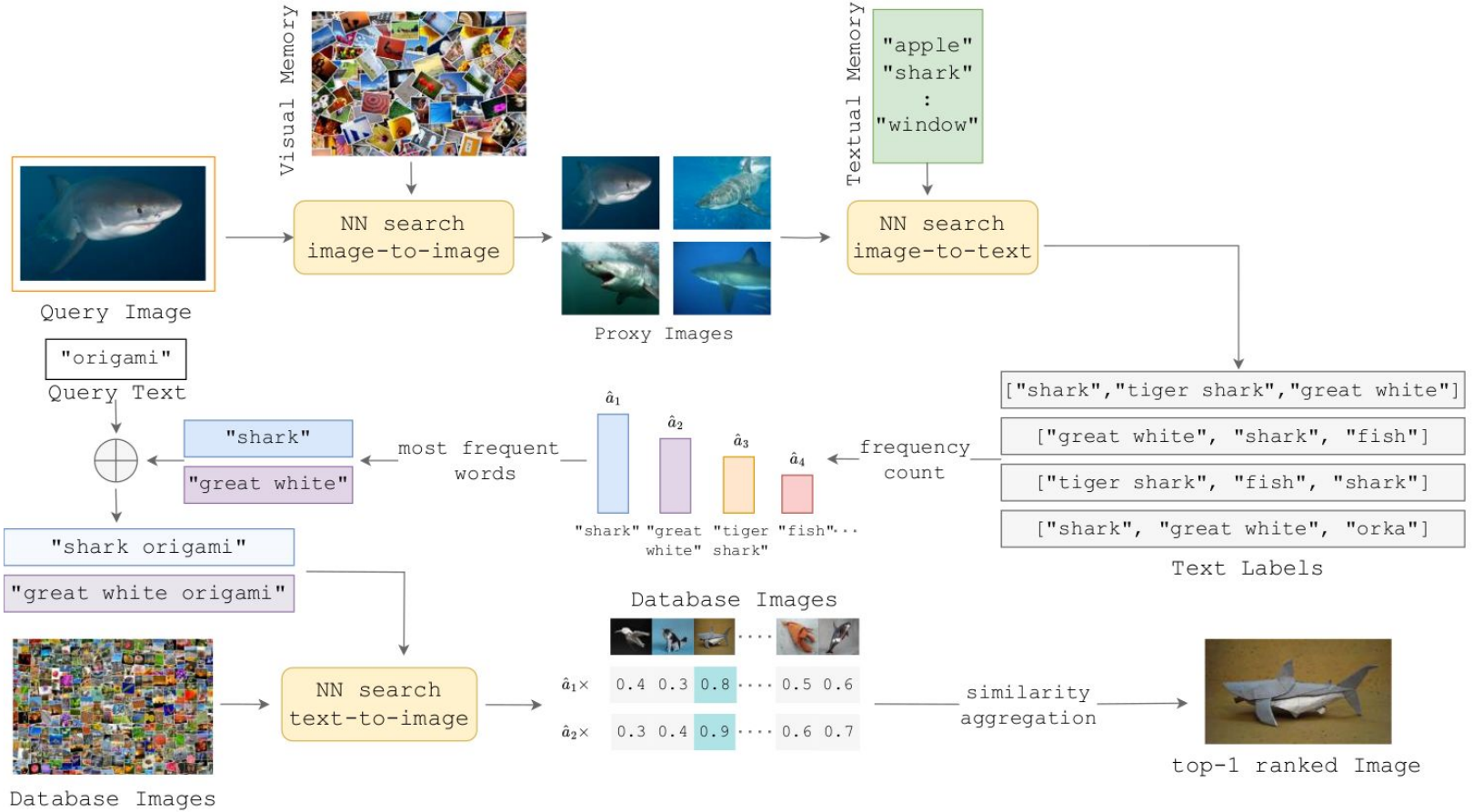
Our Full Workflow - FreeDom



Our Full Workflow - FreeDom



Our Full Workflow - FreeDom



Ablations

m	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7
L	25.1	28.1	28.0	27.4	26.3
L+	26.9	30.7	31.3	30.5	28.4
W	26.9	30.6	31.6	31.6	31.1

mAP Average Across All Datasets

Ablations

Method component

m	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7
L	25.1	28.1	28.0	27.4	26.3
L+	26.9	30.7	31.3	30.5	28.4
W	26.9	30.6	31.6	31.6	31.1

mAP Average Across All Datasets

Ablations

Number of words

m	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7
L	25.1	28.1	28.0	27.4	26.3
L+	26.9	30.7	31.3	30.5	28.4
W	26.9	30.6	31.6	31.6	31.1

mAP Average Across All Datasets

Ablations

m	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7
L	25.1	28.1	28.0	27.4	26.3
L+	26.9	30.7	31.3	30.5	28.4
W	26.9	30.6	31.6	31.6	31.1

mAP Average Across All Datasets

Ablations

Continuous Textual Inversion

m	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7
L	25.1	28.1	28.0	27.4	26.3
L+	26.9	30.7	31.3	30.5	28.4
W	26.9	30.6	31.6	31.6	31.1

mAP Average Across All Datasets

Ablations

Memory-Based Textual Inversion

m	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7
L	25.1	28.1	28.0	27.4	26.3
L+	26.9	30.7	31.3	30.5	28.4
W	26.9	30.6	31.6	31.6	31.1

mAP Average Across All Datasets

Ablations

Use of Visual Memory

m	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7
L	25.1	28.1	28.0	27.4	26.3
L+	26.9	30.7	31.3	30.5	28.4
W	26.9	30.6	31.6	31.6	31.1

mAP Average Across All Datasets

Ablations

Use Frequencies as Weights

m	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7
L	25.1	28.1	28.0	27.4	26.3
L+	26.9	30.7	31.3	30.5	28.4
W	26.9	30.6	31.6	31.6	31.1

mAP Average Across All Datasets

Ablations

Use Frequencies as Weights

m	1	3	7	10	15
SRL	19.5	19.3	18.7	18.2	17.7
L	25.1	28.1	28.0	27.4	26.3
L+	26.9	30.7	31.3	30.5	28.4
W	26.9	30.6	31.6	31.6	31.1

mAP Average Across All Datasets

Very robust region

Results

CLIP

(a) ImageNet-R

METHOD	CAR	ORI	PHO	SCU	TOY	AVG
Pic2Word	7.60	5.53	7.64	9.39	9.27	7.88
CompoDiff	13.71	10.61	8.76	15.17	16.17	12.88
WeiCom	10.07	7.61	10.06	11.26	13.38	10.47
SEARLE	18.11	9.02	9.94	17.26	15.83	14.04
MagicLens	7.79	6.33	11.02	9.94	10.57	9.13
FREEDOM	**35.97**	**11.80**	**27.97**	**36.58**	**37.21**	**29.91**

(b) MiniDomainNet

METHOD	CLIP	PAINT	PHO	SKE	AVG
Pic2Word	13.39	8.63	17.96	8.03	12.00
CompoDiff	19.06	24.27	23.41	25.05	22.95
WeiCom	7.52	7.04	15.13	4.40	8.52
SEARLE	25.04	18.72	23.75	19.61	21.78
MagicLens	24.40	17.54	28.59	9.71	20.06
FREEDOM	**41.96**	**31.65**	**41.12**	**34.36**	**37.27**

(c) NICO++

METHOD	AUT	DIM	GRA	OUT	ROC	WAT	AVG
Pic2Word	9.79	8.09	11.24	11.27	11.01	7.16	9.76
CompoDiff	10.07	7.83	10.53	11.41	11.93	10.15	10.32
WeiCom	8.58	7.39	13.04	13.17	11.32	9.73	10.54
SEARLE	13.49	13.73	17.91	17.99	15.79	11.84	15.13
MagicLens	18.76	15.17	22.14	23.61	21.99	16.30	19.66
FREEDOM	**24.35**	**24.41**	**30.06**	**30.51**	**26.92**	**20.37**	**26.10**

(d) LTLL

METHOD	TODAY	ARCHIVE	AVG
Pic2Word	17.86	24.67	21.27
CompoDiff	15.45	27.76	21.61
WeiCom	24.56	28.63	26.60
SEARLE	20.82	30.10	25.46
MagicLens	**33.77**	14.65	24.21
FREEDOM	30.95	**35.52**	**33.24**

Results

CLIP

(a) ImageNet-R

METHOD	CAR	ORI	PHO	SCU	TOY	AVG
Pic2Word	7.60	5.53	7.64	9.39	9.27	7.88
CompoDiff	13.71	10.61	8.76	15.17	16.17	12.88
WeiCom	10.07	7.61	10.06	11.26	13.38	10.47
SEARLE	18.11	9.02	9.94	17.26	15.83	14.04
MagicLens	7.79	6.33	11.02	9.94	10.57	9.13
FREEDOM	**35.97**	**11.80**	**27.97**	**36.58**	**37.21**	**29.91**

(b) MiniDomainNet

METHOD	CLIP	PAINT	PHO	SKE	AVG
Pic2Word	13.39	8.63	17.96	8.03	12.00
CompoDiff	19.06	24.27	23.41	25.05	22.95
WeiCom	7.52	7.04	15.13	4.40	8.52
SEARLE	25.04	18.72	23.75	19.61	21.78
MagicLens	24.40	17.54	28.59	9.71	20.06
FREEDOM	**41.96**	**31.65**	**41.12**	**34.36**	**37.27**

(c) NICO++

METHOD	AUT	DIM	GRA	OUT	ROC	WAT	AVG
Pic2Word	9.79	8.09	11.24	11.27	11.01	7.16	9.76
CompoDiff	10.07	7.83	10.53	11.41	11.93	10.15	10.32
WeiCom	8.58	7.39	13.04	13.17	11.32	9.73	10.54
SEARLE	13.49	13.73	17.91	17.99	15.79	11.84	15.13
MagicLens	18.76	15.17	22.14	23.61	21.99	16.30	19.66
FREEDOM	**24.35**	**24.41**	**30.06**	**30.51**	**26.92**	**20.37**	**26.10**

(d) LTLL

METHOD	TODAY	ARCHIVE	AVG
Pic2Word	17.86	24.67	21.27
CompoDiff	15.45	27.76	21.61
WeiCom	24.56	28.63	26.60
SEARLE	20.82	30.10	25.46
MagicLens	**33.77**	14.65	24.21
FREEDOM	30.95	**35.52**	**33.24**

SigLIP

(a) ImageNet-R

METHOD	CAR	ORI	PHO	SCU	TOY	AVG
Text	0.88	0.80	0.62	0.95	0.90	0.83
Image	4.97	3.70	0.84	8.18	7.40	5.02
Text × Image	6.57	4.34	4.89	6.46	7.46	5.94
Text + Image	7.88	5.84	3.08	13.50	12.71	8.60
FREEDOM	**49.46**	**27.12**	**38.11**	**47.52**	**46.90**	**41.82**

(b) MiniDomainNet

METHOD	CLIP	PAINT	PHO	SKE	AVG
Text	0.76	0.72	0.76	0.75	0.74
Image	5.07	7.53	3.68	6.15	5.61
Text × Image	3.00	2.60	4.34	3.18	3.28
Text + Image	7.79	11.33	10.80	9.02	9.74
FREEDOM	**57.14**	**45.47**	**59.71**	**52.21**	**53.63**

(c) NICO++

METHOD	AUT	DIM	GRA	OUT	ROC	WAT	AVG
Text	1.08	1.13	1.04	1.26	1.10	1.11	1.12
Image	6.19	5.19	5.42	7.67	7.44	5.62	6.25
Text × Image	2.31	2.91	3.26	3.53	3.25	2.90	3.03
Text + Image	8.35	7.19	8.08	11.42	10.57	8.12	8.95
FREEDOM	**30.28**	**29.96**	**33.86**	**37.16**	**33.14**	**26.49**	**31.81**

(d) LTLL

METHOD	TODAY	ARCHIVE	AVG
Text	3.84	5.02	4.43
Image	10.25	28.14	19.20
Text × Image	4.87	3.49	4.18
Text + Image	10.16	26.73	18.44
FREEDOM	**27.45**	**47.00**	**37.22**

Results

CLIP

(a) ImageNet-R

METHOD	CAR	ORI	PHO	SCU	TOY	AVG
Pic2Word	7.60	5.53	7.64	9.39	9.27	7.88
CompoDiff	13.71	10.61	8.76	15.17	16.17	12.88
WeiCom	10.07	7.61	10.06	11.26	13.38	10.47
SEARLE	18.11	9.02	9.94	17.26	15.83	14.04
MagicLens	7.79	6.33	11.02	9.94	10.57	9.13
FREEDOM	**35.97**	**11.80**	**27.97**	**36.58**	**37.21**	**29.91**

(b) MiniDomainNet

METHOD	CLIP	PAINT	PHO	SKE	AVG
Pic2Word	13.39	8.63	17.96	8.03	12.00
CompoDiff	19.06	24.27	23.41	25.05	22.95
WeiCom	7.52	7.04	15.13	4.40	8.52
SEARLE	25.04	18.72	23.75	19.61	21.78
MagicLens	24.40	17.54	28.59	9.71	20.06
FREEDOM	**41.96**	**31.65**	**41.12**	**34.36**	**37.27**

(c) NICO++

METHOD	AUT	DIM	GRA	OUT	ROC	WAT	AVG
Pic2Word	9.79	8.09	11.24	11.27	11.01	7.16	9.76
CompoDiff	10.07	7.83	10.53	11.41	11.93	10.15	10.32
WeiCom	8.58	7.39	13.04	13.17	11.32	9.73	10.54
SEARLE	13.49	13.73	17.91	17.99	15.79	11.84	15.13
MagicLens	18.76	15.17	22.14	23.61	21.99	16.30	19.66
FREEDOM	**24.35**	**24.41**	**30.06**	**30.51**	**26.92**	**20.37**	**26.10**

(d) LTLL

METHOD	TODAY	ARCHIVE	AVG
Pic2Word	17.86	24.67	21.27
CompoDiff	15.45	27.76	21.61
WeiCom	24.56	28.63	26.60
SEARLE	20.82	30.10	25.46
MagicLens	**33.77**	14.65	24.21
FREEDOM	30.95	**35.52**	**33.24**

SigLIP

(a) ImageNet-R

METHOD	CAR	ORI	PHO	SCU	TOY	AVG
Text	0.88	0.80	0.62	0.95	0.90	0.83
Image	4.97	3.70	0.84	8.18	7.40	5.02
Text × Image	6.57	4.34	4.89	6.46	7.46	5.94
Text + Image	7.88	5.84	3.08	13.50	12.71	8.60
FREEDOM	**49.46**	**27.12**	**38.11**	**47.52**	**46.90**	**41.82**

(b) MiniDomainNet

METHOD	CLIP	PAINT	PHO	SKE	AVG
Text	0.76	0.72	0.76	0.75	0.74
Image	5.07	7.53	3.68	6.15	5.61
Text × Image	3.00	2.60	4.34	3.18	3.28
Text + Image	7.79	11.33	10.80	9.02	9.74
FREEDOM	**57.14**	**45.47**	**59.71**	**52.21**	**53.63**

(c) NICO++

METHOD	AUT	DIM	GRA	OUT	ROC	WAT	AVG
Text	1.08	1.13	1.04	1.26	1.10	1.11	1.12
Image	6.19	5.19	5.42	7.67	7.44	5.62	6.25
Text × Image	2.31	2.91	3.26	3.53	3.25	2.90	3.03
Text + Image	8.35	7.19	8.08	11.42	10.57	8.12	8.95
FREEDOM	**30.28**	**29.96**	**33.86**	**37.16**	**33.14**	**26.49**	**31.81**

(d) LTLL

METHOD	TODAY	ARCHIVE	AVG
Text	3.84	5.02	4.43
Image	10.25	28.14	19.20
Text × Image	4.87	3.49	4.18
Text + Image	10.16	26.73	18.44
FREEDOM	**27.45**	**47.00**	**37.22**

Transferability: mAP **increase** from **4.0 to 16.4** just by changing the **backbone**!

Thanks