ConceptExpress: Harnessing Diffusion Models for Single-image Unsupervised Concept Extraction

Shaozhe Hao, Kai Han, Zhengyao Lv, Shihao Zhao, Kwan-Yee K. Wong

The University of Hong Kong

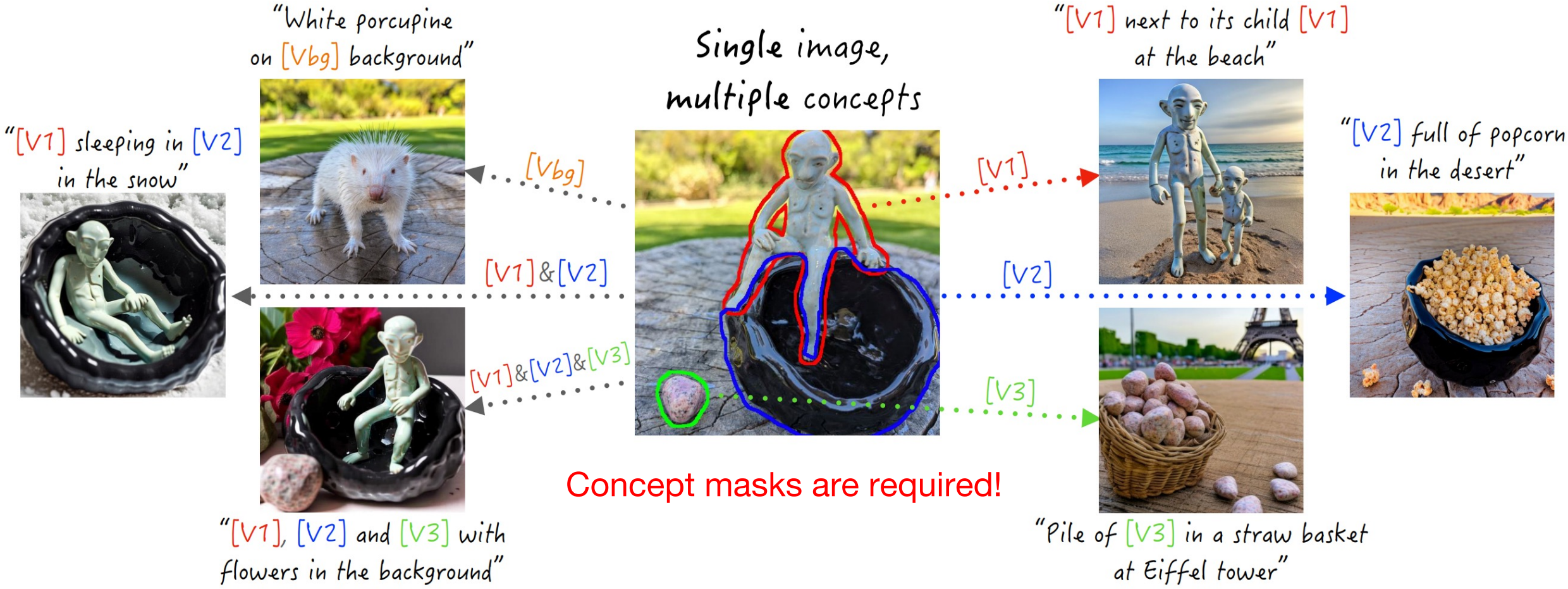
Diffusion models can learn visual concepts

Textual Inversion [Gal et al. 2022]

DreamBooth [Ruiz et al. 2022]

Learn multiple visual concepts from a single image?

Break-A-Scene [Avrahami et al. 2022]



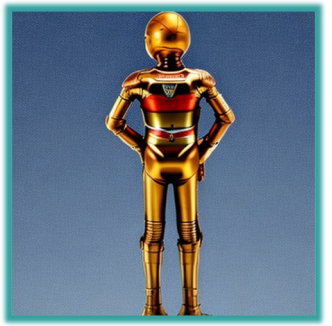
Concept masks are required!

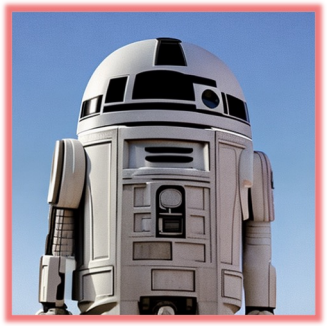
Unsupervised concept extraction (UCE)

Given a single image, can we extract the visual concepts in it without any information (**unsupervised**)?

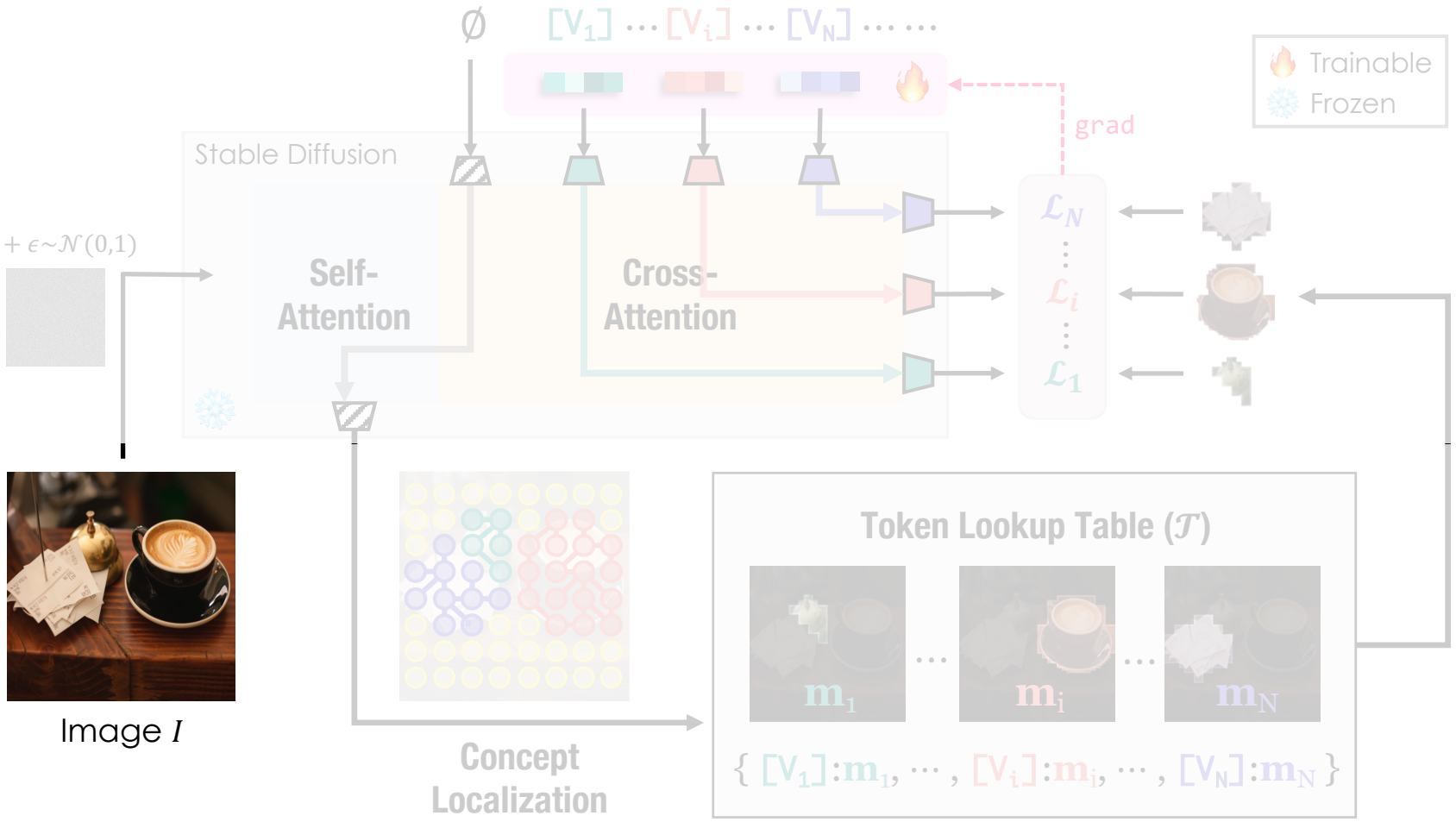
By "**Unsupervised**", we mean:

- No concept descriptors
- No object masks
- No instance numbers



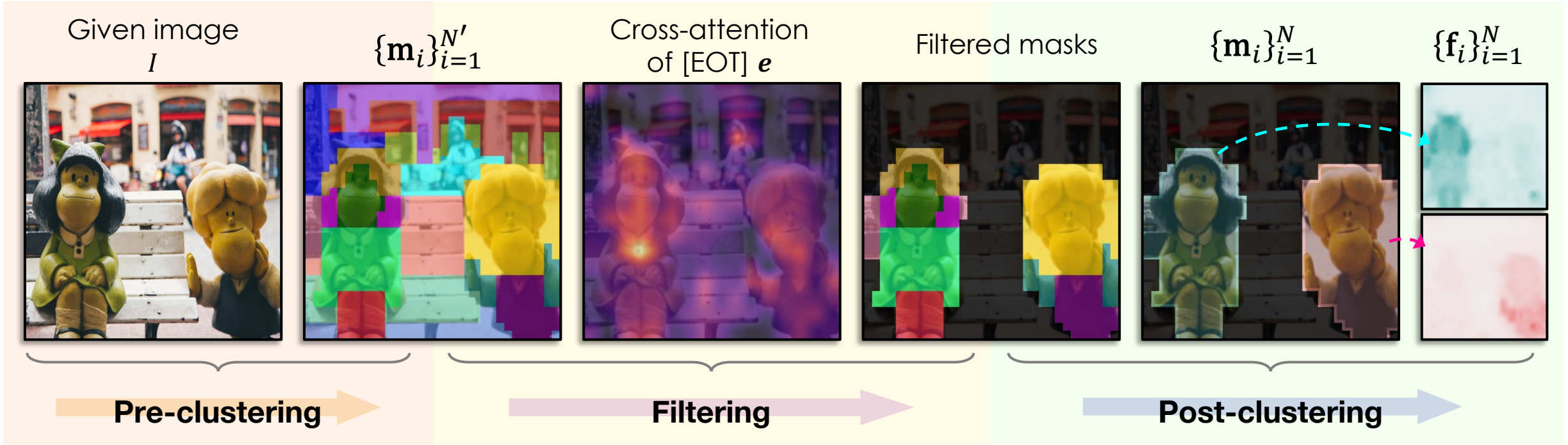


ConceptExpress



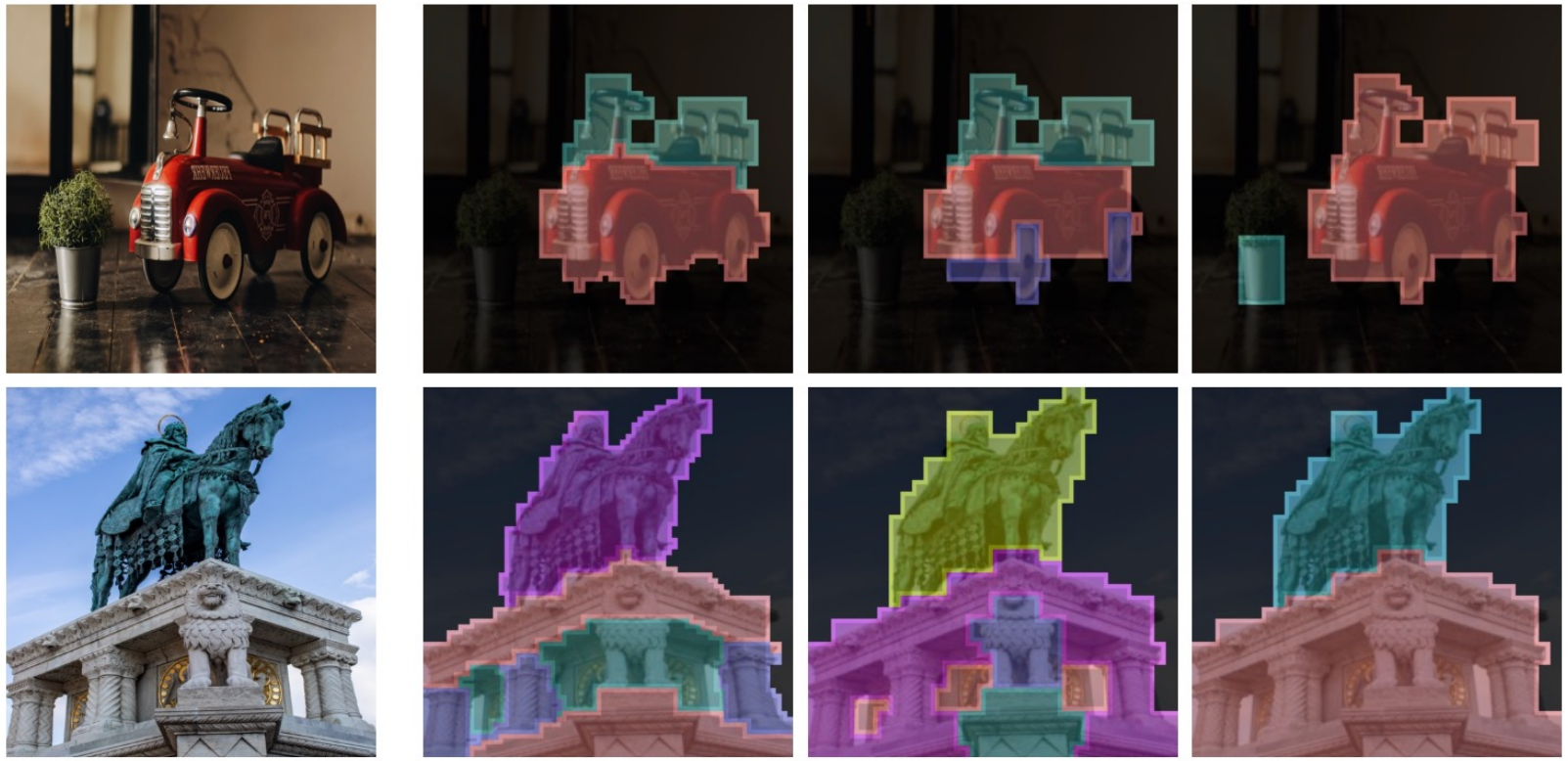
ConceptExpress - Concept localization

1. **Cluster on self-attention map**: Aggregate semantically close patches.
2. **Filter background**: Use [EOT] token to filter out background patches.
3. **Automatic stopping point**: Obtain final concept masks and feature maps.



ConceptExpress - Concept localization

Localization results using different (clustering) methods

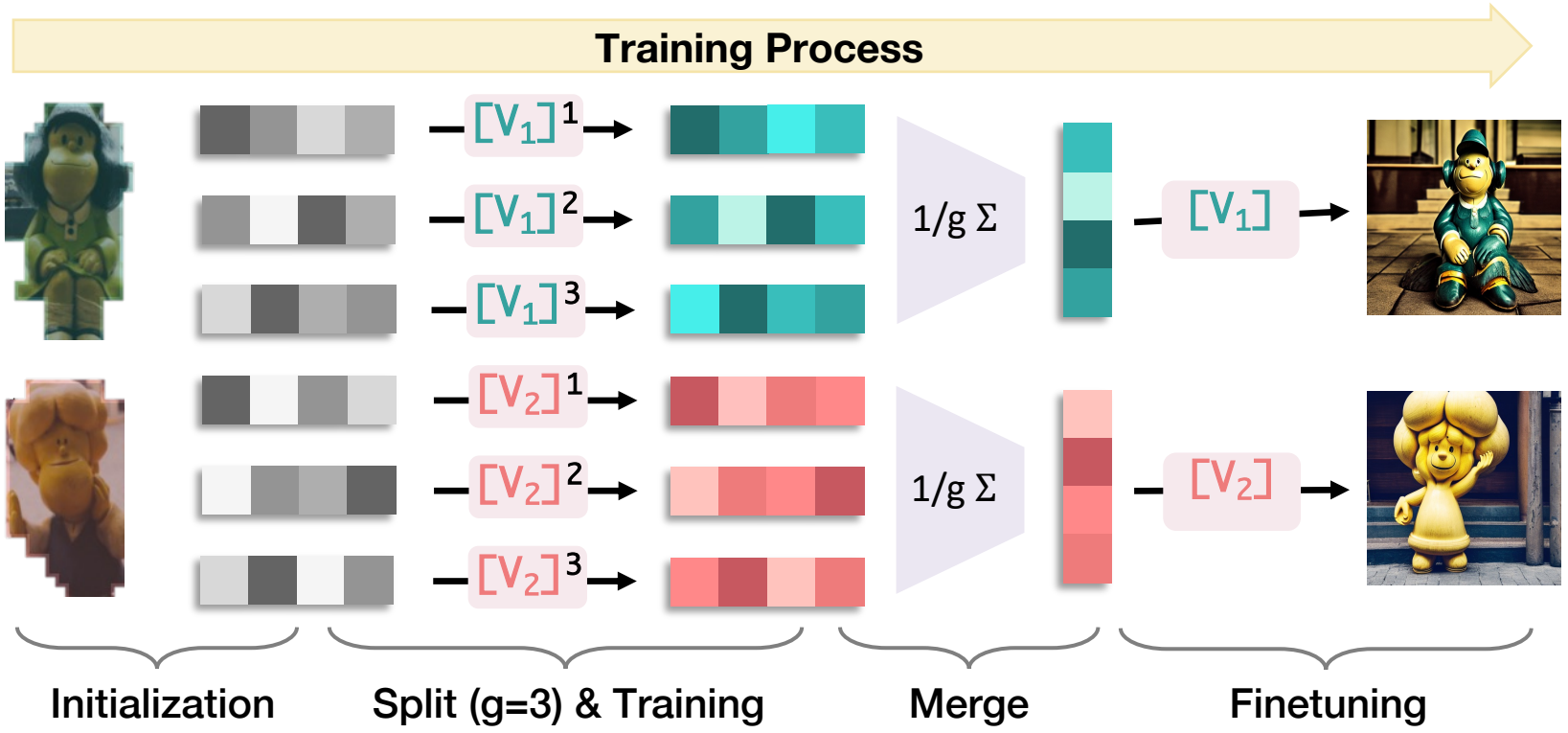


Source image | K-means | FINCH | **Ours**

ConceptExpress - Concept learning

We are not accessible to initial words for concepts!

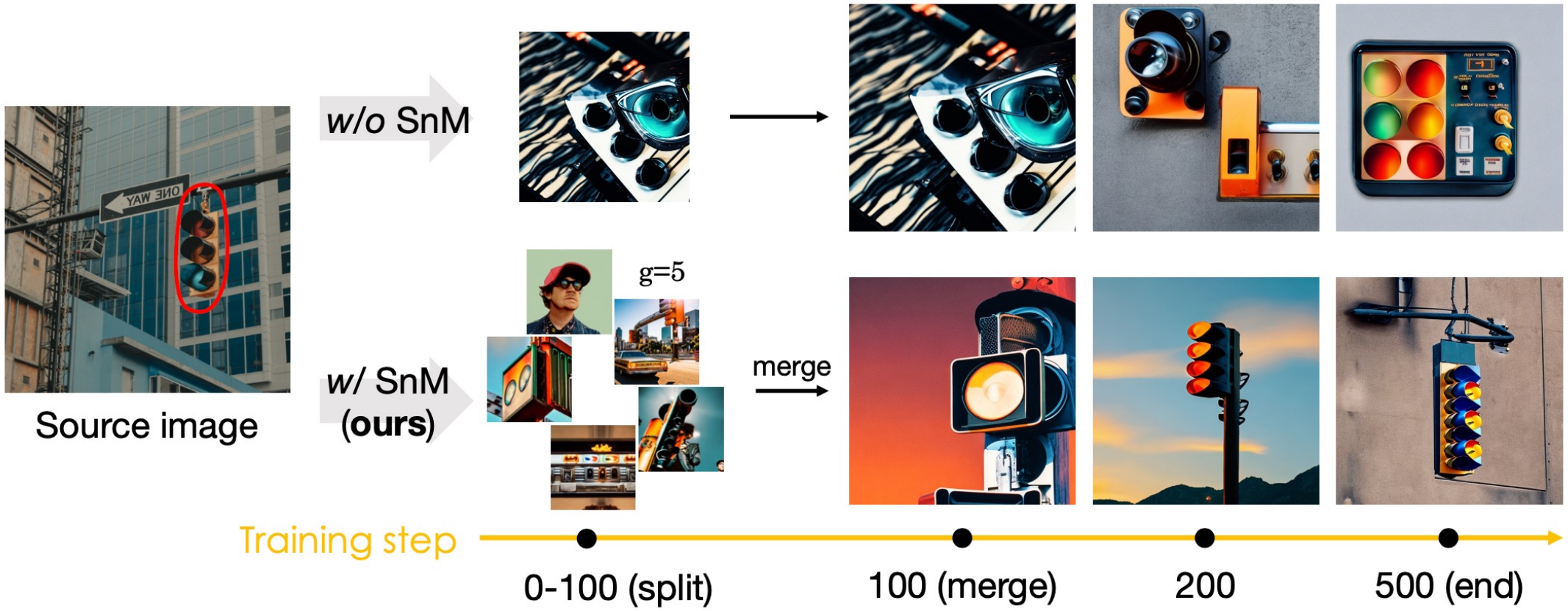
- **Split-and-merge strategy**



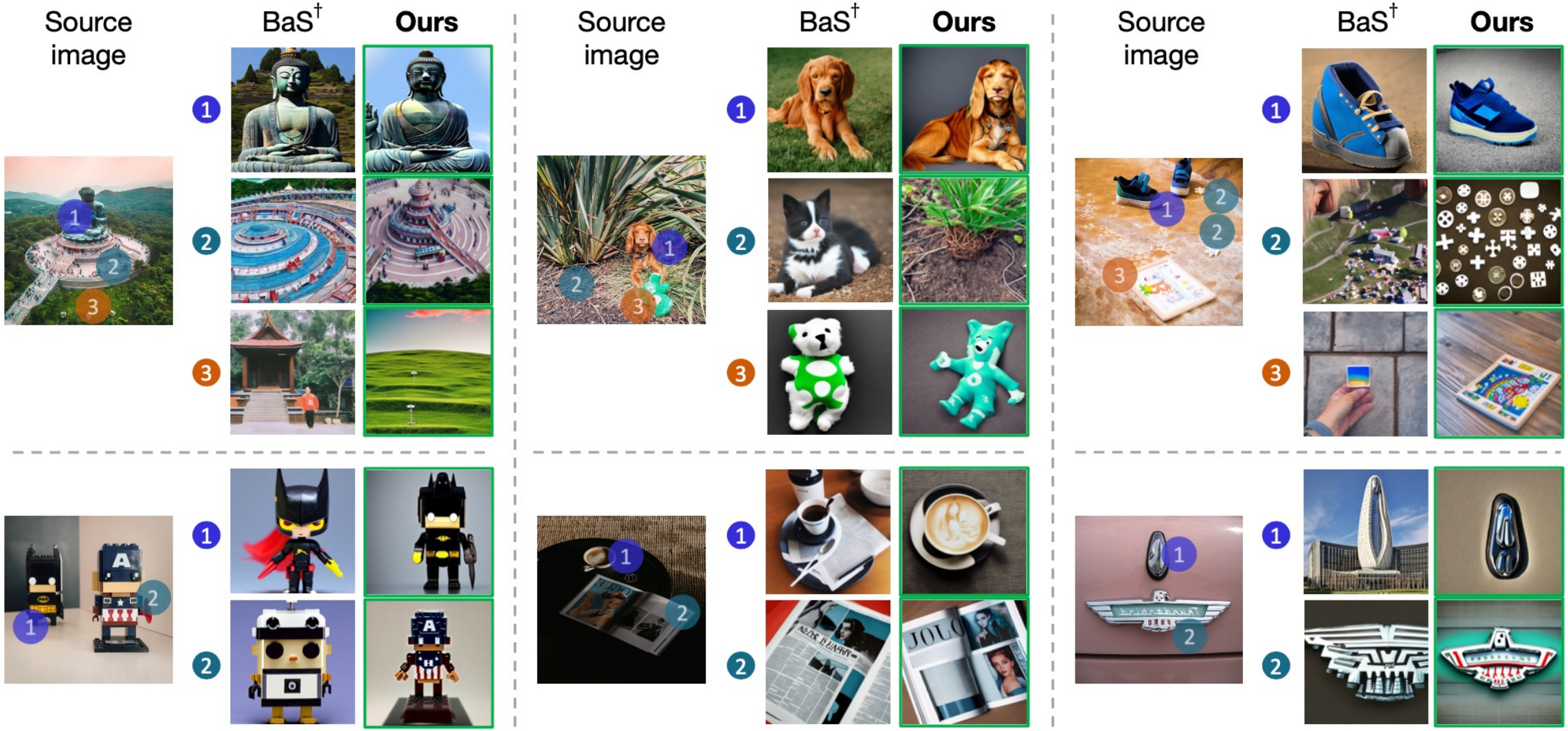
$$\mathcal{L}_i = \mathbb{E}_{z\sim\mathcal{E}(\mathcal{I}),y_i,\epsilon,t}\left[\left\|\left[\epsilon - \epsilon_\theta\left(z_t, t, c_{v_i}(y_i)\right)\right]\odot \mathbf{m}_i\right\|_2^2\right]$$

ConceptExpress - Concept learning

w/ SnM *v.s* w/o SnM



Qualitative results



BaS† denotes unsupervised version of Break-A-Scene [Gal et al. 2022].

Quantitative Evaluation

- Data
  - D1: 96 images self-collected from Unsplash.
  - D2: 7 images provided by Break-A-Scene [Gal et al. 2022].
- Metric
  - **Concept similarity**: identity similarity (SIM^I) & compositional similarity (SIM^C)
  - **Classification accuracy**: top-1 (ACC^1) & top-3 (ACC^3)

(a) Evaluation using CLIP [49].

	D_1				D_2			
Method	SIM^I	SIM^C	ACC^1	ACC^3	SIM^I	SIM^C	ACC^1	ACC^3
BaS [2]	–	–	–	–	0.686	0.696	0.467	0.599
BaS f.t. [2]	–	–	–	–	0.693	0.789	0.526	0.697
BaS† [2]	0.627	0.773	0.174	0.282	0.613	0.653	0.368	0.487
Ours	**0.689**	**0.784**	**0.263**	**0.385**	**0.715**	**0.737**	**0.566**	**0.783**

(b) Evaluation using DINO [10].

	D_1				D_2			
Method	SIM^I	SIM^C	ACC^1	ACC^3	SIM^I	SIM^C	ACC^1	ACC^3
BaS [2]	–	–	–	–	0.316	0.474	0.559	0.704
BaS f.t. [2]	–	–	–	–	0.411	0.696	0.697	0.737
BaS† [2]	0.254	0.510	0.202	0.315	0.231	0.417	0.329	0.559
Ours	**0.319**	**0.568**	**0.324**	**0.470**	**0.371**	**0.535**	**0.803**	**0.934**

Text-prompted generation

Source image

[V_i]

[V_i] in the jungle

[V_i] in the snow

[V_i] with a sunset

[V_1] and [V_2] with a wheat field

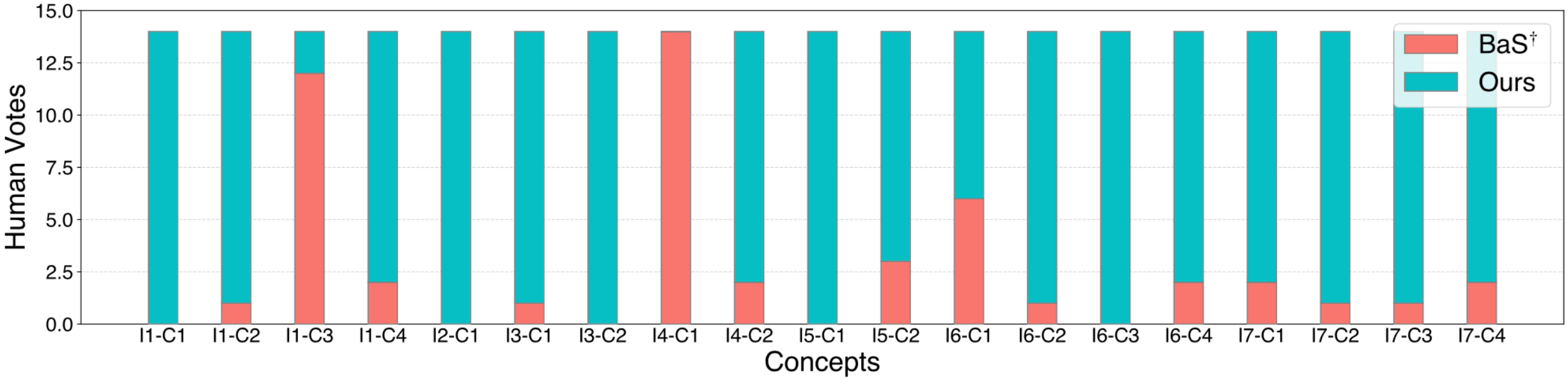
[V_1] and [V_2] among skyscrapers

[V_1] and [V_2] in a movie theater

[V_1] and [V_2] floating on top of water

[V_1] and [V_2] on a cobblestone street

User Study



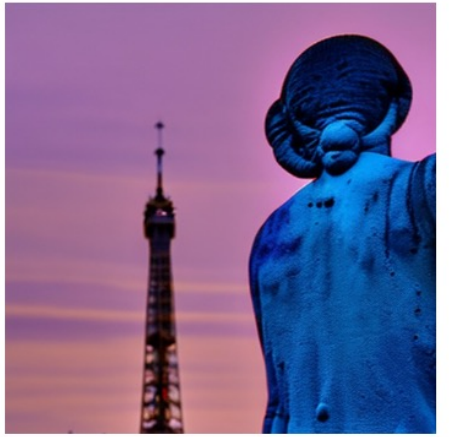
- 14 users; 7 source images; 19 concepts; 266 votes

More results

[V1]

[V1] floating on top of water

[V1] with a city in the background



[V1] and [V2]

[V2]

[V2] with a beautiful sunset

[V2] on top of pink fabric

[V1] and [V2] with a wheat field in the background

More results

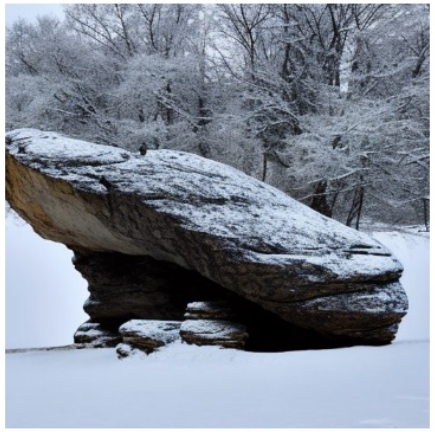
[V1]

[V1] floating on top of water

[V1] among the skyscrapers in New York city

[V1] and [V2]

[V2]



[V2] in the snow

[V2] with a tree and autumn leaves in the background

[V1] and [V2] with a beautiful sunset

[V1]

[V1] with sunflowers around it

[V1] on top of a dirt road

[V1] and [V3] with a beautiful sunset

[V1] and [V2] and [V3] in a luxurious living room

[V2]

[V2] with the Eiffel Tower in the background

[V2] in the snow

[V1] and [V2] with a mountain in the background

[V1] and [V2] and [V3] in the snow

[V3]

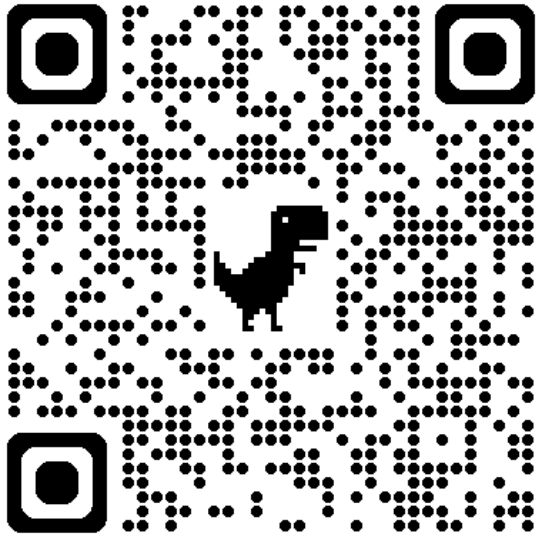
[V3] in the jungle

[V3] at the beach

[V2] and [V3] with a city in the background

[V1] and [V2] and [V3] on a cobblestone street

Thank You for Listening!



Scan QR code to try our model!

Drop by for a chat at:

Session 6 #217
16:30-18:30