❄ICE: Intrinsic Concept Extraction from a Single Image via Diffusion Models

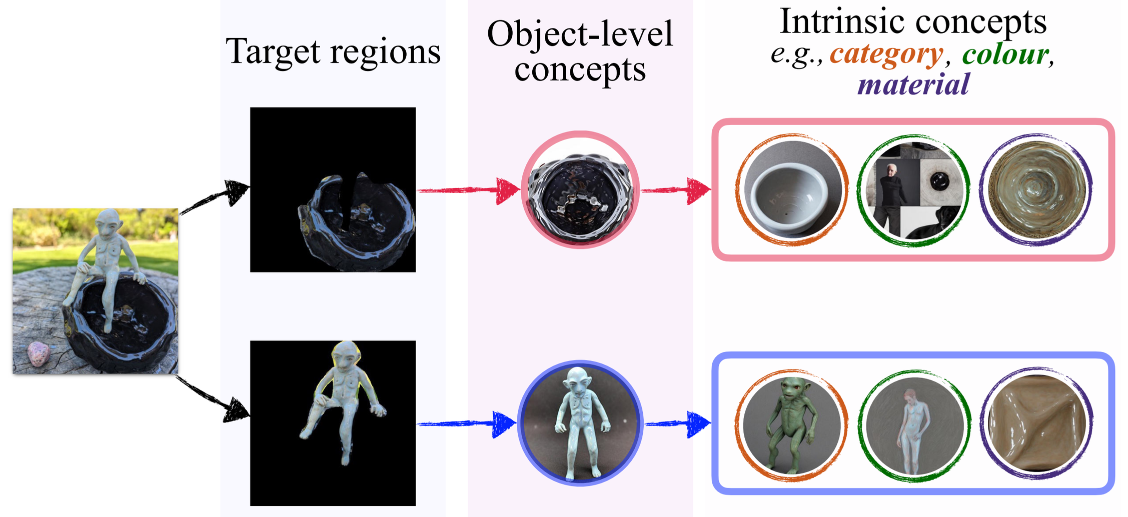
香 港 大 學
THE UNIVERSITY OF HONG KONG

Fernando Julio Cendra   Kai Han
Visual AI Lab, The University of Hong Kong

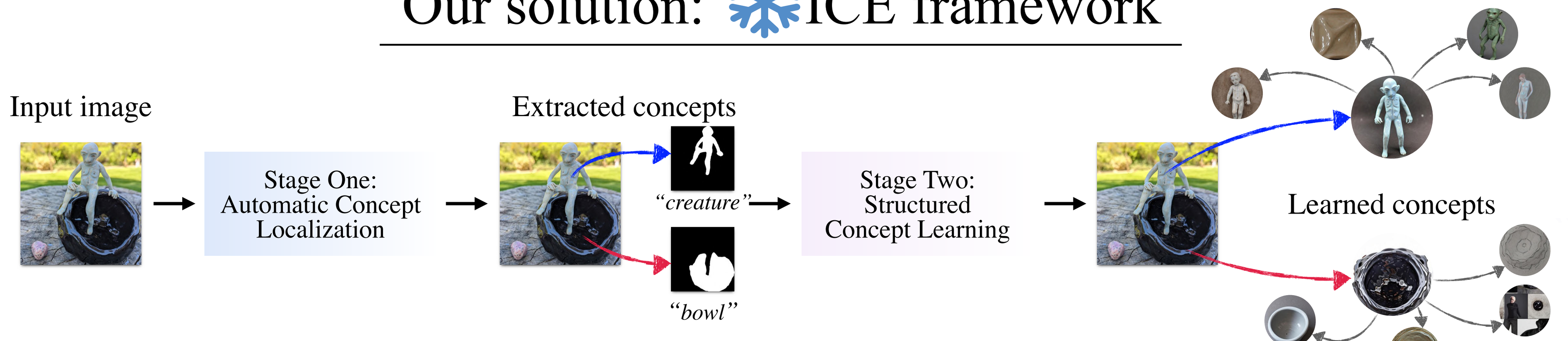
CVPR Nashville JUNE 11-15, 2025

Intrinsic Concept Extraction (ICE)

Intrinsic Concept Extraction aims at extracting **object-level concepts** and the underlying **intrinsic attributes** such as ***semantic category***, ***colour***, and ***material***. Intrinsic Concept Extraction provides a detailed and interpretable representation of visual elements, enabling a structured and comprehensive understanding of the image's components, allowing for versatile downstream generative applications.



Our solution: ❄ICE framework



Our proposed framework, **ICE**, offers a unified and structured approach to automatically and systematically discover intrinsic concepts within an image using a single T2I model.

The proposed ICE framework operates through a two-stage architecture:

- **Stage One:** Automatic Concept Localization, a training-free method utilizing a single T2I Diffusion Model.
- **Stage Two:** Structured Concept Learning, which learns concepts by finetuning the T2I Diffusion Model.

Comparison & positioning of ❄ICE

Comparison of ICE and relevant works.

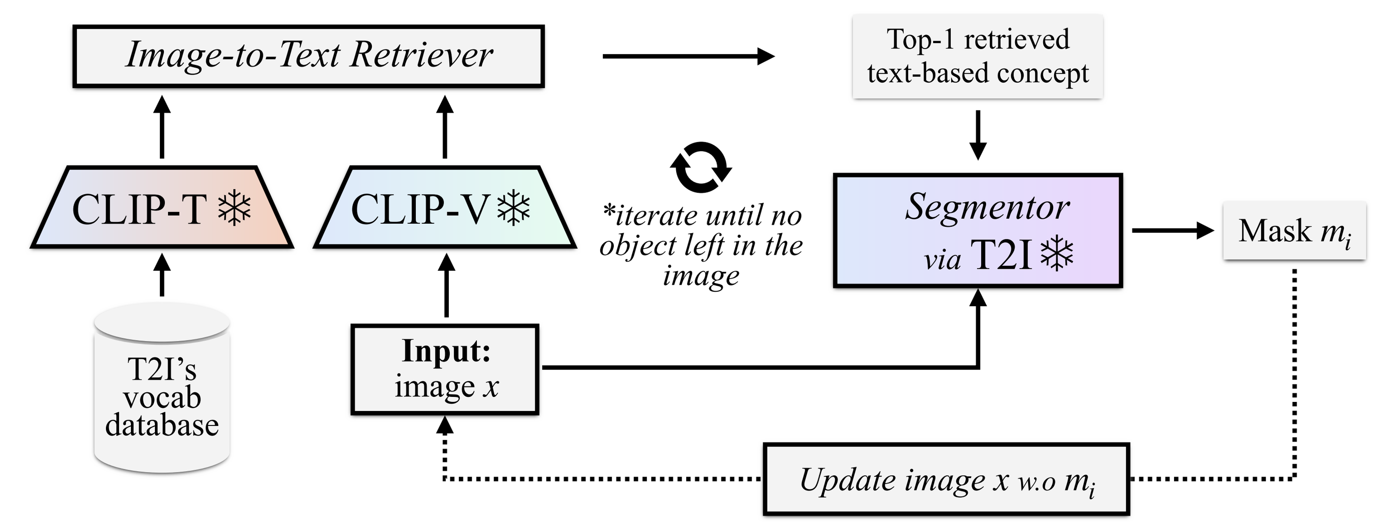
Method	Learned concept(s): Object-level concepts	Learned concept(s): Intrinsic concepts	Learned concept(s): Multi concepts	Framework details: Single image	Framework details: Extra information
Textual Inversion [9]	✔	✗	✗	✗	-
Dreambooth [28]	✔	✗	✗	✗	-
Inspiration Tree [34]	✗	✔	✗	✗	-
LangInt [17]	✔	✔	✗	✔	VQA-guided
Break-A-Scene [1]	✔	✗	✔	✔	Mask
MCPL [15]	✔	✗	✔	✔	Text
ConceptExpress [11]	✔	✗	✔	✔	-
ICE (Ours)	✔	✔	✔	✔	-

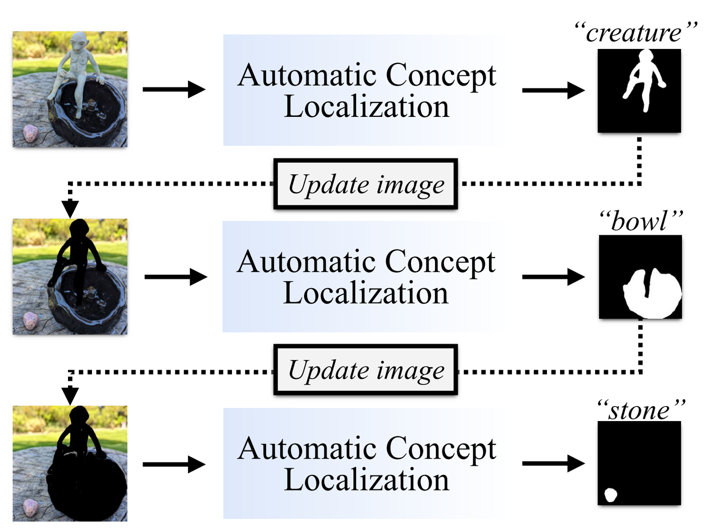
- **Unified framework:** ICE automatically discovers visual concepts using a single T2I model in an unsupervised manner.
- **Structured concept learning:** Deconstructing an image by identifying object-level concepts and further breaks down them into intrinsic attributes such as colour and material.

This solution allows for a more granular and interpretable breakdown of visual elements which leads to better personalized image generation and concept extraction.

❄ICE framework

Stage One: Automatic Concept Localization

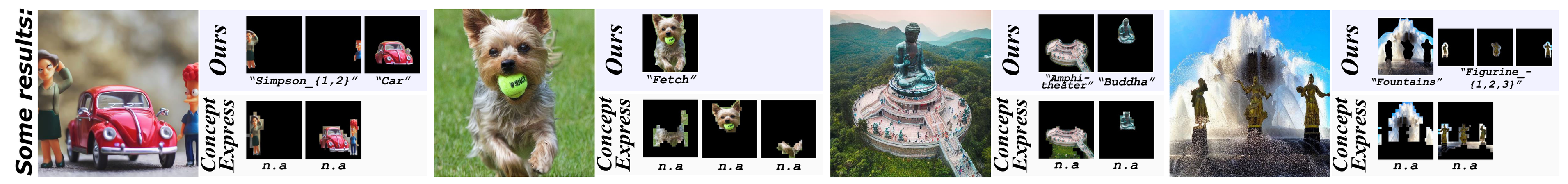




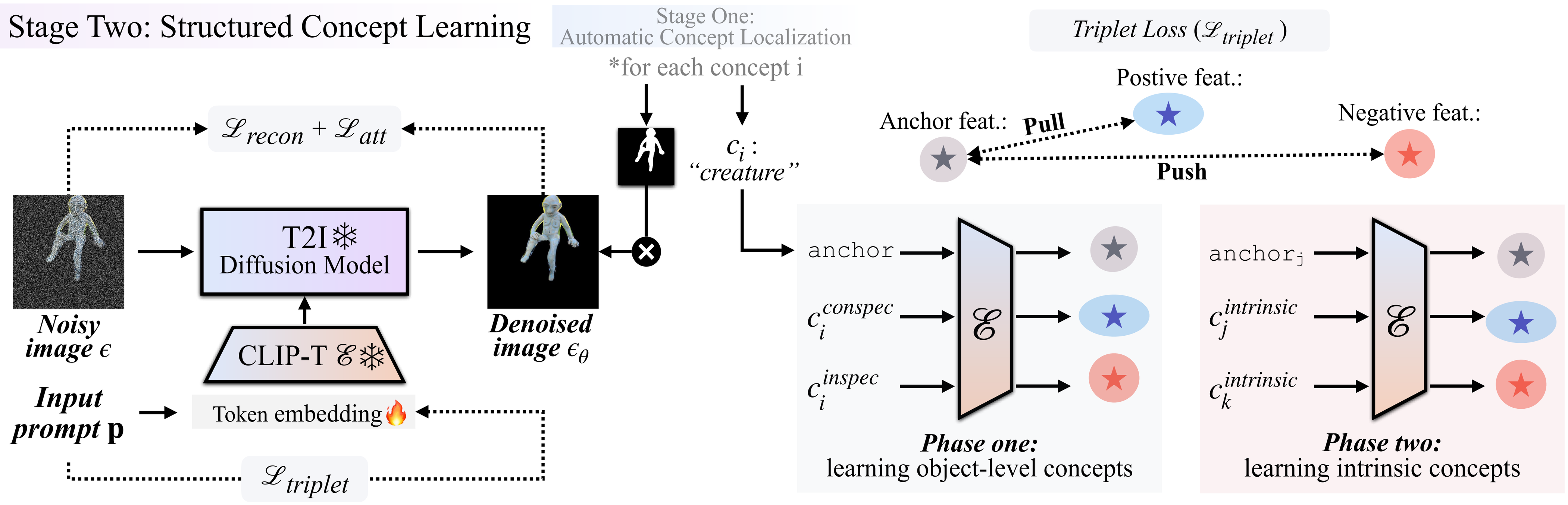
This stage is designed to extract object-level concepts from an unlabelled input image automatically.

Step 1: Retrive top-1 associated text based on input image → Step 2: Extract mask given top-1 text and input image *via* T2I model → Step 3: Update input image by removing the masked region → Step 4: Repeat until no "object" regions are left in the image

The module outputs a **set of text-based concepts** along with their **corresponding masks** for a given image.



Stage Two: Structured Concept Learning



This stage focuses on decomposing the extracted object-level concepts into the underlying intrinsic concepts, such as colour and material. To ensure an accurate decomposition of the concepts, we divide this stage into two phases: (1) learning object-level concepts and (2) learning intrinsic concepts.

For ***Phase one***, the framework learns both concept-specific and instance-specific tokens for each extracted object-level concept while the ***Phase two*** delves into decomposing the learned object-level concepts into intrinsic attributes.

Ablation study

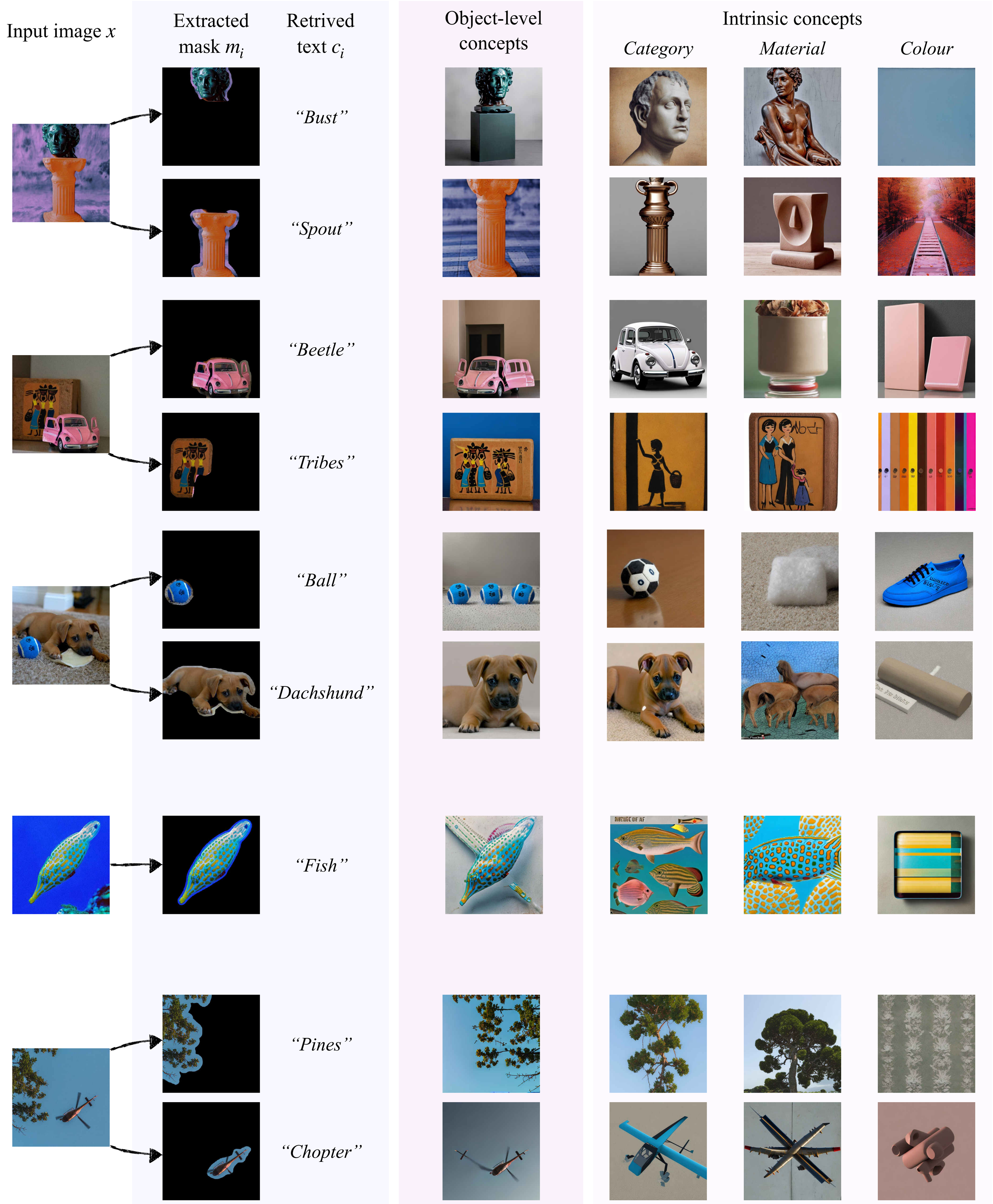
Ablation study of ICE model components.

Method	Stage One's mask	Stage Two's token init.	Stage Two learning	CLIP [25]: SIM^I (%)	CLIP [25]: SIM^C (%)	CLIP [25]: ACC^1 (%)	CLIP [25]: ACC^3 (%)	DINO [7]: SIM^I (%)	DINO [7]: SIM^C (%)	DINO [7]: ACC^1 (%)	DINO [7]: ACC^3 (%)
ConceptExpress	✗	✗	✗	0.689	0.784	0.263	0.385	0.319	0.568	0.324	0.470
ICE *w.* mask	✔	✗	✗	0.710	0.781	0.307	0.456	0.493	0.601	0.395	0.604
ICE *w.o* Stage Two	✔	✔	✗	0.726	0.807	0.301	0.452	0.501	0.621	0.411	0.604
ICE *w.o* text init.	✔	✗	✔	0.722	0.814	0.320	0.475	0.548	0.643	0.449	0.627
ICE (Ours)	✔	✔	✔	**0.738**	**0.822**	**0.325**	**0.518**	**0.677**	**0.755**	**0.476**	**0.638**

Each component significantly enhances the framework.

The complete ICE framework outperforms all it's variants.

Qualitative results



Quantitative results

Performance of ICE and relevant works on unsupervised Concept Extraction benchmarks using CLIP encoder.

Method	SIM^I (%)	SIM^C (%)	ACC^1 (%)	ACC^3 (%)
Break-A-Scene [1]	0.627	0.773	0.174	0.282
ConceptExpress[11]	0.689	0.784	0.263	0.385
ICE (Ours)	**0.738**	**0.822**	**0.325**	**0.518**

Performance of ICE and relevant works on unsupervised Concept Extraction benchmarks using DINO encoder.

Method	SIM^I (%)	SIM^C (%)	ACC^1 (%)	ACC^3 (%)
Break-A-Scene [1]	0.254	0.510	0.202	0.315
ConceptExpress [11]	0.319	0.568	0.324	0.470
ICE (Ours)	**0.677**	**0.755**	**0.476**	**0.638**