

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

## – *Supplementary Material* –

Fernando Julio Cendra      Kai Han<sup>†</sup>

Visual AI Lab, The University of Hong Kong

fcendra@connect.hku.hk, kaihax@hku.hk

**Overview.** In this supplementary material, we first present additional quantitative results in Section [S1](#) to further validate the efficacy of our ICE framework. We then showcase more qualitative results in Section [S2](#). Furthermore, beyond the intrinsic concept extraction task, we explore two additional practical applications of our framework: compositional concept generation in Section [S3.1](#) and zero-shot unsupervised segmentation in Section [S3.2](#). Next, we present the pseudocode for ICE’s Stage One: Automatic Concept Localization in Section [S4](#). Section [S5](#) provides the details of the triplet margin  $\gamma$  used in Stage Two: Structured Concept Learning. Section [S6](#) provides the list of prompt templates used during our Stage Two training. Finally, we discuss our work’s broader impacts and limitations in Section [S7](#).

### Contents

<a href="#">S1. More Quantitative Results</a>	2
<a href="#">S2. More Qualitative Results</a>	4
<a href="#">S3. Applications</a>	5
<a href="#">S4. Pseudocode for ICE Stage One: Automatic Concept Localization</a>	7
<a href="#">S5. ICE Stage Two’s choice of triplet margin <math>\gamma</math>.</a>	8
<a href="#">S6. Prompt Templates</a>	9
<a href="#">S7. Broader Impacts and Limitation</a>	10

---

<sup>†</sup>Corresponding author.

## S1. More Quantitative Results

In this section, we provide additional quantitative evaluations to further assess the performance of our ICE framework. The evaluations are divided into two subsections:

- **Intrinsic Concept benchmark**, Section S1.1.
- **Additional results on UCE benchmarks**, Section S1.2.

### S1.1. Intrinsic Concept benchmark

Since our proposed ICE framework introduces a novel approach to extracting intrinsic concepts, we design a comprehensive evaluation method to assess the quality of these intrinsic concepts. We call this Intrinsic Concept benchmark (ICBench). To achieve this, we expand upon the dataset utilised by the Unsupervised Concept Extraction (UCE) benchmarks [4], referred to as the *DI* dataset (96 images), which is based on Unsplash<sup>1</sup> images. For each extracted concept mask in every image, we provide detailed concept descriptions, denoted as  $description_j$ , encompassing  $intrinsic_j$  attributes such as object, material, and colour. These descriptions, as illustrated in Figure A, are generated using GPT-4o model [5] to ensure accurate and reliable descriptions.

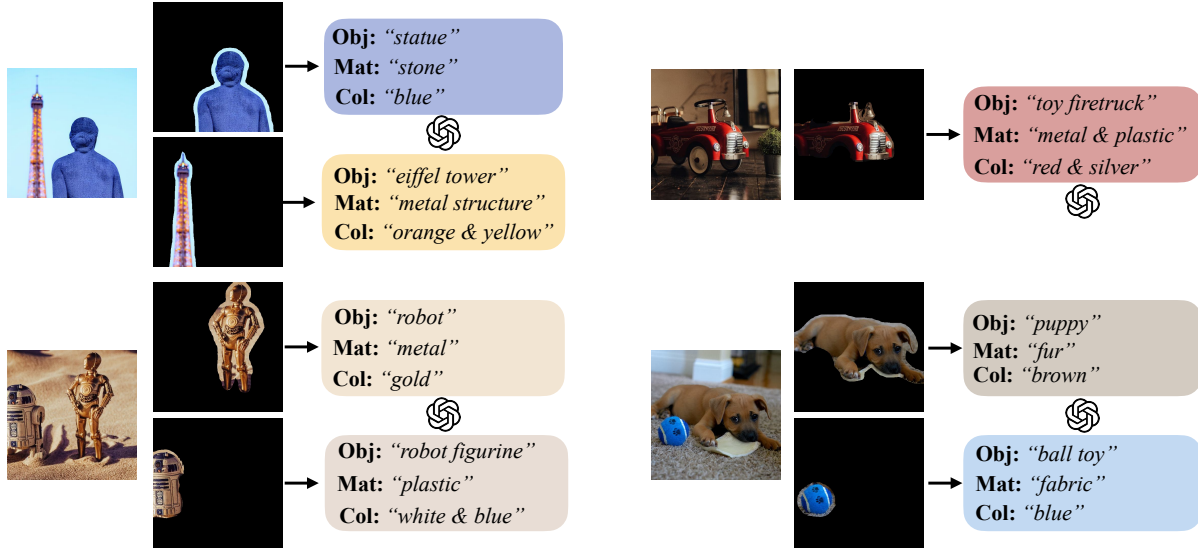


Figure A. Representative intrinsic concept evaluation descriptions generated by GPT-4o model for each segmentation mask produced by the ICE framework. Each description includes detailed attributes such as object (**Obj**), material (**Mat**), and colour (**Col**).

We employ two primary metrics to evaluate the quality of the extracted intrinsic concepts:

- $SIM_{intrinsic_j}^{T-T}$ : This metric evaluates the similarity between the GPT’s described concept  $description_j$  and the learned token  $c_j^{intrinsic}$ .
- $SIM_{intrinsic_j}^{T-V}$ : This metric assesses the similarity between the GPT’s described concept  $description_j$  and corresponding visual representations. Specifically, we generate 8 images based on learned token  $c_j^{intrinsic}$  and evaluate their similarity to the described concepts.

Given that ConceptExpress learns only a single token per concept, to facilitate a fair comparison, we adapt our evaluation by modifying ConceptExpress’s learned asset (from ICE’s Stage One mask) to:

**Prompt:** “a  $intrinsic_j$  of asset”

Table A presents the results of our ICBench. The results demonstrate that ICE achieves higher similarity scores across both metrics, indicating that our learnt tokens are more diverse and richer compared to those generated by ConceptExpress w/ ICE

<sup>1</sup><https://unsplash.com/>

mask. The higher  $\text{SIM}_{intrinsic_j}^{T-T}$  and  $\text{SIM}_{intrinsic_j}^{T-V}$  scores achieved by ICE indicate that our framework effectively captures and represents intrinsic attributes more accurately and comprehensively. This superiority is attributed to our two-stage framework which enables the extraction of meaningful intrinsic concepts.

Table A. Performance of ICE and relevant works on the Intrinsic Concept benchmark (ICBench) using CLIP [6] encoders.

Method	$\text{SIM}_{object}^{T-T}$	$\text{SIM}_{material}^{T-T}$	$\text{SIM}_{colour}^{T-T}$	$\text{SIM}_{object}^{T-V}$	$\text{SIM}_{material}^{T-V}$	$\text{SIM}_{colour}^{T-V}$
ConceptExpress w/ ICE mask	0.096	0.071	0.067	0.136	0.107	0.110
<b>ICE (Ours)</b>	<b>0.249</b>	<b>0.101</b>	<b>0.093</b>	<b>0.264</b>	<b>0.208</b>	<b>0.215</b>

## S1.2. Additional results on UCE benchmarks

In the main paper, we evaluated our ICE framework on UCE benchmarks [4] using the UCE’s *D1* dataset, which comprises 96 images sourced from Unsplash. To further validate our framework’s robustness and generalisability, we evaluate our approach with an additional dataset, referred to as the *D2* dataset, which consists of 7 diverse images. In this experiments, we use the masks localized by our framework for all methods.

Table B. Performance of ICE and relevant works on the UCE benchmarks using different encoders on the *D2* dataset.

(a) Using CLIP [6] encoder on the <i>D2</i> dataset.					(b) Using DINO [2] encoder on the <i>D2</i> dataset.				
Method	$\text{SIM}^I$	$\text{SIM}^C$	$\text{ACC}^1$	$\text{ACC}^3$	Method	$\text{SIM}^I$	$\text{SIM}^C$	$\text{ACC}^1$	$\text{ACC}^3$
ConceptExpress w/ ICE mask	0.701	0.788	0.725	0.894	ConceptExpress w/ ICE mask	0.512	0.599	0.775	0.925
<b>ICE (Ours)</b>	<b>0.713</b>	<b>0.820</b>	<b>0.781</b>	<b>0.937</b>	<b>ICE (Ours)</b>	<b>0.538</b>	<b>0.607</b>	<b>0.775</b>	<b>0.944</b>

Tables Ba and Bb present the quantitative results of ICE compared to previous approaches on the *D2* dataset. The results consistently show that ICE outperforms existing methods, underscoring the effectiveness of our two-stage framework in learning and extracting concepts. The quantitative evaluations on both CLIP [6] and DINO [2] embeddings demonstrate that ICE consistently achieves higher  $\text{SIM}^I$ ,  $\text{SIM}^C$ ,  $\text{ACC}^{\{1,3\}}$  compared to ConceptExpress w/ ICE mask. This consistent performance across both datasets and embedding models highlights the superior capability of ICE in accurately extracting and representing object-level concepts. The enhanced alignment and deeper understanding fostered by ICE’s two-stage approach significantly contribute to its improved performance, making it a more effective framework for unsupervised concept extraction task.

## S2. More Qualitative Results

In this section, we present additional qualitative results generated by our ICE framework to demonstrate its effectiveness in accurately localising and decomposing visual concepts. Figure B showcases a variety of unlabelled images along with the corresponding extracted concepts and their segmentation masks. These examples illustrate the framework’s capability to identify and segment distinct objects within an image, as well as decompose them into their intrinsic attributes such as object, colour, and material. The results highlight the precision and interpretability of the concept representations produced by ICE, reinforcing the advantages of our structured approach to learning concepts.


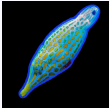


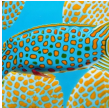

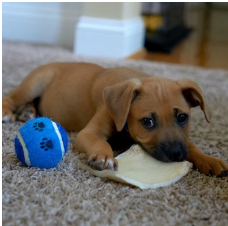
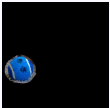








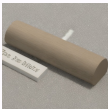

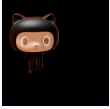



















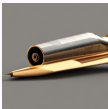






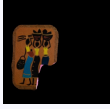

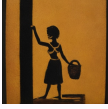


Input image $x$	Extracted mask $m_i$	Retrieved text $c_i$	Object-level concept	Intrinsic concepts		
				Object	Material	Colour
		"Fish"				
		"Ball"				
		"Dachshund"				
		"Funko"				
		"Sprout"				
		"Droid"				
		"Oro"				
		"Beetle"				
		"Tribes"				

Figure B. Additional qualitative results obtained by our ICE framework. Each row displays the original input image, the extracted relevant text-based concepts with their segmentation masks, and the decomposed intrinsic attributes.

### S3. Applications

Here we demonstrate two practical applications of our ICE framework: compositional concept generation in Section S3.1 and unsupervised segmentation in Section S3.2.

#### S3.1. Compositional concept generation

While existing visual concept learning approaches predominantly focus on object-level concepts, enabling the transfer of entire objects to generate personalised image generations, they often lack the ability to selectively manipulate specific intrinsic attributes such as material or colour. This limitation arises from the inherent entanglement of these attributes within the object-level representations, making targeted edits challenging and inefficient.

Our ICE framework addresses this limitation by decomposing object-level concepts into their intrinsic attributes through Stage Two: Structured Concept Learning. This decomposition facilitates the disentanglement of attributes like object, material, and colour, allowing for flexible manipulation of each attribute without altering the overall object identity. Figure C illustrates an example where ICE enables compositional generation of an object’s colour and material independently. By using ICE framework, we can specifically edit or alter the material while keeping the colour unchanged. This granular control over intrinsic attributes empowers users to perform precise and targeted compositional concept edits, enhancing the flexibility and utility of image generation tasks. By enabling the independent manipulation of intrinsic attributes, ICE not only enhances the customization capabilities of image generation but also paves the way for more sophisticated and user-driven image generation applications. This advancement overcomes the limitations of current object-level concept learning methods, providing a more nuanced and flexible approach to text-to-image generation.

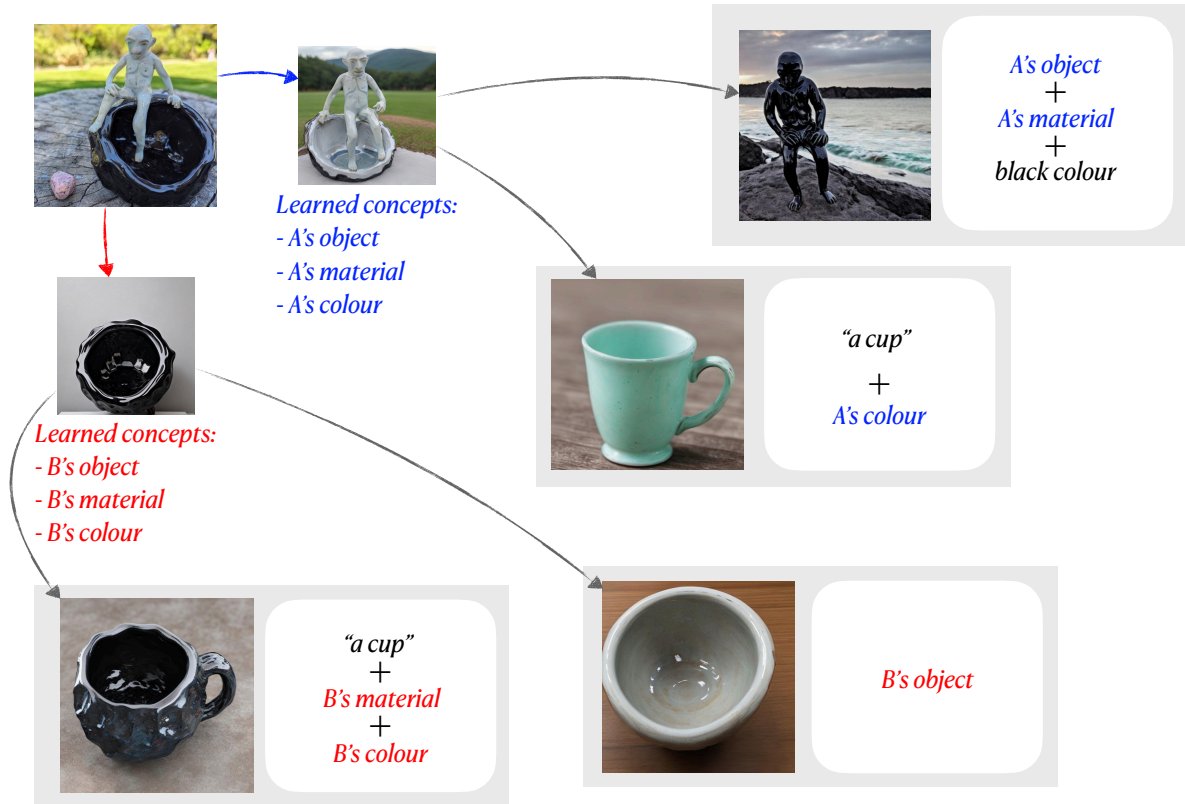


Figure C. Compositional concept generation using ICE. The first column shows the original image and identified object with intrinsic attributes. The subsequent columns demonstrate the compositional generation of individual attributes such as material and colour.

### S3.2. Zero-Shot unsupervised segmentation

Our ICE framework is divided into two stages: automatic concept localisation (Stage One) and structured concept learning (Stage Two). In this section, we showcase the potential of our automatic concept localisation module for zero-shot unsupervised segmentation. Specifically, our Stage One module utilises DiffSeg [7] as the zero-shot segmentor  $\mathcal{S}$ , enabling segmentation without any additional training or reliance on language dependencies. DiffSeg leverages a pre-trained Text-to-Image (T2I) Stable Diffusion model to perform zero-shot segmentation tasks. Specifically, it employs an iterative merging process based on measuring the distribution among self-attention maps extracted from the T2I model. We compare Stage One of our method with DiffSeg and some traditional clustering techniques such as  $k$ -means-S [7], which utilises  $k$ -means clustering with a predefined number of clusters based on the ground truth classes, and the DBSCAN [3] algorithm, known for its density-based clustering approach. These comparisons provide a comprehensive evaluation of our ICE framework’s performance in unsupervised segmentation tasks.

Table C. Zero-shot unsupervised segmentation performance comparison between ICE’s Stage One: Automatic Concept Localization with other methods on the subset of COCO-stuff [1] called COCO-stuff-27 dataset.

Method	ACC. (%)	mIoU (%)
$k$ -means-S [7]	62.6	34.7
DBSCAN	57.7	27.2
DiffSeg [7]	<b>72.5</b>	<b>43.6</b>
<b>ICE’s Stage One (Ours)</b>	<u>69.5</u>	<u>39.4</u>

As demonstrated in Table C, our ICE framework performs competitively compared to DiffSeg, although it slightly underperforms in metrics such as Accuracy (ACC) and Intersection over Union (IoU). However, it is noteworthy that the ICE’s Stage One module not only performs segmentation but also retrieves relevant text-based concepts, which are crucial for downstream tasks in concept learning. This dual functionality is pivotal for concept learning tasks. By obtaining relevant text-based concepts, ICE enables the initialization of each learnable token with its corresponding text description. This leads to more effective and meaningful token embedding initialization, enhancing the overall quality of the learned concepts. In contrast, replacing ICE’s Stage One module with DiffSeg would provide segmentation results to the lack of the text-based concept retrieval necessary for effective concept learning.



#### S4. Pseudocode for ICE Stage One: Automatic Concept Localization

This stage is designed to automatically extract object-level concepts from an unlabelled input image. It leverages off-the-shelf modules integrated within the T2I diffusion model, ensuring a training-free and seamless concept extraction process. The workflow of Stage One: Automatic Concept Localization is illustrated in Figure 4 of the main paper. The process begins with retrieving the most relevant text-based concept using the *Image-to-Text Retriever* ( $\mathcal{T}$ ). For the retrieved concept, the *Segmentor* ( $\mathcal{S}$ ) generates a corresponding segmentation mask, delineating the region of the image associated with that concept. The identified object is then masked out from the image, and the process iterates until the proportion of unmasked pixels falls below a predefined threshold. The process is summarized in Algorithm A.

---

**Algorithm A** ICE’s Stage One: Automatic Concept Localization

---

**Require:** Input image  $\mathbf{x}$   
**Require:** *Image-to-Text Retriever*  $\mathcal{T}$   
**Require:** *Segmentor*  $\mathcal{S}$   
**Require:** Pixel proportion threshold  $\tau$  (e.g., 5%)

- 1: Initialize  $\mathcal{C} \leftarrow \{\}, \mathcal{M} \leftarrow \{\}$
- 2: Initialize remaining pixel proportion  $\rho \leftarrow 100\%$
- 3: **while**  $\rho > \tau$  **do**
- 4:      $c_i \leftarrow \mathcal{T}(\mathbf{x})[\text{top-1}]$                      *// Retrieve top-1 text-based concept from image  $\mathbf{x}$*
- 5:      $\mathbf{m}_i \leftarrow \mathcal{S}(\mathbf{x}, c_i)$                      *// Generate segmentation mask for concept  $c_i$*
- 6:      $\mathcal{C} \leftarrow \mathcal{C} \cup \{c_i\}$                      *// Store the retrieved concept*
- 7:      $\mathcal{M} \leftarrow \mathcal{M} \cup \{\mathbf{m}_i\}$                      *// Store the corresponding mask*
- 8:      $\mathbf{x} \leftarrow \mathbf{x} \odot (1 - \mathbf{m}_i)$                      *// Mask out the identified object from the image*
- 9:      $\rho \leftarrow \text{PixelProportion}(\mathbf{x})$              *// Update the remaining pixel proportion*
- 10: **return**  $\mathcal{C}, \mathcal{M}$

---

The function  $\text{PixelProportion}(\mathbf{x})$  computes the percentage of unmasked pixels in the current state of the image  $\mathbf{x}$ . It is defined as:

$$\text{PixelProportion}(\mathbf{x}) = \frac{\text{No. of unmasked pixels in } \mathbf{x}}{\text{Total no. of pixels in } \mathbf{x}}. \quad (\text{a})$$

## S5. ICE Stage Two’s choice of triplet margin $\gamma$ .

In this section, we provide details of the triplet margin  $\gamma$  parameter utilized in Stage Two: Structured Concept Learning, encompassing both *Phase one* and *Phase two* learning.

**Phase one margin  $\gamma$ :** Here, we make use of margin  $\gamma$  to regulate the separation between concept-specific  $c_i^{\text{conspec}}$  and instance-specific  $c_i^{\text{inspec}}$  tokens. As shown in Figure D, we show three experiments with  $\gamma$  values of low (0.001), medium (0.05), and high (1.0). Results showed that a large margin  $\gamma = 1.0$  negatively impacted the model by pushing both  $c_i^{\text{conspec}}$  and  $c_i^{\text{intrinsic}}$  outside their optimal distributions, leading to poorer concept alignment and reconstruction quality. Conversely, a very small margin  $\gamma = 0.001$  caused the concept-specific and intrinsic tokens to be too close, making it difficult for the model to distinguish between general and instance-specific attributes, thus hindering accurate reconstruction. Empirically, we find that setting  $\gamma = 0.05$  provides the best balance, ensuring adequate separation between  $c_i^{\text{conspec}}$  and  $c_i^{\text{intrinsic}}$  while maintaining their alignment within the desired distributions.

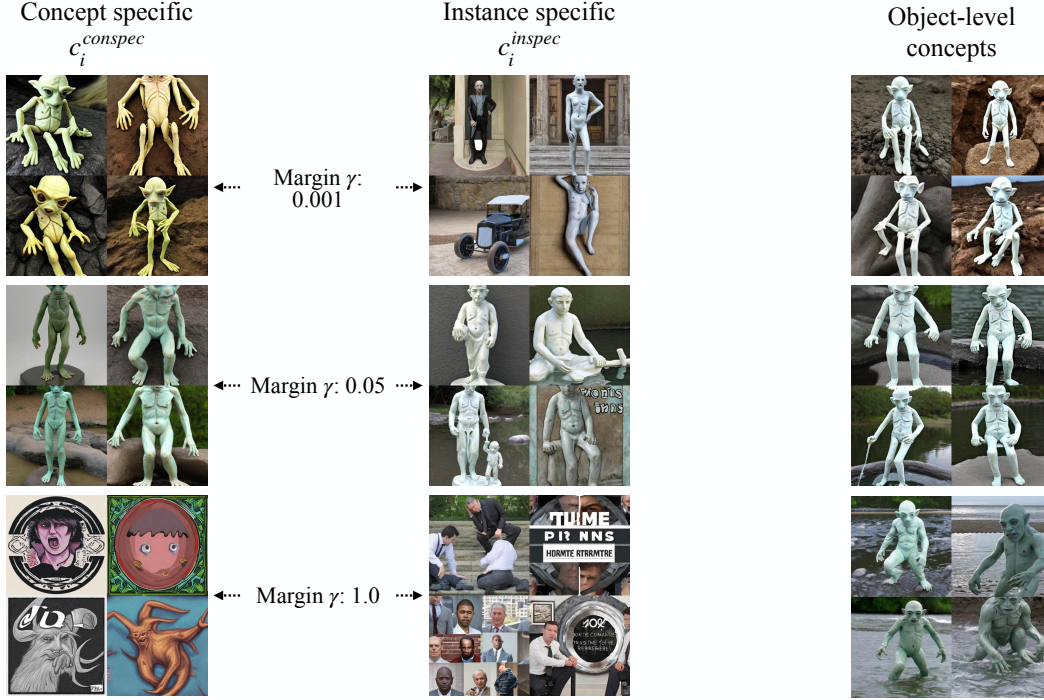


Figure D. Effect of *Phase one* triplet margin  $\gamma$ . We find that a margin  $\gamma$  of 0.05 provides the best balance in separating concept-specific and instance-specific attributes.

**Phase two margin  $\gamma$ :** In addition to the *Phase one* margin, we introduce a *Phase two* margin specifically for intrinsic concept learning. Given  $j$  intrinsic concepts, we calculate the text embedding distances between each pair of intrinsic concepts. For example, we measure the distance between the textual features of the words “colour” and “material”. This *Phase two* margin ensures that intrinsic concepts are well-separated from one another in the embedding space, preventing overlap and enhancing the model’s ability to capture distinct intrinsic properties. By enforcing sufficient distance between different intrinsic concepts, the *Phase two* margin promotes more precise and meaningful intrinsic concept representations.



## S6. Prompt Templates

We provide the complete list of prompt templates utilized during ICE’s Stage Two: Structured Concept Learning. For each training step, a prompt is randomly selected from this list. This random selection process ensures a diverse range of inputs, enhancing the model generalizably and robustness.

```
“a photo of a {}”  
“a rendering of a {}”  
“a cropped photo of the {}”  
“a photo of a {}”  
“a rendering of a {}”  
“a cropped photo of the {}”  
“the photo of a {}”  
“a photo of a clean {}”  
“a photo of a dirty {}”  
“a dark photo of the {}”  
“a photo of my {}”  
“a photo of the cool {}”  
“a close-up photo of a {}”  
“a bright photo of the {}”  
“a cropped photo of a {}”  
“a photo of the {}”  
“a good photo of the {}”  
“a photo of one {}”  
“a close-up photo of the {}”  
“a rendition of the {}”  
“a photo of the clean {}”  
“a rendition of a {}”  
“a photo of a nice {}”  
“a good photo of a {}”  
“a photo of the nice {}”  
“a photo of the small {}”  
“a photo of the weird {}”  
“a photo of the large {}”  
“a photo of a cool {}”  
“a photo of a small {}”
```

## **S7. Broader Impacts and Limitation**

The development of unsupervised compositional concept discovery frameworks like ICE has significant implications for the field of computer vision and beyond. By enabling more granular and interpretable concept learning, ICE enhances the capabilities of generative AI applications in tasks such as image content personalization. These advancements can lead to more intuitive and user-friendly applications, such as advanced image generation tools and more accurate compositional concept generation. However, the use of pre-trained T2I diffusion models in ICE implies reliance on datasets that may contain inherent biases. These biases can be inadvertently perpetuated or even amplified in the generated concepts, leading to unfair or discriminatory outcomes. It is imperative to address these biases through careful dataset curation and model auditing to ensure the equitable and responsible deployment of ICE in real-world applications.

## References

- [1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 2018. [6](#)
- [2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. [3](#)
- [3] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996. [6](#)
- [4] Shaozhe Hao, Kai Han, Zhengyao Lv, Shihao Zhao, and Kwan-Yee K Wong. Conceptexpress: Harnessing diffusion models for single-image unsupervised concept extraction. In *ECCV*, 2024. [2](#), [3](#)
- [5] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. [2](#)
- [6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. [3](#)
- [7] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. In *CVPR*, 2024. [6](#)