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

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

The University of Hong Kong
{szhao,shzhao,kykwong}@cs.hku.hk kaihanx@hku.hk cszy98@gmail.com

Abstract. While personalized text-to-image generation has enabled the learning of a single concept from multiple images, a more practical yet challenging scenario involves learning multiple concepts within a single image. However, existing works tackling this scenario heavily rely on extensive human annotations. In this paper, we introduce a novel task named Unsupervised Concept Extraction (UCE) that considers an unsupervised setting without any human knowledge of the concepts. Given an image that contains multiple concepts, the task aims to extract and recreate individual concepts solely relying on the existing knowledge from pretrained diffusion models. To achieve this, we present ConceptExpress that tackles UCE by unleashing the inherent capabilities of pretrained diffusion models in two aspects. Specifically, a concept localization approach automatically locates and disentangles salient concepts by leveraging spatial correspondence from diffusion self-attention; and based on the lookup association between a concept and a conceptual token, a concept-wise optimization process learns discriminative tokens that represent each individual concept. Finally, we establish an evaluation protocol tailored for the UCE task. Extensive experiments demonstrate that ConceptExpress is a promising solution to the UCE task. Our code and data are available at: https://github.com/haoosz/ConceptExpress

Keywords: Unsupervised concept extraction · Diffusion model

1 Introduction

After observing an image containing multiple concepts, a skilled painter can recreate each individual concept within the complex scene. This remarkable cognitive ability prompts us to raise an intriguing question: *Do text-to-image generative models also possess the capability to extract and recreate concepts?* In this paper, we try to provide an answer to this question by harnessing the potential of Stable Diffusion [54] in concept extraction.

Diffusion models [23, 45, 50, 54, 56, 61] have exhibited unprecedented performance in photorealistic text-to-image generation. Although diffusion models

⋆ Corresponding authors

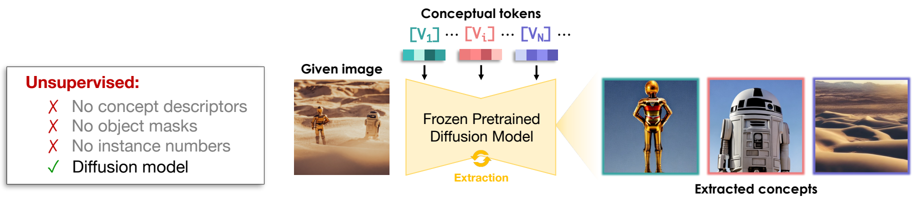

Fig. 1: Unsupervised concept extraction. We focus on the unsupervised problem of extracting multiple concepts from a single image. Given an image that contains multiple concepts (e.g., Star Wars characters C-3PO, R2-D2, and desert), we aim to harness a frozen pretrained diffusion model to automatically learn the conceptual tokens. Using the learned conceptual tokens, we can regenerate the extracted concepts with high quality, as shown in the rightmost column. In this process, no human knowledge or aids are available, and we only rely on the inherent capabilities of the pretrained Stable Diffusion [54].

are trained solely for the purpose of text-to-image generation, extensive evidence suggests their underlying capabilities in various tasks, including classification [37], segmentation [28,44,65,68,75], and semantic correspondence [22,39,80]. This indicates that diffusion models embed significant world knowledge, potentially enabling them to perceive and recreate concepts akin to skilled painters. Motivated by this insight, we delve into this problem and explore the untapped potential of Stable Diffusion [54] in concept extraction. While recent research [2, 26] has made initial attempts in exploring concept extraction using Stable Diffusion, existing approaches heavily rely on external human knowledge for supervision during the learning process. For example, Break-A-Scene [2] demands pre-annotated object masks, while MCPL [26] requires accurate concept-descriptive captions. However, these human aids are both costly and often inaccessible. This critical constraint renders existing approaches infeasible, as none of them extract concepts without using any prior knowledge of the concepts.

To bridge this gap, we introduce a novel and challenging task named *Unsupervised Concept Extraction* (UCE). Given an image containing multiple objects, UCE aims to automatically extract the object concepts such that they can be used to generate new images. In UCE, we consider a strict and realistic "unsupervised" setting, in which there is no prior knowledge about the image or the concepts present within it. Specifically, "unsupervised" emphasizes **(1)** no concept descriptors for proper word embedding initialization, **(2)** no object masks for concept localization and disentanglement, and **(3)** no instance number for a definite number of concepts to be extracted. We illustrate UCE in Fig. 1.

To tackle this problem, we introduce **ConceptExpress**, the first method designed for unsupervised concept extraction. ConceptExpress unleashes the inherent capabilities of pretrained Stable Diffusion, enabling it to disentangle each concept in the compositional scene and learn discriminative conceptual tokens that represent each individual concept. ConceptExpress presents two major innovations. (1) For concept disentanglement, we propose a concept localization approach that automatically locates salient concepts within the image. This approach involves clustering spatial points on the self-attention map, building upon the observation that Stable Diffusion has learned good unsupervised spatial correspondence in the self-attention layers [65]. Our approach has three sequential steps, namely pre-clustering, filtering, and post-clustering, seamlessly integrating a parameter-free hierarchical clustering method [57]. By utilizing the end-of-text cross-attention map as a magnitude filter, we filter out non-salient backgrounds. Additionally, our approach automatically determines the number of concepts based on self-adaptive clustering constraints. (2) For conceptual token learning, we employ concept-wise masked denoising optimization by reconstructing the located concept. This optimization is based on a token lookup table that associates each located concept with its corresponding conceptual token. To address the issue of absence of initial words, which can detrimentally impact optimization [16], we introduce a split-and-merge strategy for robust token initialization, mitigating performance degradation. To prevent undesired cross-attention activation with the wrong concept, we incorporate regularization to align cross-attention maps with the desired concept activation exhibited in self-attention maps.

To evaluate the new UCE task, we construct a new dataset that contains various multi-concept images, and introduce an evaluation protocol including two metrics tailored for unsupervised concept extraction. We use concept similarity, including identity similarity and compositional similarity, to measure the absolute similarity between the source and the generated concepts. We also use classification accuracy to assess the degree of concept disentanglement. Through comprehensive experiments, our results demonstrate that ConceptExpress successfully tackles the challenge of unsupervised concept extraction, as evidenced by both qualitative and quantitative evaluations.

2 Related Work

Text-to-image synthesis In the realm of GANs [8,20,29–31], plenty of works have gained remarkable advancements in text-to-image generation [52,63,76,77,79,85] and text-driven image manipulation [1,18,47,74], significantly pushing forward image synthesis conditioned on plain text. Content-rich text-to-image generation is achieved by auto-regressive models [51,78] that are trained on large-scale text-image data. Based on the pretrained CLIP [49], Crowson *et al.* [14] optimizes the generated image at test time using CLIP similarity without any training. Diffusion-based methods [23] have pushed the boundaries of text-to-image generation to a new level, *e.g.*, DALL·E 2 [50], Imagen [56], GLIDE [45], and LDM [54]. Based on the implementation of LDMs [54], Stable Diffusion (SD), large-scale trained on LAION-5B [58], achieves unprecedented text-to-image synthesis performance. Diffusion models are widely used for various tasks such as controllable generation [81,84], global [9,66] and local editing [3,5,13,32,46,69], video generation [24,60,73] and editing [43,82], inpainting [41], and scene generation [4,6].

Generative concept learning Recently, many works [12,16,17,21,25,36,38,42,48,55,59,64,72] have emerged, aiming to learn a generative concept from multiple images. For example, Textual Inversion [16] learns an embedding vector that represents a concept in the textual embedding space. Liu *et al.* [40]

extended it to multi-concept discovery using composable diffusion models [15]. Their work operates in an unsupervised setting like ours. However, there is a major difference: they extract concepts from multiple images, each containing only one concept, whereas our focus is on extracting multiple concepts from a single image. Our work is closely related to Break-A-Scene [2] which relies heavily on human-annotated masks that are not available in our setting. The concurrent works, MCPL [26] and DisenDiff [83], address a similar problem, but they require either a concept-descriptive text caption or specific class names, which renders them infeasible for our task. There are also works related to generative concepts include concept erasing [19, 35], decomposition [11, 67], manipulation [71], and creative generation [53].

Attention-based segmentation Pre-trained Stable Diffusion [54] possesses highly informative semantic representations within its attention layers. This property effectively enables its cross-attention layers to indicate the interrelations between text and image tokens [62], and its self-attention layers to capture the spatial correspondence among image tokens. Consequently, prior works [7,28, 44,65,68,75] have explored the utilization of the pre-trained Stable Diffusion for semantic segmentation, showing remarkable performance in unsupervised zero-shot segmentation. Diffsegmenter [68] and FTM [75] use cross-attention to initialize segmentation maps, and then extract affinity weights from self-attention for further refinement. DiffSeg [65] achieves unsupervised zero-shot segmentation by clustering aggregated self-attention maps. Their investigation of the self-attention property inspires our concept localization approach.

3 Unsupervised Concept Extraction

We aim to learn discriminative tokens that can represent multiple instance-level concepts from a single image in an *unsupervised* manner. Specifically, given an image $\mathcal{I}$ containing multiple salient instances, we use a pretrained text-to-image diffusion model to discover a set of conceptual tokens $\mathcal{S} = \{[\mathtt{V_i}]\}_{i=1}^{N}$ and their corresponding embedding vectors $\mathcal{V} = \{v_i\}_{i=1}^{N}$, which capture discriminative concepts from $\mathcal{I}$. The concept number N is automatically determined in the discovery process. By prompting the i-th token $[\mathtt{V_i}] \in \mathcal{S}$, we can recreate the corresponding concept extracted from $\mathcal{I}$. We present ConceptExpress to tackle this problem. Fig. 2 gives an overview of ConceptExpress.

3.1 Preliminary

Text-to-image diffusion model [54] is composed of a pretrained autoencoder with an encoder $\mathcal{E}$ to extract latent codes and a corresponding decoder $\mathcal{D}$ to reconstruct images, a CLIP [49] text encoder that extracts text embeddings, and a denoising U-Net ϵ_θ with text-conditional cross-attention blocks. Textual inversion [16] represents a particular concept using a learnable embedding vector $v_\star$, which is optimized using a standard latent denoising loss with ϵ_θ frozen, written as

$$\mathcal{L} = \mathbb{E}_{z\sim\mathcal{E}(\mathcal{I}),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_{v_\star}(y))\|_2^2\right], \tag{1}$$

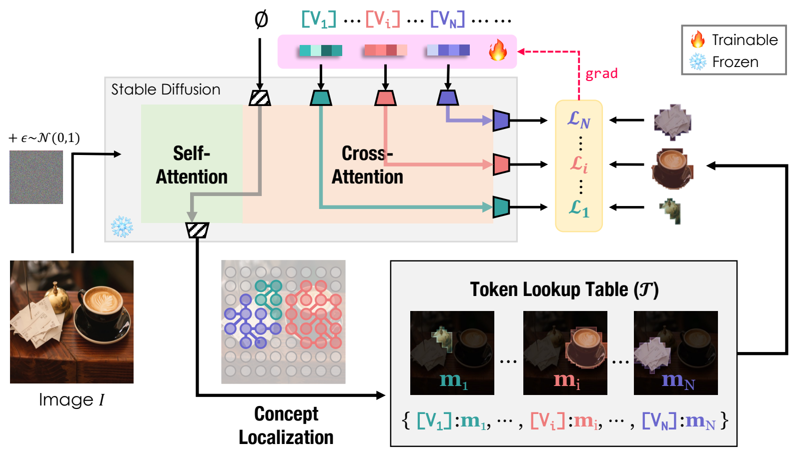

Fig. 2: Overview of ConceptExpress. ConceptExpress takes a multi-concept image $\mathcal{I}$ as input and learns a set of conceptual tokens. ConceptExpress consists of three key components. First, it leverages self-attention maps from the unconditional token $\varnothing$ to locate the latent concepts. Second, it constructs a token lookup table that associates each concept mask with its corresponding conceptual token $[\mathtt{V_i}]$. Finally, it optimizes each conceptual token using a masked denoising loss. The learned conceptual tokens can then be used to generate images that represent each individual concept. See Sec. 3 for more details of the method.

where t is the timestep, z_t is the latent code at timestep t, ϵ is the randomly sampled Gaussian noise, y is the text prompt, and $c_{v_\star}$ is the text encoder parameterized by the learnable $v_\star$. ConceptExpress advances further by learning multiple embedding vectors in an unsupervised setting.

FINCH [57] is an efficient parameter-free hierarchical clustering method. Given a set of n sample points in d dimensions, denoted as $\mathcal{S}=\{s_i \mid s_i \in \mathbb{R}^d\}_{i=1}^n$, we construct an adjacent matrix G for paired samples as

$$\mathrm{G}(i,j) = \begin{cases} 1 & \text{if } \kappa_i = j \text{ or } \kappa_j = i \text{ or } \kappa_i = \kappa_j \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

where κ_i represents the index of the closest sample to $s_i \in \mathcal{S}$ under a specific distance metric. To obtain a sample partition, we group the connected components within the undirected graph defined by the adjacency matrix G. Each connected component in the graph represents a cluster, and the centroids of the clusters are treated as super sample points for constructing a new adjacent matrix. This process enables iterative hierarchical clustering until all samples are grouped. As a result, multiple clustering levels of varying granularity are generated.

3.2 Automatic Latent Concept Localization

We begin by locating instance-level concepts within the diffusion latent space. In pretrained diffusion models, self-attention possesses good properties of spatial correspondence which offers the inherent benefit as an unsupervised semantic

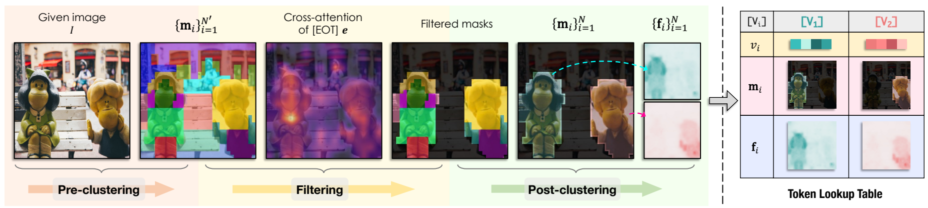

Fig. 3: Visualization. Left: we visualize the concept localization process, which involves: (1) pre-clustering that groups together semantically related regions; (2) filtering that removes non-salient regions that are not visually significant; and (3) post-clustering that integrates salient regions into instance-level concepts. **Right**: we visualize the token lookup table, which establishes a one-to-one correspondence between the conceptual token [$\mathtt{V_i}$] and the learnable embedding vector v_i, the latent mask $\mathbf{m}_i$, and the attention map $\mathbf{f}_i$.

segmenter [65]. With this insight, we propose an approach to automatically locating concepts by subtly leveraging self-attention.

Let $\mathbf{A}_l \in \mathbb{R}^{(h_l \times w_l)\times(h_l \times w_l)}$ denote the self-attention map from the l-th layer of the U-Net, where the feature map has a spatial resolution $h_l \times w_l$. To aggregate self-attention maps from different layers into an identical resolution $h \times w$, we follow the practice in [65] to interpolate the last two dimensions, duplicate the first two dimensions, and average all maps. The aggregated attention, denoted as $\mathbf{A} \in \mathbb{R}^{(h\times w)\times(h\times w)}$, can be represented as a set of $h \times w$ spatial samples, each of which is an $h \times w$ dimensional distribution, *i.e.*, $\mathcal{A}=\{a_i \mid a_i \in \mathbb{R}^{h\times w}\}_{i=1}^{h\times w}$. By clustering on $\mathcal{A}$, we can naturally derive latent masks that align with the semantic segmentation of the original image. This is because latent patches sharing similar semantics tend to possess consistent self-attention activations. The masks are formed by combining spatial samples belonging to the same cluster, effectively representing specific segments in the image. However, accurately locating instance-level concepts and effectively filtering out the background remain challenging when our goal is to disentangle multiple instances rather than solely segmenting semantics. To tackle this challenge, we adapt the hierarchical clustering algorithm FINCH [57] to generate latent masks that satisfy our needs.

Pre-clustering We first apply FINCH algorithm on $\mathcal{A}$. Since a_i is normalized and treated as a distribution, κ_i can be determined using a distribution distance metric, specifically the mean KL divergence, *i.e.*,

$$d(a_i, a_j) = \left(D_{KL}\left(a_i, a_j\right) + D_{KL}\left(a_j, a_i\right)\right)/2, \tag{3}$$

$$\kappa_i = \arg\min_j \{d(a_i, a_j) \mid a_j \in \mathcal{A}\}. \tag{4}$$

We set the upper limit of the number of discovered concepts to $N_{\max}$. We then identify the clustering level with the cluster number N' closest to but greater than $N_{\max}$. At this level, we construct a mask for each cluster from all spatial points within the cluster. We denote the resulting masks as $\{\mathbf{m}_i \mid \mathbf{m}_i \in \{0,1\}^{h\times w}\}_{i=1}^{N'}$. Since spatial samples within the same cluster share consistent semantics, the distribution distance between them serves as an effective indicator for distinguishing between different semantic instances. Therefore, we use the largest intra-cluster distance at this level, denoted as δ, as a self-adaptive threshold to determine the final clustering level in the post-clustering phase.

Filtering The obtained masks cover all areas on the latent map, encompassing both the foreground instances with clear semantics and the indistinct background regions. In diffusion models, the cross-attention map of the end-of-text token (`[EOT]`) demonstrates robust foreground localization capabilities [21], where salient regions exhibit higher magnitudes and vice versa. This characteristic makes it suited for automatically distinguishing between distinct instances and indistinct backgrounds. Let $\mathbf{e} \in \mathbb{R}^{h\times w}$ denote the cross-attention map of `[EOT]`. Based on $\mathbf{e}$, we discard masks whose masked regions satisfy

$$\frac{\|\mathrm{vec}(\mathbf{m}_i \odot \mathbf{e})\|_1}{\|\mathrm{vec}(\mathbf{m}_i)\|_1} < \frac{\|\mathrm{vec}(\mathbf{e})\|_1}{h \times w} \tag{5}$$

where $\mathrm{vec}(\cdot)$, $\|\cdot\|_1$, and $\odot$ denote matrix vectorization, ℓ_1 norm, and Hadamard product respectively. By applying this criterion, we filter out those masks whose masked regions show magnitudes lower than the average level, indicating that they correspond to indistinct regions. This criterion helps identify and exclude masks that correspond to indistinct regions in the `[EOT]` cross-attention map.

Post-clustering After filtering, we reapply FINCH to the remaining clusters iteratively. Additionally, we introduce two extra constraints to determine the stopping point in the clustering procedure. **(1)** To enhance the proximity of semantic relationships within the same mask, we set $\mathrm{G}(i,j)$=0 if the distance $d(a_i, a_j)$ exceeds δ, which is determined in the level with N' clusters of pre-clustering. By removing such connections, we hinder the grouping of strong semantic variations within the same mask. **(2)** We forbid non-adjacent masks from grouping together, *i.e.*, masks that are not spatially adjacent to each other cannot be clustered together, regardless of their connectivity in G. With these two constraints, the clustering will automatically terminate and yield N masks that locate the latent spaces corresponding to the N target concepts. The mean attention activations of each concept region is precisely the centroid of each cluster, given by

$$\mathbf{f}_i = \mathbf{m}_i^{1\times(h\times w)} \cdot \mathbf{A}^{(h\times w)\times(h\times w)} \;/\; \|\mathrm{vec}(\mathbf{m}_i)\|_1 \tag{6}$$

where the centroid $\mathbf{f}_i \in \mathbb{R}^{1\times(h\times w)}$ represents the average attention activations of i-th masked latent region to the entire $h\times w$ latent space. The latent masks and their corresponding attention activations are ready for token optimization. The concept localization process is visualized in Fig. 3 (left).

3.3 Concept-wise Masked Denoising

We construct a token lookup table

$$\mathcal{T}_{\text{lookup}} := \{[\mathtt{V_i}] : (v_i, \mathbf{m}_i, \mathbf{f}_i) \mid i = 1, 2, \cdots, N\} \tag{7}$$

where the i-th conceptual token $[\mathtt{V_i}]$ corresponds to a learnable embedding vector v_i, a latent mask $\mathbf{m}_i \in \{0,1\}^{h\times w}$, and a mean attention map $\mathbf{f}_i \in \mathbb{R}^{h\times w}$. We visualize the token lookup table in Fig. 3 (right). We employ the masked denoising loss [2] to optimize each token $[\texttt{V}_\texttt{i}] \in \mathcal{T}_{\text{lookup}}$:

$$\mathcal{L}_i = \mathbb{E}_{z\sim\mathcal{E}(\mathcal{I}),y_i,\epsilon,t}\left[\|[\epsilon - \epsilon_\theta\left(z_t, t, c_{v_i}(y_i)\right)] \odot \mathbf{m}_i\|_2^2\right] \tag{8}$$

where y_i is the text prompt "`a photo of` $[\texttt{V}_\texttt{i}]$" and v_i is the only trainable parameter. Masked denoising forces the new token to learn exclusively within specific latent regions that contain concept-wise information.

Robust token initialization To learn the concept embedding above, one may think to directly apply textual inversion [16]. However, the use of suitable words for initializing the conceptual tokens is crucial for successful textual inversion. In the case of an unsupervised setting, where specific words for initializing concept tokens are not available, the performance of [16] will deteriorate sharply. To resolve this problem, we propose a *split-and-merge* strategy that randomly initializes multiple tokens for each concept, which are later merged into a single token after several warm-up steps. Multiple tokens can explore a broader concept space, providing a greater opportunity for convergence into an embedding vector that can more precisely represent the underlying concept. Formally, we randomly initialize g tokens $\{[\texttt{V}_\texttt{i}]^\texttt{j}\}_{j=1}^{g}$ for each concept and extend the token lookup table as

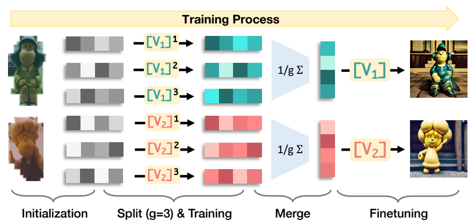

Fig. 4: Split-and-merge. During the training process, we sequentially initialize conceptual tokens, train the split tokens, merge the tokens by averaging, and further fine-tune the merged tokens. Finally, the merged tokens are well-learned and effectively represent individual concepts.

$$\mathcal{T}_{\text{lookup}}^{\text{split}} := \left\{[\texttt{V}_\texttt{i}]^\texttt{j} : (v_i^j, \mathbf{m}_i, \mathbf{f}_i) \mid i{=}1, ..., N; j{=}1, ..., g\right\}, \tag{9}$$

where v_i^j is the j-th randomly initialized embedding vector corresponding to the conceptual token $[\texttt{V}_\texttt{i}]^\texttt{j}$. At the early training steps, we optimize the loss in Eq. (8) on $[\texttt{V}_\texttt{i}]^\texttt{j} \in \mathcal{T}_{\text{lookup}}^{\text{split}}$, *i.e.*, $\mathcal{L}_{i,j}$, to learn the $g \times N$ tokens. In addition, leveraging the constraint that embeddings for the same concept should exhibit a closer embedding distance, we incorporate a contrastive loss for each token $[\texttt{V}_\texttt{i}]^\texttt{j}$ as

$$\mathcal{L}_{i,j}^{con} = -\frac{1}{g{\times}N}\ log\frac{\sum_{v_i^q \in \mathcal{V}_i \setminus \{v_i^j\}} \exp(v_i^j \cdot v_i^q/\tau)}{\sum_{v_m^n \in \mathcal{V} \setminus \{v_i^j\}} \exp(v_i^j \cdot v_m^n/\tau)}, \tag{10}$$

where τ is the temperature, $\mathcal{V}$ is the full set of embedding vectors, and $\mathcal{V}_i$ is the subset of embedding vectors that correspond to the i-th concept. Eq. (10) enforces tokens representing the same concept to be closer to each other, inducing these randomly initialized embedding vectors to converge to a shared space during several warm-up training steps. Afterward, we merge the tokens by computing the mean value of the g embeddings associated with each concept. The token lookup table is reset to Eq. (7), where the token embeddings are good initializers to robustly represent the corresponding concepts. In the subsequent training steps, we use the denoising loss described in Eq. (8) to optimize the merged tokens that represent each concept on a one-to-one basis. We depict the training process using split-and-merge in Fig. 4.

Attention alignment Although each conceptual token is optimized to reconstruct the masked region, there is a lack of direct alignment between the tokens and individual concepts within a compositional scene. This absence of alignment leads to inaccurate cross-attention activation for the learned conceptual tokens, which hinders the performance of compositional generation. To address this problem, for each token in the lookup table, we align its cross-attention map with the mean attention $\mathbf{f}_i$ of the corresponding masked region using a location-aware earth mover's distance (EMD) regularization. The earth moving cost is computed as the Euclidean distance between the 2D locations on two attention maps. Let the cross attention map of the token $[\mathtt{V_i}]^{\mathtt{j}}$ be $\mathbf{c}_{[\mathtt{V_i}]^{\mathtt{j}}} \in \mathbb{R}^{h\times w}$, where j can be omitted after token merging. The regularization loss is formulated as

$$\mathcal{L}_{i,j}^{reg} = \mathtt{EMD}(\mathbf{c}_{[\mathtt{V_i}]^{\mathtt{j}}}, \mathbf{f}_i) \tag{11}$$

which softly guides the cross-attention map to match the desired concept activations exhibited in the self-attention map.

3.4 Implementation Details

We train the tokens in two phases for a total of 500 steps, with a learning rate of 5e-4. In the first 100 steps, we optimize the tokens $[\mathtt{V_i}]^{\mathtt{j}} \in \mathcal{T}_{\text{lookup}}^{\text{split}}$ using

$$\mathcal{L} = \frac{1}{g\times N}\sum_{i=1}^{N}\sum_{j=1}^{g}(\mathcal{L}_{i,j} + \alpha\mathcal{L}_{i,j}^{con} + \beta\mathcal{L}_{i,j}^{reg}). \tag{12}$$

We then merge the tokens, deriving $[\mathtt{V_i}] \in \mathcal{T}_{\text{lookup}}$, and optimize them in the subsequent 400 steps using

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(\mathcal{L}_i + \beta\mathcal{L}_i^{reg}). \tag{13}$$

We use Stable Diffusion v2-1 [54] as our base model. We set α=1e-3, β=1e-5, τ=0.07, and g=5. All experiments are conducted on a single RTX 3090 GPU. In our implementation, self-attention used in concept localization is computed using the unconditional text prompt $\varnothing$ at timestep 0, which induces minimal textual intervention and maximal denoising of the given image.

4 Experiments

4.1 Dataset and Baseline

Dataset In our work, we do not rely on predefined object masks or manually selected initial words for training images. This allows us to gather high-quality images from the Internet without human annotations to form our dataset. Specifically, we collect a set D_1 of 96[1] images from Unsplash[2], ensuring that each

[1] 96 is considerably large compared to the dataset sizes in the previous works, such as 30 in DreamBooth [55], 50 in Break-A-Scene [2], and 10 in DisenDiff [83].

[2] `https://unsplash.com/`

image contains at least two distinct instance-level concepts. The collected images encompass a wide range of object categories, including animals, characters, toys, accessories, containers, sculptures, buildings, landscapes, vehicles, foods, and plants. For a fair comparison, we construct a set D_2 using the 7 images provided by [2]. For evaluation, we generate 8 testing images for each training image using the prompt "a photo of [$\mathtt{V_i}$]".

Baseline To the best of our knowledge, Break-A-Scene [2] is the only work that is closely related to the problem in this paper. However, Break-A-Scene operates under a strongly supervised setting, which requires significant prior knowledge of the training image, including the number of concepts, object masks, and properly selected initial words. To ensure a fair and meaningful comparison, we adapt Break-A-Scene to our unsupervised setting. Specifically, we disable the use of manually picked initial words and instead apply random initialization. Additionally, we leverage the instance masks identified by our method as the annotated masks for Break-A-Scene. Finally, we exclusively train the learnable tokens without fine-tuning the diffusion model. We use this adapted version of Break-A-Scene, denoted as BaS†, as the baseline method for comparison.

4.2 Evaluation Metric

We establish an evaluation protocol including two metrics described as follows.

Concept similarity To quantify how well the model is able to recreate the concepts accurately, we evaluate the concept similarity, including identity similarity (SIM^{I}) and compositional similarity (SIM^{C}). Identity similarity measures the similarity between each concept in the training image and the concept-specific generated images. We employ CLIP [49] and DINO [10] to compute the similarities. To ensure that the similarity is computed specifically for the i-th concept, we obtain concept-wise masks with SAM [33] by identifying the specific SAM mask associated with our extracted concept. Specifically, for each concept, we prompt SAM with 3 randomly sampled points on our extracted mask to produce SAM masks. The training image is then masked with the SAM mask corresponding to the i-th concept. The metric of identity similarity provides a crucial criterion for evaluating the intra-concept performance of unsupervised concept extraction. Compositional similarity measures the CLIP or DINO similarity between the source image and the generated image, conditioned on the prompt "a photo of [$\mathtt{V_1}$] and [$\mathtt{V_2}$] ... [$\mathtt{V_N}$]". This metric quantifies the degree to which the source image can be reversed using the extracted concepts.

Classification accuracy To assess the extent of disentanglement achieved for each concept within the full set of extracted concepts, we establish a benchmark that evaluates concept classification accuracy. Specifically, we first employ a vision encoder, such as CLIP [49] or DINO [10], to extract feature representations for each concept from the SAM-masked training images. In total, we obtain 264 concepts in D_1 and 19 concepts in D_2. We use these concept features as prototypes to construct a concept classifier. We then employ the same vision encoder to extract query features for all generated images, each associated with

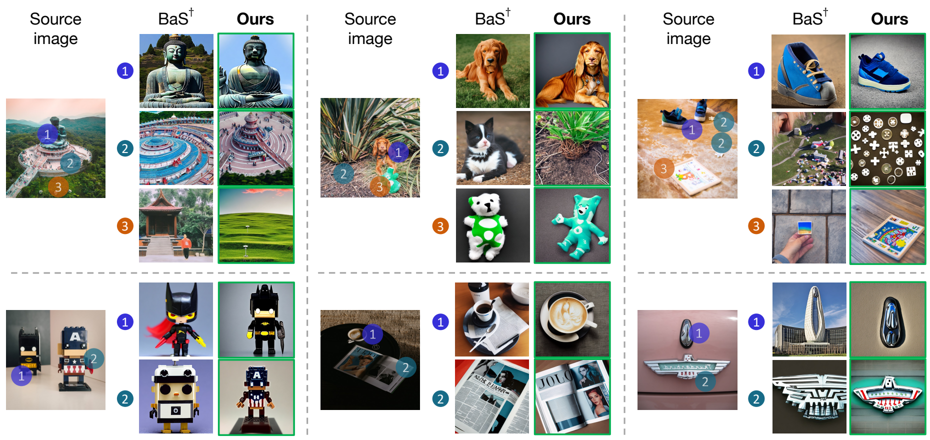

Fig. 5: Comparison with BaS† [2]. We compare the concept extraction results of BaS† and ConceptExpress in 6 examples. For each example, we show the source image and the generated concept images. We annotate concepts in serial numbers for legibility.

Table 1: Quantitative comparison. For reference, we also provide the results of original Break-A-Scene [2] on D_2 (marked in grey), by using mask and initializer supervision (BaS) and further finetuning (BaS f.t.).

(a) Evaluation using CLIP [49].

Method	D_1 SIM^{I}	D_1 SIM^{C}	D_1 ACC^1	D_1 ACC^3	D_2 SIM^{I}	D_2 SIM^{C}	D_2 ACC^1	D_2 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].

Method	D_1 SIM^{I}	D_1 SIM^{C}	D_1 ACC^1	D_1 ACC^3	D_2 SIM^{I}	D_2 SIM^{C}	D_2 ACC^1	D_2 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**

a specific concept category. Finally, we evaluate the top-k classification accuracy of the query features using our concept classifier. We report classification results for k=1, 3, denoting the top-k accuracy as ACC^k. This metric effectively assesses the inter-concept performance of unsupervised concept extraction.

4.3 Performance

Quantitative comparison We compare ConceptExpress with BaS† based on concept similarity and classification accuracy metrics. The quantitative comparison results on the two datasets are reported in Tabs. 1a and 1b, respectively with CLIP [49] and DINO [10] as the visual encoder. Notably, ConceptExpress outperforms BaS† by a significant margin on all evaluation metrics. It achieves higher concept similarity SIM^{I} and SIM^{C}, indicating a closer alignment with the source concepts. It also achieves higher classification accuracy ACC^1 and ACC^3, indicating a more significant level of disentanglement among the individually extracted concepts. These results highlight the limitations of the existing concept extraction approach [2] and establish ConceptExpress as the state-of-the-art method for the UCE problem.

Qualitative comparison We show several generation samples of ConceptExpress and BaS† in Fig. 5. ConceptExpress presents overall better generation

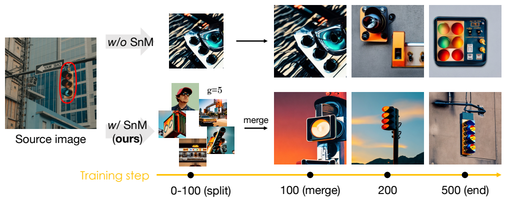

Fig. 6: Generation results of Split-and-merge (SnM) ablation. We show the generated concept "traffic light" throughout the training process, with (bottom) and without (top) SnM that utilizes g= 5 diverse tokens.

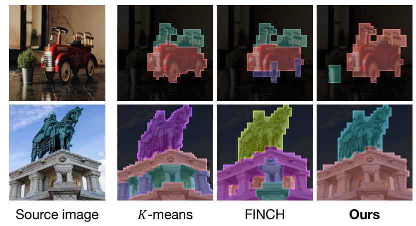

Fig. 7: Comparison on self-attention clustering. Each located concept is enclosed within a distinct colored region.

fidelity and quality than BaS†. We observe some defects in the generations of BaS†. For example, in the top left ❸ and the top center ❷, the generation of BaS† deviates from the source concept. In addition, BaS† fails to preserve the characteristics of the source concept in the top center ❸ and the bottom left ❶ ❷. ConceptExpress effectively overcomes the defects of wrong identity and poor preservation observed in the generations in BaS†. ConceptExpress consistently generates high-quality images that precisely align with the source concepts.

4.4 Ablation Study

We conduct a quantitative ablation study on the training components in Tab. 2.

Effectiveness of split-and-merge strategy (SnM) By comparing Rows (0) and (1), we validate the benefit of the split-and-merge strategy to initializer-absent training. The split-and-merge strategy effectively improves identity similarity and classification accuracy while slightly sacrificing compositional similarity due to its strong focus on a single concept. In Fig. 6, we present the generated images at different training steps, which reveals how SnM rectifies the training direction. The results illustrate that SnM effectively expands the concept space, allowing learnable tokens to explore a wider range of concepts, ultimately resulting in a more faithful concept indicator.

Effectiveness of regularization By comparing Rows (0) and (2) in Tab. 2, we observe that regularizing the attention map can enhance the generation performance of individual concepts. Row (3) is our full method which further improves the performance regarding all metrics compared to incorporating each component in Rows (1) and (2). The thorough ablation study indicates the effectiveness of each training component in ConceptExpress.

4.5 Concept Localization Analysis

Self-attention clustering To validate the significance of our three-phase method for concept localization, we compare it with k-means and our base method FINCH [57]. Since k-means requires a predefined cluster number and FINCH requires a stopping point, we set a proper cluster number of 7 for them. After clustering, we apply the proposed filtering method for fair comparison. We

Table 2: Ablation study. Concept-wise optimization, split-and-merge strategy, and regularization are respectively abbreviated as CwO, SnM, and Reg. The results are evaluated on D_2 using DINO.

	CwO	SnM	Reg	SIM^{I}	SIM^{C}	ACC^1	ACC^3
(0)	✓			0.344	**0.549**	0.625	0.776
(1)	✓	✓		0.362	0.519	0.750	0.901
(2)	✓		✓	0.364	0.490	0.724	0.895
(3)	✓	✓	✓	**0.371**	0.535	**0.803**	**0.934**

Table 3: Comparison of concept localization. "#clust" predefines the cluster number for k-kmeans and FINCH. The best value is highlighted in **bold**, while the second-best value is underlined.

Method	k-means				FINCH				Ours	MC [70]
#clust	4	5	6	7	4	5	6	7	auto	given
IoU	52.8	51.4	48.7	47.6	50.7	<u>54.8</u>	54.6	51.5	**57.3**	58.1
Recall	70.3	90.1	95.5	95.7	83.7	93.4	<u>96.9</u>	**97.9**	89.1	97.3
Precision	92.7	88.0	81.5	75.4	**98.7**	90.7	85.8	77.6	<u>93.7</u>	77.0

compare the visualized results in Fig. 7. We can observe that using k-means or FINCH will miss some concepts (the 1st row) or split a single concept (the 2nd row). In contrast, our method effectively locates complete and intact concepts, automatically determining a reasonable concept number.

Concept localization benchmark To quantitatively evaluate the concept localization performance, we establish a benchmark by (1) building a test dataset of multi-concept images along with ground-truth concept masks, and (2) devising tailored metrics for concept localization. The test dataset is sourced from CLEVR [27], a synthetic dataset featuring clean backgrounds and clear, distinct objects. In this dataset, each object explicitly represents a concept, thereby eliminating potential discrepancies in human-defined concepts in natural images. By comparing the predicted masks with the ground-truth masks in the test set using the Hungarian algorithm [34], we can evaluate three metrics: (1) Intersection over Union (IoU) that assesses segmentation accuracy, (2) Recall that evaluates the proportion of true concepts the model can discover, and (3) Precision that evaluates the correctness of the discovered concepts. We provide additional details of this evaluation benchmark for concept localization in Appendix B.

Quantitative evaluation Based on the established benchmark, we evaluate our method compared to k-means and FINCH with various predefined cluster numbers in Tab. 3. We also report the results of a training-free segmentation method MaskCut (MC) introduced in CutLER [70] for reference, which performs comparably to our method. When using k-means and FINCH, the predefined cluster number can significantly impact performance, and no fixed number can consistently achieve desired performance across all metrics. In contrast, our method performs well in terms of IoU and Precision, with a slight trade-off in Recall. One possible reason is that, unlike specifying the cluster number, our model automatically determines it. Therefore, there may be cases where two concepts are merged into one, resulting in one concept being unable to be matched to the ground truth, potentially reducing recall. Nevertheless, as the only method capable of automatically determining the number of concepts, our method achieves the best overall performance compared to all other clustering techniques.

4.6 Unsupervised *vs.* Supervised

Although ConceptExpress is an unsupervised model, it would be intriguing to compare ConceptExpress to some supervised methods. Motivated by this, we

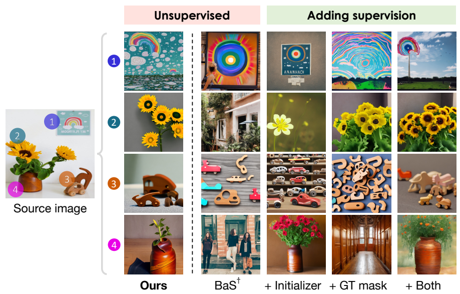

Fig. 8: Comparison with supervised methods. We compare ConceptExpress and BaS† to the supervised methods of Break-A-Scene [2] with added initializers, ground truth masks, and both of them.

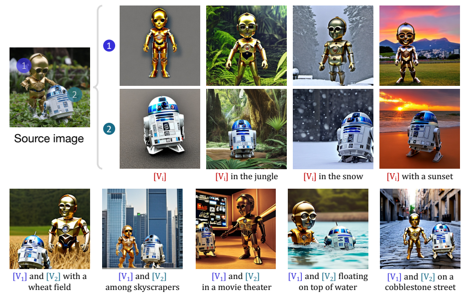

Fig. 9: Text-prompted generation. We show the generation results prompted by various text contexts using a single conceptual token (top) and multiple conceptual tokens (bottom).

experiment by providing initial words and ground-truth object masks (obtained by SAM [33]) for the supervised method Break-A-Scene [2]. We compare our method, BaS†, and three supervised methods by adding different supervision to Break-A-Scene, as shown in Fig. 8. We can observe that adding initial words guides the generation towards the specified category, while adding ground-truth object masks enhances the preservation of texture details. However, even with these two settings, the generated results of Break-A-Scene still fall short compared to our unsupervised model. Despite being trained in a completely unsupervised manner, our model performs on par with the fully supervised setting, where both types of supervision are used.

4.7 Text-prompted Generation

With the extracted generative concepts, we can perform text-prompted generation. In Fig. 9, we showcase the results conditioned on various text prompts using both individual concepts and compositional concepts. The results demonstrate that the learned conceptual tokens can generate images with high text fidelity, aligning faithfully with the text prompt. Furthermore, the images generated with the conceptual tokens also preserve consistent concept identity with the source concepts in both individual and compositional generation. Please refer to Appendix F for additional photorealistic results of text-prompted generation.

5 Conclusion

In this paper, we introduce Unsupervised Concept Extraction (UCE) that aims to leverage diffusion models to learn individual concepts from a single image in an unsupervised manner. We present ConceptExpress to tackle the UCE problem by harnessing the capabilities of pretrained diffusion models to locate concepts and learn their corresponding conceptual tokens. Moreover, we establish an evaluation protocol for the UCE problem. Extensive experiments highlight ConceptExpress as a promising solution to the UCE task.

Acknowledgement This work is partially supported by the Hong Kong Research Grants Council - General Research Fund (Grant No.: 17211024).

References

1. Abdal, R., Zhu, P., Femiani, J., Mitra, N., Wonka, P.: CLIP2StyleGAN: Unsupervised extraction of stylegan edit directions. In: ACM SIGGRAPH (2022)
2. Avrahami, O., Aberman, K., Fried, O., Cohen-Or, D., Lischinski, D.: Break-a-scene: Extracting multiple concepts from a single image. In: SIGGRAPH Asia (2023)
3. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. In: ACM SIGGRAPH (2023)
4. Avrahami, O., Hayes, T., Gafni, O., Gupta, S., Taigman, Y., Parikh, D., Lischinski, D., Fried, O., Yin, X.: Spatext: Spatio-textual representation for controllable image generation. In: CVPR (2023)
5. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: CVPR (2022)
6. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation. In: ICML (2023)
7. Baranchuk, D., Voynov, A., Rubachev, I., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. In: ICLR (2022)
8. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: ICLR (2019)
9. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR (2023)
10. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
11. Chefer, H., Lang, O., Geva, M., Polosukhin, V., Shocher, A., Irani, M., Mosseri, I., Wolf, L.: The hidden language of diffusion models. arXiv preprint arXiv:2306.00966 (2023)
12. Chen, W., Hu, H., Li, Y., Rui, N., Jia, X., Chang, M.W., Cohen, W.W.: Subject-driven text-to-image generation via apprenticeship learning. In: NeurIPS (2023)
13. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. In: ICLR (2022)
14. Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L., Raff, E.: VQGAN-CLIP: Open domain image generation and editing with natural language guidance. In: ECCV (2022)
15. Du, Y., Li, S., Mordatch, I.: Compositional visual generation with energy based models. In: NeurIPS (2020)
16. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: ICLR (2023)
17. Gal, R., Arar, M., Atzmon, Y., Bermano, A.H., Chechik, G., Cohen-Or, D.: Encoder-based domain tuning for fast personalization of text-to-image models. arXiv preprint arXiv:2302.12228 (2023)
18. Gal, R., Patashnik, O., Maron, H., Bermano, A.H., Chechik, G., Cohen-Or, D.: Stylegan-nada: Clip-guided domain adaptation of image generators. ACM Transactions on Graphics (TOG) (2022)
19. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. In: ICCV (2023)

20. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
21. Hao, S., Han, K., Zhao, S., Wong, K.Y.K.: Vico: Detail-preserving visual condition for personalized text-to-image generation. arXiv preprint arXiv:2306.00971 (2023)
22. Hedlin, E., Sharma, G., Mahajan, S., Isack, H., Kar, A., Tagliasacchi, A., Yi, K.M.: Unsupervised semantic correspondence using stable diffusion. arXiv preprint arXiv:2305.15581 (2023)
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
24. Ho, J., Salimans, T., Gritsenko, A.A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. In: NeurIPS (2022)
25. Jia, X., Zhao, Y., Chan, K.C., Li, Y., Zhang, H., Gong, B., Hou, T., Wang, H., Su, Y.C.: Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. arXiv preprint arXiv:2304.02642 (2023)
26. Jin, C., Tanno, R., Saseendran, A., Diethe, T., Teare, P.: An image is worth multiple words: Learning object level concepts using multi-concept prompt learning. arXiv preprint arXiv:2310.12274 (2023)
27. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In: CVPR (2017)
28. Karazija, L., Laina, I., Vedaldi, A., Rupprecht, C.: Diffusion models for zero-shot open-vocabulary segmentation. arXiv preprint arXiv:2306.09316 (2023)
29. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. In: NeurIPS (2021)
30. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
31. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: CVPR (2020)
32. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: CVPR (2023)
33. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
34. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly (1955)
35. Kumari, N., Zhang, B., Wang, S.Y., Shechtman, E., Zhang, R., Zhu, J.Y.: Ablating concepts in text-to-image diffusion models. In: ICCV (2023)
36. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: CVPR (2023)
37. Li, A.C., Prabhudesai, M., Duggal, S., Brown, E., Pathak, D.: Your diffusion model is secretly a zero-shot classifier. In: ICCV (2023)
38. Li, D., Li, J., Hoi, S.C.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In: NeurIPS (2023)
39. Li, X., Lu, J., Han, K., Prisacariu, V.: Sd4match: Learning to prompt stable diffusion model for semantic matching. arXiv preprint arXiv:2310.17569 (2023)
40. Liu, N., Du, Y., Li, S., Tenenbaum, J.B., Torralba, A.: Unsupervised compositional concepts discovery with text-to-image generative models. In: ICCV (2023)
41. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: CVPR (2022)

42. Ma, Y., Yang, H., Wang, W., Fu, J., Liu, J.: Unified multi-modal latent diffusion for joint subject and text conditional image generation. arXiv preprint arXiv:2303.09319 (2023)
43. Molad, E., Horwitz, E., Valevski, D., Acha, A.R., Matias, Y., Pritch, Y., Leviathan, Y., Hoshen, Y.: Dreamix: Video diffusion models are general video editors. arXiv preprint arXiv:2302.01329 (2023)
44. Ni, M., Zhang, Y., Feng, K., Li, X., Guo, Y., Zuo, W.: Ref-diff: Zero-shot referring image segmentation with generative models. arXiv preprint arXiv:2308.16777 (2023)
45. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: ICML (2022)
46. Patashnik, O., Garibi, D., Azuri, I., Averbuch-Elor, H., Cohen-Or, D.: Localizing object-level shape variations with text-to-image diffusion models. In: ICCV (2023)
47. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: StyleCLIP: Text-driven manipulation of stylegan imagery. In: ICCV (2021)
48. Qiu, Z., Liu, W., Feng, H., Xue, Y., Feng, Y., Liu, Z., Zhang, D., Weller, A., Schölkopf, B.: Controlling text-to-image diffusion by orthogonal finetuning. In: NeurIPS (2023)
49. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
50. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125 (2022)
51. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML (2021)
52. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: ICML (2016)
53. Richardson, E., Goldberg, K., Alaluf, Y., Cohen-Or, D.: Conceptlab: Creative generation using diffusion prior constraints. arXiv preprint arXiv:2308.02669 (2023)
54. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
55. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR (2023)
56. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022)
57. Sarfraz, S., Sharma, V., Stiefelhagen, R.: Efficient parameter-free clustering using first neighbor relations. In: CVPR (2019)
58. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: NeurIPS (2022)
59. Shi, J., Xiong, W., Lin, Z., Jung, H.J.: Instantbooth: Personalized text-to-image generation without test-time finetuning. arXiv preprint arXiv:2304.03411 (2023)
60. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. In: ICLR (2022)

61. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
62. Tang, R., Pandey, A., Jiang, Z., Yang, G., Kumar, K., Lin, J., Ture, F.: What the daam: Interpreting stable diffusion using cross attention. In: ACL (2023)
63. Tao, M., Tang, H., Wu, F., Jing, X.Y., Bao, B.K., Xu, C.: DF-GAN: A simple and effective baseline for text-to-image synthesis. In: CVPR (2022)
64. Tewel, Y., Gal, R., Chechik, G., Atzmon, Y.: Key-locked rank one editing for text-to-image personalization. In: ACM SIGGRAPH (2023)
65. Tian, J., Aggarwal, L., Colaco, A., Kira, Z., Gonzalez-Franco, M.: Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. arXiv preprint arXiv:2308.12469 (2023)
66. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: CVPR (2023)
67. Vinker, Y., Voynov, A., Cohen-Or, D., Shamir, A.: Concept decomposition for visual exploration and inspiration. arXiv preprint arXiv:2305.18203 (2023)
68. Wang, J., Li, X., Zhang, J., Xu, Q., Zhou, Q., Yu, Q., Sheng, L., Xu, D.: Diffusion model is secretly a training-free open vocabulary semantic segmenter. arXiv preprint arXiv:2309.02773 (2023)
69. Wang, S., Saharia, C., Montgomery, C., Pont-Tuset, J., Noy, S., Pellegrini, S., Onoe, Y., Laszlo, S., Fleet, D.J., Soricut, R., et al.: Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In: CVPR (2023)
70. Wang, X., Girdhar, R., Yu, S.X., Misra, I.: Cut and learn for unsupervised object detection and instance segmentation. In: CVPR (2023)
71. Wang, Z., Gui, L., Negrea, J., Veitch, V.: Concept algebra for text-controlled vision models. arXiv preprint arXiv:2302.03693 (2023)
72. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. arXiv preprint arXiv:2302.13848 (2023)
73. Wu, J.Z., Ge, Y., Wang, X., Lei, W., Gu, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. arXiv preprint arXiv:2212.11565 (2022)
74. Xia, W., Yang, Y., Xue, J.H., Wu, B.: Tedigan: Text-guided diverse face image generation and manipulation. In: CVPR (2021)
75. Xiao, C., Yang, Q., Zhou, F., Zhang, C.: From text to mask: Localizing entities using the attention of text-to-image diffusion models. arXiv preprint arXiv:2309.04109 (2023)
76. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018)
77. Ye, H., Yang, X., Takac, M., Sunderraman, R., Ji, S.: Improving text-to-image synthesis using contrastive learning. arXiv preprint arXiv:2107.02423 (2021)
78. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. Transactions on Machine Learning Research (2022)
79. Zhang, H., Koh, J.Y., Baldridge, J., Lee, H., Yang, Y.: Cross-modal contrastive learning for text-to-image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2021)
80. Zhang, J., Herrmann, C., Hur, J., Cabrera, L.P., Jampani, V., Sun, D., Yang, M.H.: A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. In: NeurIPS (2023)

81. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543 (2023)
82. Zhang, Y., Wei, Y., Jiang, D., ZHANG, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. In: ICLR (2024)
83. Zhang, Y., Yang, M., Zhou, Q., Wang, Z.: Attention calibration for disentangled text-to-image personalization. In: CVPR (2024)
84. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Uni-controlnet: All-in-one control to text-to-image diffusion models. In: NeurIPS (2023)
85. Zhu, M., Pan, P., Chen, W., Yang, Y.: DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. In: CVPR (2019)