# SELDOM: Scene Editing via Latent Diffusion with Object-centric Modifications

Richard E. L. Higgins
University of Michigan
relh@umich.edu

David F. Fouhey
New York University
david.fouhey@nyu.edu

## Abstract

*We introduce SELDOM to composably edit scenes—mixing and matching objects with backgrounds, camera changes, and object-centric edits. SELDOM is a 3D-aware diffusion editing method which conditions on sequences of "neural nouns and verbs". Neural nouns represent scene state and are visual features extracted from source image(s). Neural verbs are learnt representations for image edits, formed either by explicitly parsing prompts or implicitly attending to them. Neural verbs combine with their associated neural nouns to convey object-centric transformations. Finally, a sequence of these tokens is composed with scene background tokens and used as conditioning for a fine-tuned latent diffusion model. Our factorization affords test-time compositionality, allowing us to compose edited objects from multiple datasets into a single scene. We further demonstrate our model's ability to photo-edit: SELDOM can convincingly edits scenes to change object hue and lighting, scale and rotate objects, apply diverse language-based edits, and control camera rotation and translation.*

## 1. Introduction

Fig. 1 shows two source scenes, a shoe on a sidewalk and a cereal box in a store. How might one edit the first scene to rotate the camera, change the color of the shoe, and insert the cereal box? How might one edit any scene (or scenes) to compose objects, potentially across datasets, with multiple transformations, and with a new camera orientation? Extracting 3D-aware object representations from 2D images, editing them, and then composing them can enable deep-learning photo editing and provide rich visual representations for downstream tasks. Our goal is to produce such a representation.

Current text-based image editing methods are impressive. But prompts like "move the chair to the right of the couch" (and language generally) lack specificity compared with the precise control of 3D objects [51] one might desire. Conversely, non text-based, explicitly 3D-aware methods, such as Neural Assets [107], excel at precise spatial transformations but lack the capability to perform general edits involving attributes, styles, or actions (e.g., changing object color or state). Existing methods for text-based control using diffusion [2, 7, 31, 81] partially address precision limitations, but are still limited by the vagueness of language. Moreover, most image-editing approaches are limited in that they do not directly offer the ability to retain subject identity or compose edits.

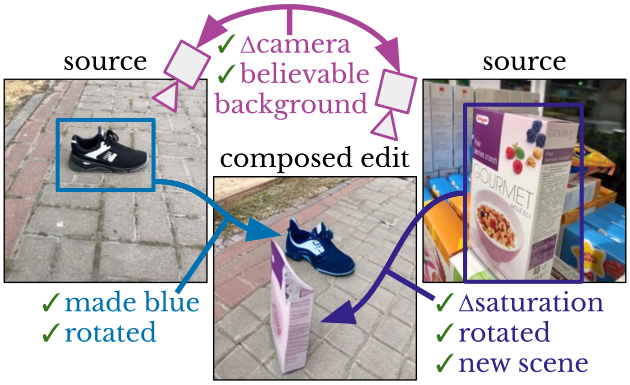


Figure 1. Our method, SELDOM, uses neural nouns and verbs for compositional image edits within and across scenes via latent diffusion. Here the shoe from the first source scene is made blue and the camera is rotated. The box from the second source scene is rotated, slightly saturated, and inserted into the first scene. SELDOM performs edits, composes the second object into the first scene, and rotates the camera viewpoint.

Prior image editing work often uses a single source image (concatenated with to-be-denoised latents) and a text prompt. But how might one combine multiple visual entities across scenes (as in Fig. 1)? Should one just concatenate/stack more images? Single-image visual conditioning is a limitation that prevents the broad type of composability a simple text prompt already supports. Furthermore, unlike rich text encodings, images (and their minimalist latent conditionings) are raw rather than representa-

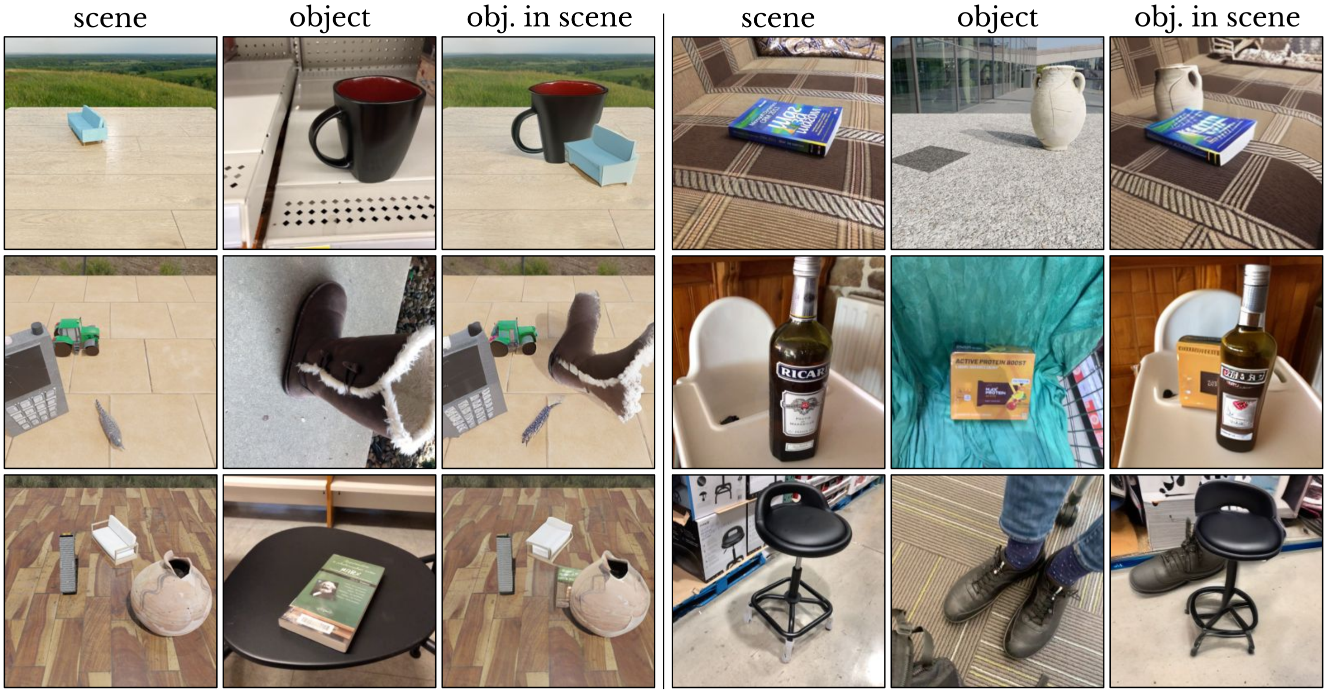


Figure 2. **Object Compositionality**: Each triplet (scene, object, obj. in scene) shows two source images and SELDOM's output applying object-centric editing and then recomposing the scene. SELDOM first extracts features and edits from multiple images and prompts, then applies the edits, and then recompose the scene as a sequence of neural nouns and verbs. SELDOM edits entities: rotating, translating, changing attributes such as color, inserting synthetic objects into real scenes and vice versa, and finally also changing the camera orientation. The left triplets insert Objectron objects into OBJect scenes. The right triplets are scenes from the Objectron dataset.

tional. Diffusion systems may be limited when concurrently parsing their visual scene while generating outputs. When it comes to editing, encoders like CLIP [80] are limited as they associate captions—not instructional prompts describing edits—with images. Yet latent diffusion [81] can separately generate images corresponding to the before and after of instructional prompts, e.g., a whole vs. chopped carrot or a lounging vs. prowling cat. These before and afters don't preserve identity but *their existence* shows that the model can understand these states. This highlights the power of text-based generation for showing what *is*—just not what's *changing*. Still, text-based editing methods sometimes fail to preserve object identity, fail to represent inarticulatable changes, and can be restricted to conditioning on a single scene. We explore visual representations for moving between states, learning features for what-can-be instead of what-is. We pursue an object-centric composable editing approach, breaking scenes down into objects, background, and camera viewpoint before applying verbs individually and re-recomposing them. We learn a disentangled, visual understanding of verbs/actions for image generation.

Our method is SELDOM, Scene Editing via Latent Diffusion with Object-centric Modifications. SELDOM bridges the gap between broad language-based editability and precise 3D control. Our approach allows us to insert specific objects into new scenes, perform 3D transformations, and edit objects—so as to apply attribute edits such as color changes—all while preserving object identity. SELDOM extracts visual "neural nouns", textual "neural verbs", and then associates them based on similarity. Sequences of "neural nouns" describe the *state* of the scene, similar to a non-editing text prompt, while "neural verbs" capture the edits to be performed, similar to a text instruction. The association between neural nouns and verbs controls which verbs can affect which entities, and the edited scene is a result of conditioning a diffusion model with a sequence of these associated tokens. Our contributions are:

- We introduce neural nouns and verbs, a composable means of image editing that balances precise 3D control with the broad expressivity of text-based prompts.
- We improve upon the state-of-the-art composable pose editing method Neural Assets [107] without needing 3D bounding boxes and while supporting the application of verbs beyond just pose.
- SELDOM believably edits and insert objects (both real and synthetic) across different datasets and scenes.
- SELDOM performs language-based photo edits such as hue shift, scaling, brightness, and camera control.

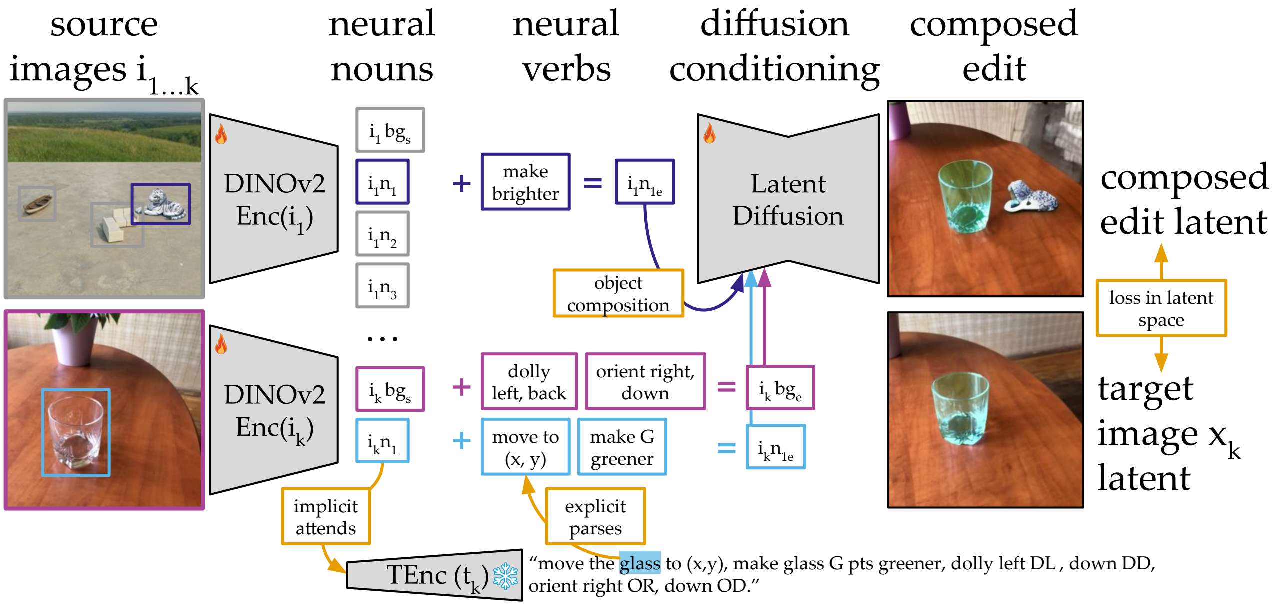


Figure 3. For image edits we require a source image and an edit prompt. We extract visual features from the source image, extracting 2 × 2 sets of ROI aligned features for each object of interest. We generate conditioning for our UNet by concatenating these features with either implicit or explicit *neural verbs*, magnitudes, and location information. We optionally edit and compose objects across scenes and datasets, as shown with the tiger figurine inserted into the edited table scene.

## 2. Related Work

**Diffusion-Based Image Editing.** Diffusion models have substantially advanced image editing and synthesis [34, 81, 85], building upon earlier GAN-based methods [26, 39, 60, 121] for style transfer and generation. Recent works focus on text guidance [2, 7, 31, 56, 58, 71, 83], introduce semantic and latent constraints [52, 67, 105], or unify visual and textual editing [63, 109]. Concurrent improvements in diffusion training/inference [43, 101] yield higher fidelity. Beyond static images, further approaches incorporate actions or multi-frame sequences [29, 92], refine prompts for action changes [18, 87], or embed state evolution [112]. Unlike SELDOM, few prompt-based editing works focus on precise 3D-aware control [59] or extract verbs for composing with object-centric representations.

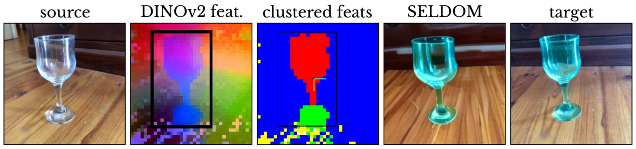
Related efforts use pretrained diffusion models for editing without large-scale fine-tuning. Recent work explores image/denoising deltas [32, 66], localizes regions from unconditional–conditional differences [19], masks cross-attention to avoid attribute leakage [65], and transfers transformations from exemplar pairs [14, 68, 88, 119]. Others retrieve pivot embeddings for real-image inversions [4, 44, 61, 93] or preserve structures via optimization of timesteps/noise and self-attention features [12, 62]. These works extract interesting representations for edits, but are often limited to changes that fit the granularity of text prompts the models were trained with, unlike the composable neural verb representations of SELDOM.

**Personalization and Identity Preservation.** Preserving subject identity and composing multiple concepts is also important. Recent works preserve subject identity and compose multiple concepts from references [3, 48, 57, 108, 120], including fine-grained object- or part-level personalization [20, 77, 111]. Many personalization works preserve identity like SELDOM, but lack precise control over the generated 3D content [76] (or are limited to a single entity).

**Object-centric and Multi-Subject Editing.** To organize scenes at an object level, prior work aligns text tokens with segments for multi-object generation [28, 35, 102, 110], handles object insertions or replacements [69, 91, 95], or performs region-based manipulation of objects [10, 15, 25, 49]. However, unlike SELDOM, these multi-object generation methods often lack identity preservation or editability.

**Scene Factorization and Compositional Editing.** Earlier object-factorization methods [8, 22, 27, 55] have inspired slot-based diffusion architectures [40, 41, 72, 106] and concept-disentangling techniques [51, 84], with further extensions in unsupervised or contrastive learning [11, 16, 23, 74, 75, 78, 94, 118]. Slot based approaches are inherently compositional, but most don't focus on learning composable *neural verb* representations like SELDOM.

**3D-Aware Synthesis and Camera Control.** Methods have adapted 2D diffusion priors to 3D without explicit 3D training data [13, 53, 54, 76, 79, 100], often guided by geome-



“move the cup to (0.45, 0.44) and rotate by (-0.02°, -10.44°, -7.03°), color greener by 30.00. camera dolly left by 0.11, down by 0.01, back by 0.01, orient right by 27.89°, up 1.25°, right 2.52°.”

Figure 4. SELDOM Implicit output generated using only an input image and the text prompt shown (no bounding boxes – visual features are clustered and attend to the prompt).

try critics or multi-view consistency [9, 17, 97]. Additional approaches focus on camera orientation [24, 37, 64], re-rendering or camera transforms [30, 99, 107], or lifting 2D features into 3D for downstream tasks [113]. Yet 3D-aware methods often don’t support a wide variety of actions and/or lack the ability to control the camera—SELDOM aims to bridge this gap.

**Mechanistic Interpretability.** Works dissect the internals of diffusion, revealing how attention layers encode structure and object identity [50], measuring text–image faithfulness [38], and collecting real-world edits from reddit [89]. Others formalize composition or interpret classifier-free guidance [5, 6], and align or compare diffusion with advanced vision-language models [42, 47, 103, 104, 114]. These works enlighten us about the internals of diffusion-based image-generation, but unlike SELDOM they generally don’t evaluate on real world, non-toy datasets.

## 3. Method

SELDOM replaces text-based conditioning in diffusion with a sequence of learnt “neural nouns and verbs”. Our goal is an object-centric representation that supports general text-based editing, e.g., “make the horse a camel”, while supporting precise 3D-aware object/camera viewpoint edits, e.g., “also rotate the camera 21.3 degrees clockwise”. We extract visual features from input images to represent objects and the background, and then apply actions (verbs) to them. A sequence of extracted features and actions is used to condition image generation. We train using pairs of before and after images, coupled with a text prompt that describes the change, e.g., “rotate the car by 35 degrees”.

**Architecture.** We fine-tune a pre-trained Stable Diffusion 2.1 [81] model to accept neural nouns and verbs as conditioning instead of CLIP [80] text encodings for a prompt.

**Neural Nouns.** A neural noun represents an object in the source image(s). We extract visual features (using DINOv2 [73]) from object bounding boxes (or feature clustering) and use region-of-interest (ROI) pooling to reduce the spatially variable number of $K = 384$ dimensional features down to a $2 \times 2 \times K$ grid, which we use as a neural noun.

**Neural Verbs.** A neural verb is a conceptual representation of the edit to be performed. We build neural verbs in one of two ways. *Explicit* neural verbs involve directly parsing the text to build verbs. Explicit DINOv2 verbs are a concatenation of cached delta visual features from the training set, a magnitude, and a bounding box. Building a neural verb starts by explicitly parsing the verb and magnitude from a prompt using spaCy [36], e.g., “rotate” and “35 degrees” from “rotate the car by 35 degrees”. Next, the text string for the verb is used as a key into a pre-computed database that fetches the median delta neural noun for that verb from the training set to use as a visual representation of that verb, a $2 \times 2 \times K$ vector, where $K$ is the DINOv2 feature dimension. The flattened delta neural noun, magnitude, and bounding box are concatenated and projected to form the neural verb. We also parse neural verbs *implicitly*, with an example show in Figure 4. Implicit DINOv2 neural verbs are more promising, as they are more general than parsing prompts. Implicit neural verbs start by using a linear layer to project the neural noun to the dimensionality of a frozen text encoder (either CLIP [80] or T5 [70], before using it as a query for multi-head attention while the text encoder output serves as keys and values.

**Background and Camera.** To compare to Neural Assets [107] with our explicit neural verbs, we first create a neural noun to represent the background of the image by pooling the remaining DINO features down to a $2 \times 2 \times K$ grid. We then project the delta camera pose as the associated neural verb for this background neural noun. For implicit neural verbs, the pooled background features projected to the text dimensionality attend to a text prompt describing the camera change.

**Conditioning.** The model is conditioned on a sequence of tokens. Each neural noun represents an entity in the image, with an additional entry for the scene background. Composing entities explicitly as individual tokens is a powerful break from the contextual text encodings of prompts. Our neural nouns are size $4(X + 1) \times 1024$, where X is the number of entities in the image. The additional 1 describes camera changes, and the 4 represents the flattened $2 \times 2$ grid of DINOv2 features. The sequence of neural verbs is a $(X + 1) \times 1024$ sequence of tokens. The neural verbs are then repeated 4 times to a final size of $4(X + 1) \times 1024$ to correspond with the 4 DINOv2 features per neural noun. The 2048 dimensional sequence of paired and concatenated nouns and verbs is projected to the dimensionality of the CLIP text encoder output and used in lieu of it.

**Inference.** At test time, only a source image and prompt are needed. SELDOM supports multiple source image(s) and general text-editing prompts. To build neural nouns, we either cluster the fine-tuned DINOv2 features, use an object detector, or use ground truth bounding boxes. To build neural verbs, we parse the prompt for the verb, magnitude, and

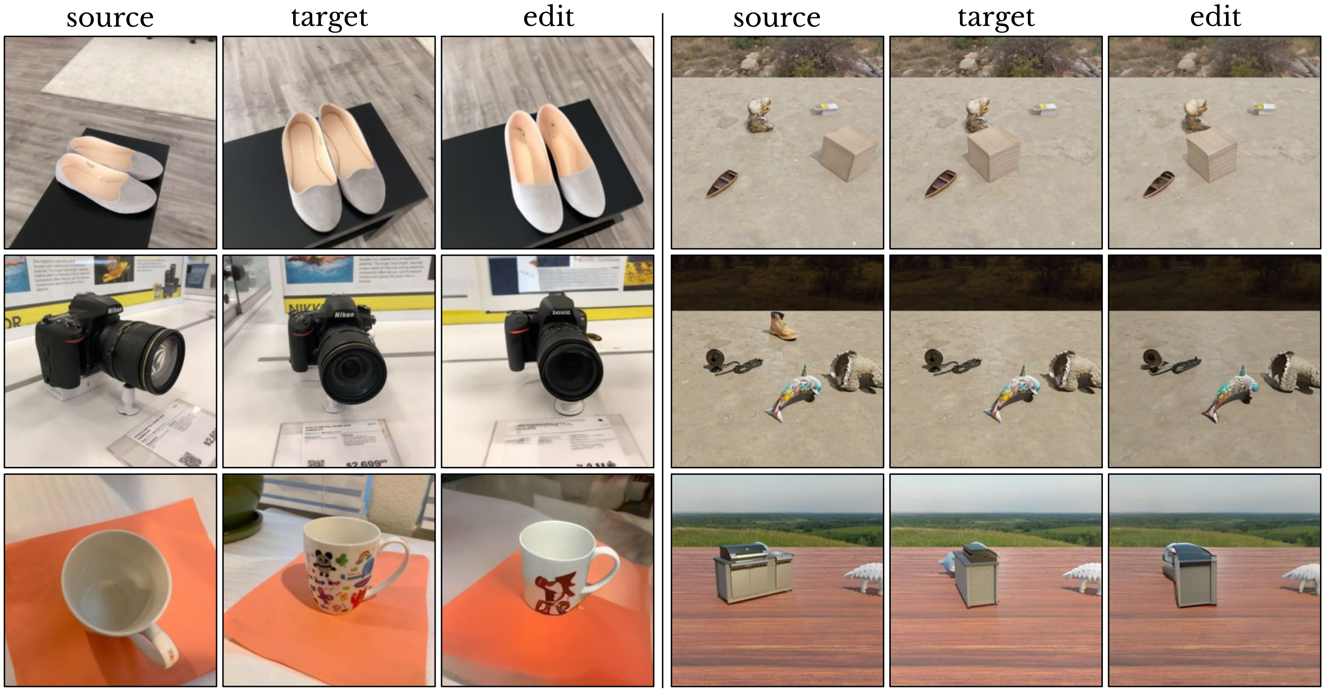


Figure 5. **SELDOM Outputs**: Triplets of source, target, and SELDOM editing result (one model produces all outputs). The left three rows show triplets that all include the move, rotate, and camera motion verbs for the Objectron [1] dataset: (top-left) a rotated pair of shoes, with an inferred shelf bottom; (middle-left) a rotation to the front of the camera, successfully darkening the transparent lens and generating a plausible backcard for the camera; (bottom-left) a top-view of a mug without the side visible, and SELDOM acceptably hallucinates a dragon. The right three rows show triplets from the OBJect [59] dataset with translation, removal, and rotation edits respectively: (top-right) a box is shifted left; (middle-right) a boot is removed; (bottom-right) a grill is rotated, revealing a hidden object that was occluded.

desired 2D bounding box and compose them with retrieved delta DINOv2 features for the verb. Alternatively, our implicit method projects DINOv2 features as queries against the frozen text encoder output.

**Training.** A training sample is a source image, target image, and a text prompt describing the change to be performed. Our explicit variant of SELDOM requires source and target bounding boxes for entities.

**Classifier-free Guidance for Editing.** Classifier-free guidance [33] for image generation typically involves combining an unconditional and conditional denoising step. This combination is purported to both denoise towards the distribution of natural images and the direction of the conditioning. For compositional image editing compared with generation, it is unclear if this same intuition is best applied. We explore three variants of classifier-free guidance. The first is normal classifier-free guidance, with a guidance factor of 2.0. The second is an "edit" approach in which, with probability $p = 0.1$, we both change the target to be the source image and drop the neural verbs (but not the neural nouns) from the conditioning. This encourages the network to learn that neural nouns represent the current state and setting neural verbs to zeros should be a no-op. The third variant, dubbed "raw-edit", drops neural verb conditioning similarly, but is not classifier-free guidance because it only uses conditional generation.

**Loss.** Latent diffusion [81] encodes images $\mathbf{x}_0$ into latent representations $\mathbf{z}_0$ using an autoencoder $E$:

$$\mathbf{z}_0 = E(\mathbf{x}_0)$$

where $\mathbf{x}_0$ is the original image and $\mathbf{z}_0$ its latent encoding.

Then, at timestep $t$, a noisy latent $\mathbf{z}_t$ is defined by:

$$\mathbf{z}_t = \alpha_t \mathbf{z}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \text{with} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}),$$

where scalar coefficients $\alpha_t$ and $\sigma_t$ control the amount of signal and noise according to a predetermined schedule.

We fine-tune the diffusion model using $v$-prediction [82], predicting the velocity vector: $\mathbf{v} = \alpha_t \mathbf{z}_0 - \sigma_t \boldsymbol{\epsilon}$.

The training objective is the mean squared error (MSE) between the predicted velocity $v_\theta(\mathbf{z}_t, t, c)$ and the true velocity $\mathbf{v}$:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0, c, t, \boldsymbol{\epsilon}} \left[ \| v_\theta(\mathbf{z}_t, t, c) - \mathbf{v} \|^2 \right],$$

where $v_\theta$ is the neural network parameterized by weights $\theta$, and $c$ denotes conditioning prompt embeddings. The expectation is taken over batches of input images $\mathbf{x}_0$, conditioning embeddings $c$, timesteps $t$, and noise realizations $\boldsymbol{\epsilon}$.

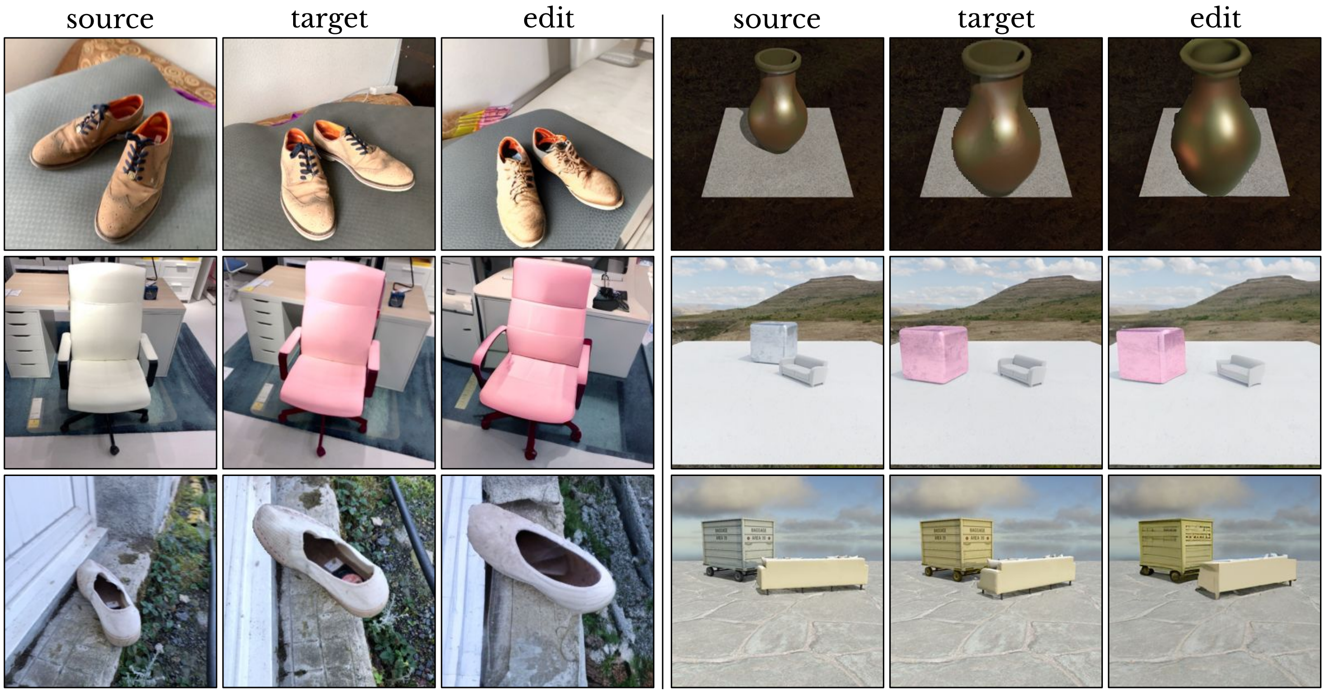


Figure 6. **Photo-Editing:** Existing actions from source target pairs are augmented with photo-edits on the Objectron [1] (three left columns) and OBJect [59] (three right columns) datasets. The Objectron triplets show: (top-left) a shoe is brightened and the camera orbits, (middle-left) a white chair is turned pink and the camera is moved, (bottom-left) a shoe on a doorstep is scaled up as the camera approaches. The OBJect triplets show: (top-right) a simple scale-up, (middle-right) a silver cube made pink and moved forward and to the left, (bottom-right) a box cart turned yellow.

**Hardware and Training Details.** We fine-tune models for 1-2 days and 100,000 steps, similar to Neural Assets [107]. We use NVIDIA A40, A100, and H100 GPUs. We use batch sizes ranging from 2 to 50. We use a learning rate of $1 \times 10^{-4}$ and optimize the model using the Schedule-Free [21] LR adjuster with an Adam optimizer [45].

## 4. Experiments

SELDOM both follows precise instructions, e.g., "rotate the car 35.3 degrees", and performs general text-based image edits, e.g., "paint the door red". SELDOM can compose objects across scenes and edit photos using verbs.

### 4.1. Datasets

We train on the Objectron [1] and OBJect [59] datasets, as they have granular prompts, e.g., "rotate the hat by 35°."

**Objectron.** [1] consists of short video clips of common objects like chairs and tables, annotated with 3D bounding boxes and camera poses. It provides explicit knowledge of 3D camera extrinsics, allowing for precise conditioning based on camera movements. Additionally, Objectron includes multiple verbs that affect the same object, e.g., move to location and rotate, allowing us to better test SELDOM.

**OBJect.** [59] is a dataset comprising 100,000 procedurally generated synthetic scenes, each containing 1 to 4 objects, designed for 3D-aware image editing tasks such as rotation, removal, insertion, and translation. This dataset includes multiple objects in a scene, with one edited per sample, allow disentangled neural verb learning.

**Pseudolabels.** Using SAM-generated segmentation masks for objects in both datasets, we create modified target images with different hue, scale, and brightness, to provide a greater diversity of "image editing" targets.

### 4.2. Metrics

Similar to prior work in image generation, we evaluate our method using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [98], and Learned Perceptual Image Patch Similarity (LPIPS) [116]. PSNR measures reconstruction quality between the generated and target images. SSIM evaluates perceptual similarity between images. LPIPS measures perceptual similarity using deep features. For a fair comparison with Neural Assets [107], we evaluate within object bounding boxes.

### 4.3. Results

**Quantitative Results.** Table 1 evaluates SELDOM against three baselines from the state-of-the-art Neural As-

Table 1. Quantitative comparison on the OBJect unseen split within bounding boxes for translation, rotation, and removal tasks. ↑ indicates higher is better, ↓ indicates lower is better. SELDOM performs on-par or better than Neural Assets [107].

| **Method** | **PSNR ↑** | | | **SSIM ↑** | | | **LPIPS ↓** | | |
|---|---|---|---|---|---|---|---|---|---|
| | Translate | Rotate | Remove | Translate | Rotate | Remove | Translate | Rotate | Remove |
| Chained [107] | 14.1 | 12.9 | 12.1 | 0.33 | 0.27 | 0.38 | 0.47 | 0.54 | 0.46 |
| 3DIT [59] | 15.2 | 16.3 | 24.7 | 0.29 | 0.37 | 0.57 | 0.48 | 0.45 | 0.26 |
| Neural Assets [107] | 20.1 | **18.4** | 28.4 | 0.43 | 0.38 | 0.61 | 0.27 | 0.37 | 0.17 |
| SELDOM (Ours) | **20.6** | 17.6 | **31.0** | **0.52** | **0.40** | **0.77** | **0.07** | **0.16** | **0.03** |

Table 2. Quantitative comparison on the Objectron validation split within bounding boxes for the joint translation and rotation task. ↑ indicates higher is better, ↓ indicates lower is better. Neural Assets [107] uses 3D bboxes but SELDOM only uses 2D bboxes.

| **Method** | **PSNR ↑** | **SSIM ↑** | **LPIPS ↓** |
|---|---|---|---|
| Chained [107] | 11 | 0.1 | 0.4 |
| 3DIT [59] | 11 | 0.1 | 0.4 |
| Neural Assets [107] | **16** | **0.45** | **0.15** |
| SELDOM (Ours) | 14.4 | 0.35 | 0.25 |

Table 3. SELDOM ablations evaluated on a subset of OBJect [59] seen and Objectron val datasets. We compare Δ VAE latents vs. Δ DINOv2 features in explicit neural verbs. Next, we compare slot-based vs. DINOv2 feature *implicit* neural verbs. Then we compare slot-based vs. DINOv2 feature *explicit* neural verbs. Finally, we compare variations of classifier-free guidance.

| **Conditioning** | **PSNR ↑** | **SSIM ↑** | **LPIPS ↓** |
|---|---|---|---|
| Explicit VAE | 22.3 | 0.59 | 0.12 |
| Explicit DINOv2 | **25.9** | **0.66** | **0.09** |
| Implicit Slot | 20.2 | **0.53** | 0.13 |
| Implicit DINOv2 | **20.6** | 0.51 | **0.12** |
| Explicit Slot | 20.0 | 0.46 | 0.16 |
| Explicit DINOv2 | **25.9** | **0.66** | **0.09** |
| Imp. DINOv2 Normal | 20.1 | 0.49 | **0.12** |
| Imp. DINOv2 Edit | 20.0 | 0.47 | 0.14 |
| Imp. DINOv2 Raw | **20.6** | **0.51** | **0.12** |

sets [107]. We perform on par with or better than Neural Assets, improving PSNR, SSIM, and LPIPS on the OBJect [59] unseen subset, except for on PSNR rotation. Compellingly, SELDOM does not make use of 3D bounding boxes like Neural Assets does, working well without the canonical orientation implied by such bounding boxes. By avoiding 3D bounding boxes, SELDOM is capable of training and inference on any image dataset that has object detections (rather than needing 3D bounding box annotations).

**Qualitative Results.** Figure 2 shows the object compositionality capabilities of SELDOM using image triplets (scene, object, obj. in scene) with two source images and SELDOM's generated output placing the object in the scene. These triplets show SELDOM is capable of inter-dataset edits, as well as meaningful 3D-aware positioning (including occlusion). Unlike most image-editing methods, SELDOM is capable of jointly composing image and camera changes, as Figure 2 shows. No target features are available for the changed viewpoint; the model has simply learnt how to rotate real world Objectron scenes. Figure 5 shows results from SELDOM on the Objectron and OBJect datasets that directly compare to Neural Assets, in that they don't mix-and-match objects across datasets or apply general verbs (as SELDOM is capable of). Figure 6 introduces SELDOM doing more than just manipulate the poses of objects. SELDOM's general verb-based approach enables photo editing capabilities, including color changes, scaling, and other edits on the Objectron (three left columns) and OBJect (right three columns) datasets. Meanwhile, Figure 7 shows failures of SELDOM, while Figure 8 shows how SELDOM can be applied to other image editing datasets.

**Variations.** Table 3 explores SELDOM variants. First, we compare the use of Δ VAE latents vs Δ DINOv2 features for explicit neural verbs. The comparison explores the visual representation's value in capturing the verb-related delta visual features (from a cached median lookup from the training dataset) between the source and target latent diffusion VAE features or the DINOv2 features. This reveals whether the DINOv2 features improve the ability for the model to recognize the verb being applied, and we only show modest improvements, demonstrating that identifying the verb is not challenging. Next, we compare implicit methods that attend to text prompts, specifically a slot-based method vs. a DINOv2 feature-based *implicit* neural verb formulation. The slot-based variant involves initializing $n = 5$ slots of the same dimensionality as the text encoder. Each slot attends to the text prompt and is matched to a neural noun based on similarity. Promisingly, our slot-based variant forms a verb representation *solely from text*, separating concerns of edits and state. We achieve mixed

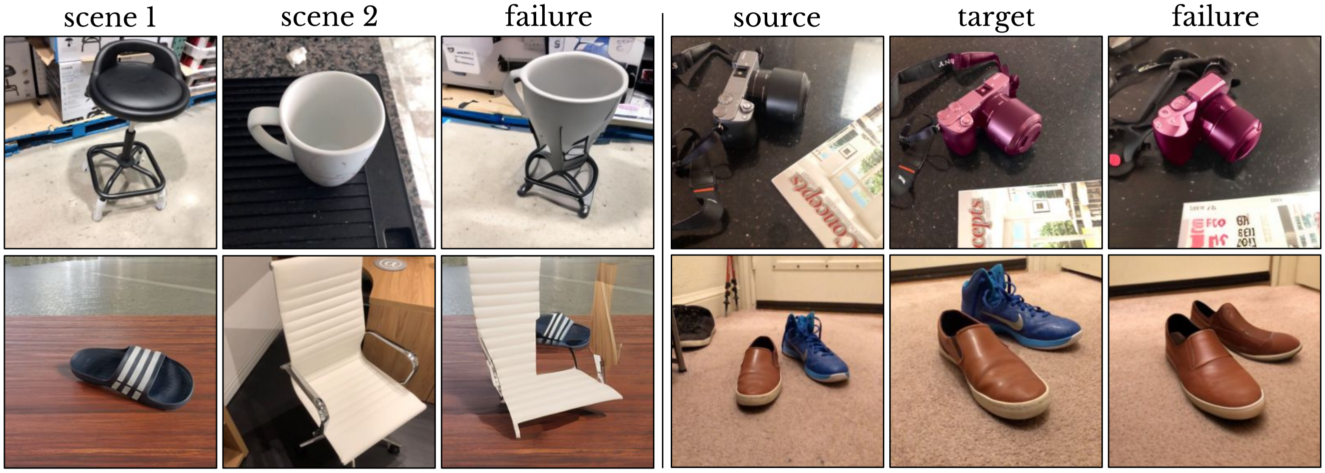
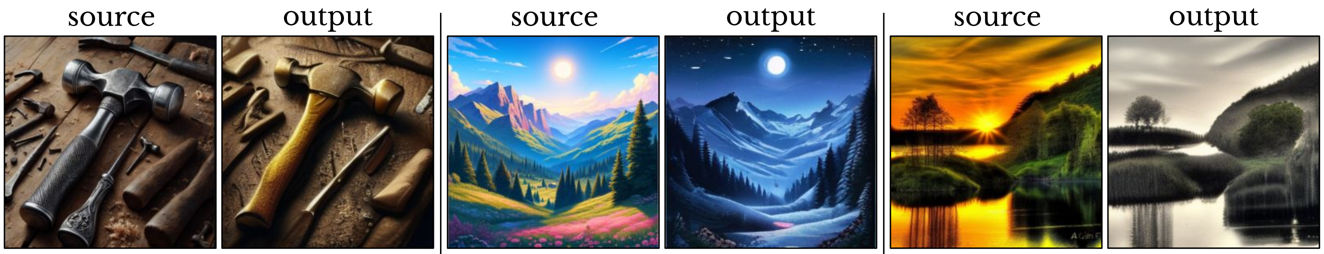


Figure 7. **Failure Cases:** The left three columns show object compositionality failures, the right three columns show different failure modes: (top-left) an object composition successfully rotates the chair, mug, and camera but creates a hybrid chair-mug due to recomposing them in the same location; (bottom-left) a cross-dataset failure from inserting a chair into a scene with an artifact behind the chair; (top-right) a synthetically generated colored target is wrong but SELDOM correctly colors only the camera body; (bottom-right) the model has generally seen paired shoes, and is tricked by two mismatched shoes, turning the blue shoe into a brown one.



"change the handle to brown wood, color yellower by 10, brighten 1.19, scale 1.35."

"transition day to night, add a twilight sky... stars in the sky with a full moon...add highlights to suggest bioluminescent flora...color bluer by 10."

"change to a winter scene."

Figure 8. **SELDOM on Image-Editing Datasets:** We train SELDOM on multiple image-editing datasets and show results: (left-pair) a hammer from the HQEdit dataset is edited, augmented with our synthetic verbs of yellowing, brightening, and scaling, (middle-pair) a scene-wide transition, showing SELDOM is capable of operating with non object-centric prompts, (right-pair) a similar scene change.

performance, showing that both verb slots and implicit dino queries can learn to model verbs. We then compare explicit slot-based and dino-based verb formulations. The explicit slot-based variant parses the prompt for verbs, projects them to queries, and attends to the full text prompt. We compare to the explicit DINOv2 version as outlined in the neural verb description of Section 3. Finally, we compare three means of applying classifier-free guidance, "normal", "edit", and "rawedit" as described in Section 3. We adopt "rawedit" for implicit DINOv2 verbs.

**Testing SELDOM on Image Editing Datasets.** To demonstrate the broad applicability of SELDOM, we train a variant across multiple general image-editing datasets, specifically: HQ-Edit [96], MagicBrush [115], HIVE [117], Instruct-Pix2Pix [7], AURORA [46], COIN [90], ChangeIt [86], and GenHowTo [87]. We demonstrate the generalizability of SELDOM in Figure 8.

## 5. Conclusion

We introduce SELDOM for object-centric image editing. We first decompose a scene into objects, extract and apply verbs to objects, and finally recompose them in a scene. Our approach eschews the paradigm of single-image conditioning to instead compose objects, object-centric verb-based edits, and scene/camera control across datasets. We find a single fine-tuned Stable Diffusion 2.1 model is capable of both general text edits such as "make the camera pink" and "rotate the camera 23.4 degrees", as well as inserting real objects into synthetic scenes and vice-versa.

## References

[1] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7822–7831, 2021. 5, 6

[2] Omri Avrahami, Tali Dekel Cohen, and Dani Lischinski. Blended diffusion for text-driven editing of natural images. *arXiv preprint arXiv:2111.14818*, 2022. 1, 3

[3] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023. 3

[4] Manuel Brack, Felix Friedrich, Katharina Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinario Passos. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8861–8870, 2024. 3

[5] Arwen Bradley and Preetum Nakkiran. Classifier-free guidance is a predictor-corrector. *arXiv preprint arXiv:2408.09000*, 2024. 4

[6] Arwen Bradley, Preetum Nakkiran, David Berthelot, James Thornton, and Joshua M. Susskind. Mechanisms of projective composition of diffusion models. *arXiv preprint arXiv:2502.04549*, 2025. 4

[7] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. 1, 3, 8

[8] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matthew Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019. 3

[9] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22560–22570, 2023. 4

[10] Tuhin Chakrabarty, Kanishk Singh, Arkadiy Saakyan, and Smaranda Muresan. Learning to follow object-centric image editing instructions faithfully. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9630–9646, 2023. 3

[11] David M. Chan, Rodolfo Corona, Joonyong Park, Cheol Jun Cho, Yutong Bai, and Trevor Darrell. Analyzing the language of visual tokens. *arXiv preprint arXiv:2411.05001*, 2024. 3

[12] Sherry X. Chen, Yaron Vaxman, Elad Ben Baruch, David Asulin, Aviad Moreshet, Kuo-Chin Lien, Misha Sra, and Pradeep Sen. TiNO-Edit: Timestep and noise optimization for robust diffusion-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6108–6117, 2024. 3

[13] Xinyu Chen, Zexiang Zhang, Zhaoxi Zhang, Zhen Li, and Dahua Lin. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023. 3

[14] Xi Chen, Yutong Feng, Mengting Chen, Yiyang Wang, Shilong Zhang, Yu Liu, Yujun Shen, and Hengshuang Zhao. Zero-shot image editing with reference imitation. *arXiv preprint arXiv:2406.07547*, 2024. 3

[15] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6593–6602, 2024. 3

[16] Yinbo Chen, Rohit Girdhar, Xiaolong Wang, Sai Saketh Rambhatla, and Ishan Misra. Diffusion autoencoders are scalable image tokenizers. *arXiv preprint arXiv:2501.18593*, 2025. 3

[17] Ta-Ying Cheng, Matheus Gadelha, Thibault Groueix, Matthew Fisher, Radomir Mech, Andrew Markham, and Niki Trigoni. Learning continuous 3d words for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6753–6762, 2024. 4

[18] Hyungjin Chung, Dohun Lee, and Jong Chul Ye. Acdc: Autoregressive coherent multimodal generation using diffusion correction, 2024. 3

[19] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 3

[20] Aleksandar Cvejic, Abdelrahman Eldesokey, and Peter Wonka. Partedit: Fine-grained image editing using pre-trained diffusion models. *arXiv preprint arXiv:2502.04050*, 2025. 3

[21] Aaron Defazio, Xingyu Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The road less scheduled, 2024. 6

[22] Dave Epstein, Taesung Park, Richard Zhang, Eli Shechtman, and Alexei A. Efros. BlobGAN: Spatially disentangled scene representations. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[23] Ke Fan, Zechen Bai, Tianjun Xiao, Tong He, Max Horn, Yanwei Fu, Francesco Locatello, and Zheng Zhang. Adaptive slot attention: Object discovery with dynamic slot number. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23062–23071. IEEE, 2024. 3

[24] Daniel Geng and Andrew Owens. Motion guidance: Diffusion-based image editing with differentiable motion estimators. *arXiv preprint arXiv:2401.18085*, 2024. 4

[25] Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Xingqian Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair diffusion: A comprehensive multimodal object-level image editor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8609–8618, 2024. 3

[26] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 3

[27] Klaus Greff, Raphael L Kaufman, Rishabh Kabra, Nicholas Watters, Christopher P Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2424–2433, 2019. 3

[28] Qin Guo and Tianwei Lin. Focus on your instruction: Fine-grained and multi-instruction image editing by attention modulation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 6986–6996. IEEE, 2024. 3

[29] Yujun Guo, Ziyu Zhang, Jianmin Zhang, Dong Chen, Lu Yuan, and Fang Wen. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3

[30] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 4

[31] Amir Hertz, Ron Mokady, Tomer Tenenbaum, Kfir Aberman, Or Perel, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1, 3

[32] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2328–2337, 2023. 3

[33] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[34] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

[35] Jiun Tian Hoe, Xudong Jiang, Chee Seng Chan, Yap-Peng Tan, and Weipeng Hu. Interactdiffusion: Interaction control in text-to-image diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 6180–6189. IEEE, 2024. 3

[36] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017. 4

[37] Chen Hou, Guoqiang Wei, Yan Zeng, and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024. 4

[38] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 4

[39] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 3

[40] Jindong Jiang, Fei Deng, Gautam Singh, and Sungjin Ahn. Object-centric slot diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8563–8601, 2023. 3

[41] Jindong Jiang, Krishnakant Singh, Simone Schaub-Meyer, and Stefan Roth. Object-centric slot diffusion. *arXiv preprint arXiv:2303.10834*, 2023. 3

[42] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, Oriane Siméoni, Huy V. Vo, Patrick Labatut, and Piotr Bojanowski. Dinov2 meets text: A unified framework for image- and pixel-level vision-language alignment. *arXiv preprint arXiv:2412.16334*, 2024. 4

[43] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 3

[44] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6007–6017, 2023. 3

[45] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[46] Benno Krojer, Dheeraj Vattikonda, Luis Lara, Varun Jampani, Eva Portelance, Christopher Pal, and Siva Reddy. Learning action and reasoning-centric image editing from videos and simulations. *arXiv preprint arXiv:2407.03471*, 2024. 8

[47] Martha Lewis, Nihal Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen Bach, and Ellie Pavlick. Does CLIP bind concepts? probing compositionality in large image models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1487–1500, St. Julian's, Malta, 2024. Association for Computational Linguistics. 4

[48] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8640–8650, 2024. 3

[49] Yuanze Lin, Yi-Wen Chen, Yi-Hsuan Tsai, Lu Jiang, and Ming-Hsuan Yang. Text-driven image editing via learnable regions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[50] Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7817–7826, 2024. 4

[51] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel,*

*October 23–27, 2022, Proceedings, Part XVII*, pages 423–439. Springer, 2022. 1, 3

[52] Xihui Liu, Yixuan Zhang, Zhijie Zhang, Jun He, Niloy J Mitra, Chunhua Wang, Philip HS Torr, Li Zhang, and Shuai Zheng. More control for free! image synthesis with semantic diffusion guidance. *arXiv preprint arXiv:2112.05744*, 2021. 3

[53] Xingang Liu, Zexiang Lin, Jiacheng He, Xiaohui Wang, Zhe Li, Zhijian Zhang, Ke Zeng, Zhen Li, Hongsheng Li, and Dahua Lin. Zero-1-to-3: Zero-shot one image to 3d object. *arXiv preprint arXiv:2303.11328*, 2023. 3

[54] Zhen Liu, Yao Feng, Michael J. Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu. Meshdiffusion: Score-based generative 3d mesh modeling. In *International Conference on Learning Representations*, 2023. 3

[55] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, pages 11525–11538, 2020. 3

[56] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. *arXiv preprint arXiv:2201.09865*, 2022. 3

[57] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 3

[58] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 3

[59] Oscar Michel, Anand Bhattad, Eli VanderBilt, Ranjay Krishna, Aniruddha Kembhavi, and Tanmay Gupta. Object 3dit: Language-guided 3d-aware image editing. In *Advances in Neural Information Processing Systems*, 2023. 3, 5, 6, 7

[60] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3

[61] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047, 2023. 3

[62] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[63] Jiteng Mu, Nuno Vasconcelos, and Xiaolong Wang. Editar: Unified conditional generation with autoregressive models. *arXiv preprint arXiv:2501.04699*, 2025. 3

[64] Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Multidiff: Consistent novel view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10258–10268, 2024. 4

[65] Sunung Mun, Jinhwan Nam, Sunghyun Cho, and Jungseul Ok. Addressing attribute leakages in diffusion-based image editing without training. *arXiv preprint arXiv:2412.04715*, 2024. 3

[66] Hyelin Nam, Gihyun Kwon, Geon Yeong Park, and Jong Chul Ye. Contrastive denoising score for text-guided latent diffusion image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9192–9201, 2024. 3

[67] Hyelin Nam, Jaehoon Lee, and Seungryong Kim. Contrastive denoising score for text-guided latent diffusion image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[68] Thao Nguyen, Yuheng Li, Utkarsh Ojha, and Yong Jae Lee. Visual instruction inversion: Image editing via image prompting. *Advances in Neural Information Processing Systems*, 36:9598–9613, 2023. 3

[69] Trong-Tung Nguyen, Duc-Anh Nguyen, Anh Tran, and Cuong Pham. Flexedit: Flexible and controllable diffusion-based object-centric image editing. *arXiv preprint arXiv:2403.18605*, 2024. 3

[70] Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*, 2021. 4

[71] Alexander Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[72] Weili Nie, Sifei Liu, Morteza Mardani, Chao Liu, Benjamin Eckart, and Arash Vahdat. Compositional text-to-image generation with dense blob representations. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. 3

[73] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4

[74] Core Francisco Park, Maya Okawa, Andrew Lee, Ekdeep Singh Lubana, and Hidenori Tanaka. Emergence of hidden capabilities: Exploring learning dynamics in concept space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[75] Jicheol Park, Dongwon Kim, Boseung Jeong, and Suha Kwak. PLOT: Text-based person search with part slot attention for corresponding part discovery. In *Computer Vision – ECCV 2024, Part XXI*, pages 474–490. Springer, 2024. 3

[76] Gaurav Parmar, Jong-Chyi Park, Jun-Yan Zhu, Alexei A Efros, and Richard Zhang. Zero-shot text-guided object generation with dream fields. *arXiv preprint arXiv:2209.12208*, 2023. 3

[77] Or Patashnik, Rinon Gal, Daniil Ostashev, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. Nested attention: Semantic-aware attention values for concept personalization. *arXiv preprint arXiv:2501.01407*, 2025. 3

[78] Thinh Pham, Chi Tran, and Dat Quoc Nguyen. MISCA: A joint model for multiple intent detection and slot filling with intent-slot co-attention. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12641–12650, Singapore, 2023. Association for Computational Linguistics. 3

[79] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2023. 3

[80] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 2, 4

[81] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1, 2, 3, 4, 5

[82] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models, 2022. 5

[83] Artur Shagidanov, Hayk Poghosyan, Xinyu Gong, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Grounded-instruct-pix2pix: Improving instruction based image editing with automatic target grounding. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6585–6589. IEEE, 2024. 3

[84] Junjie Shentu, Matthew Watson, and Noura Al Moubayed. Attencraft: Attention-guided disentanglement of multiple concepts for text-to-image customization. *arXiv preprint arXiv:2405.17965*, 2024. 3

[85] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[86] Tomáš Souček, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic. Look for the change: Learning object states and state-modifying actions from untrimmed web videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13956–13966, 2022. 8

[87] Tomáš Souček, Dima Damen, Michael Wray, Ivan Laptev, and Josef Sivic. Genhowto: Learning to generate actions and state transformations from instructional videos, 2024. 3, 8

[88] Ashutosh Srivastava, Tarun Ram Menta, Abhinav Java, Avadhoot Jadhav, Silky Singh, Surgan Jandial, and Balaji Krishnamurthy. Reedit: Multimodal exemplar-based image editing with diffusion models. *arXiv preprint arXiv:2411.03982*, 2024. 3

[89] Peter Sushko, Ayana Bharadwaj, Zhi Yang Lim, Vasily Ilin, Ben Caffee, Dongping Chen, Mohammadreza Salehi, Cheng-Yu Hsieh, and Ranjay Krishna. REALEDIT: Reddit edits as a large-scale empirical dataset for image transformations. *arXiv preprint arXiv:2502.03629*, 2025. 4

[90] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 8

[91] Yoad Tewel, Rinon Gal, Dvir Samuel, Yuval Atzmon, Lior Wolf, and Gal Chechik. Add-it: Training-free object insertion in images with pretrained diffusion models. *arXiv preprint arXiv:2411.07232*, 2024. 3

[92] Maria Mihaela Trusca, Mingxiao Li, and Marie-Francine Moens. Action-based image editing guided by human instructions. *arXiv preprint arXiv:2412.04558*, 2024. 3

[93] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, 2023. 3

[94] Mostofa Rafid Uddin and Min Xu. Dualcontrast: Unsupervised disentangling of content and transformations with implicit parameterization. *arXiv preprint arXiv:2405.16796*, 2024. 3

[95] Kai Wang, Fei Yang, Shiqi Yang, Muhammad Atif Butt, and Joost van de Weijer. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3

[96] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds, 2024. 8

[97] Ruicheng Wang, Jianfeng Xiang, Jiaolong Yang, and Xin Tong. Diffusion models are geometry critics: Single image 3d editing using pre-trained diffusion priors. In *European Conference on Computer Vision*, pages 441–458. Springer, 2024. 4

[98] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[99] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. *arXiv preprint arXiv:2312.03641*, 2023. 4

[100] Zhaoyang Wang, Zexiang Zhang, Zhaoxi Zhang, Zhen Li, and Dahua Lin. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. *arXiv preprint arXiv:2304.00503*, 2023. 3

[101] Zhendong Wang, Yifan Jiang, Huangjie Zheng, Peihao Wang, Pengcheng He, Zhangyang Wang, Weizhu Chen, Mingyuan Zhou, et al. Patch diffusion: Faster and more

data-efficient training of diffusion models. *Advances in neural information processing systems*, 36, 2024. 3

[102] Zirui Wang, Zhizhou Sha, Zheng Ding, Yilin Wang, and Zhuowen Tu. Tokencompose: Text-to-image diffusion with token-level supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8553–8564, 2024. 3

[103] Hadi Wazni, Kin Ian Lo, and Mehrnoosh Sadrzadeh. Verbclip: Improving verb understanding in vision-language models with compositional structures. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 195–201, Bangkok, Thailand, 2024. Association for Computational Linguistics. 4

[104] Chen Wu and Fernando De la Torre. Contrastive prompts improve disentanglement in text-to-image diffusion models. *arXiv preprint arXiv:2402.13490*, 2024. 4

[105] Chen Henry Wu and Fernando De la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7378–7387, 2023. 3

[106] Ziyi Wu, Jingyu Hu, Wuyue Lu, Igor Gilitschenski, and Animesh Garg. Slotdiffusion: Object-centric generative modeling with diffusion models, 2023. 3

[107] Ziyi Wu, Yulia Rubanova, Rishabh Kabra, Drew A Hudson, Igor Gilitschenski, Yusuf Aytar, Sjoerd van Steenkiste, Kelsey R Allen, and Thomas Kipf. Neural assets: 3d-aware multi-object scene synthesis with image diffusion models. *arXiv preprint arXiv:2406.09292*, 2024. 1, 2, 4, 6, 7

[108] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, pages 1–20, 2024. 3

[109] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024. 3

[110] Shan Yang. MFTF: Mask-free training-free object level layout control diffusion model. *arXiv preprint arXiv:2412.01284*, 2024. 3

[111] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3

[112] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024. 3

[113] Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. *Improving 2D Feature Representations by 3D-Aware Fine-Tuning*, page 57–74. Springer Nature Switzerland, 2024. 4

[114] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa F. Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 4

[115] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems*, 2023. 8

[116] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. 6

[117] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. Hive: Harnessing human feedback for instructional visual editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9026–9036, 2024. 8

[118] Yan Zhang, David W. Zhang, Simon Lacoste-Julien, Gertjan J. Burghouts, and Cees G. M. Snoek. Unlocking slot attention by changing optimal transport costs. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 41931–41951. PMLR, 2023. 3

[119] Chenyang Zhu, Kai Li, Yue Ma, Longxiang Tang, Chengyu Fang, Chubin Chen, Qifeng Chen, and Xiu Li. Instantswap: Fast customized concept swapping across sharp shape differences. *arXiv preprint arXiv:2412.01197*, 2024. 3

[120] Jingyuan Zhu, Huimin Ma, Jiansheng Chen, and Jian Yuan. Isolated diffusion: Optimizing multi-concept text-to-image generation training-freely with isolated diffusion guidance. *IEEE Transactions on Visualization and Computer Graphics*, page 1–14, 2024. 3

[121] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. 3