An Image is Worth Multiple Words: Discovering Object Level Concepts using Multi-Concept Prompt Learning

Chen Jin [1] **Ryutaro Tanno** [2] **Amrutha Saseendran** [1] **Tom Diethe** [1] **Philip Teare** [1]

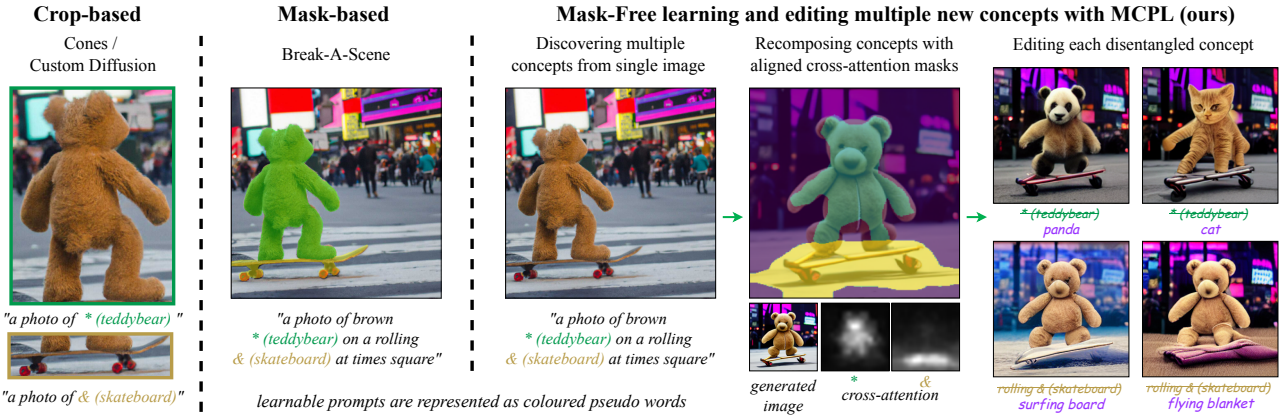

Figure 1. **Language driven multi-concepts learning and applications.** Custom Diffusion and Cones learn concepts from crops of objects, while Break-A-Scene uses masks. In contrast, our method learns object-level concepts using only image-sentence pairs, aligning the cross-attention of each learnable prompt with a semantically meaningful region, and enabling mask-free local editing.

Abstract

Textural Inversion, a prompt learning method, learns a singular text embedding for a new "word" to represent image style and appearance, allowing it to be integrated into natural language sentences to generate novel synthesised images. However, identifying multiple unknown object-level concepts within one scene remains a complex challenge. While recent methods have resorted to cropping or masking individual images to learn multiple concepts, these techniques often require prior knowledge of new concepts and are labour-intensive. To address this challenge, we introduce *Multi-Concept Prompt Learning (MCPL)*, where multiple unknown "words" are simultaneously learned from a single sentence-image pair, without any imagery annotations. To enhance the accuracy of word-concept correlation and refine attention mask boundaries, we propose three regularisation techniques: *Attention Masking*, *Prompts Contrastive Loss*, and *Bind Adjective*. Extensive quantitative comparisons with both real-world categories and biomedical images demonstrate that our method can learn new semantically disentangled concepts. Our approach emphasises learning solely from textual embeddings, using less than 10% of the storage space compared to others.

1. Introduction

Language-driven vision concept discovery is a human-machine interaction process in which the human describes an image, leaving out multiple unfamiliar concepts. The machine then learns to link each new concept with a corresponding learnable prompt (pseudo words in Figure 1) from the sentence-image pair. Such capacity would accelerate the scientific knowledge discovery process either from experimental observations or mining existing textbooks. It also facilitates hypothesis generation through local image editing without concrete knowledge of the new vision concept.

Recent research ((Gal et al., 2022; Ruiz et al., 2022)) shows that the appearance and style of an image can be encapsulated as a cohesive concept via a learned prompt ("word") optimised in the frozen embedding space of a pre-trained

[1]Centre for AI, DS&AI, AstraZeneca, UK [2]Google DeepMind, UK. Correspondence to: Chen Jin <chen.jin@astrazeneca.com>, Philip Teare <philip.teare@astrazeneca.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

text-to-image diffusion model. To bring the learning down to multiple objects in a scene, Custom Diffusion (Kumari et al., 2023) and Cones (Liu et al., 2023), make use of the crops of objects, while Break-A-Scene (Avrahami et al., 2023) uses masks, as shown in Figure 1. These approaches enhance the integration of multiple learned concepts via fine-tuning and storing the diffusion model weights. Yet, two limitations hinder the scalability of these approaches in discovering unknown concepts: 1) the imagery annotations need prior knowledge about the concept and are labour-intensive. 2) costly to update and store Diffusion Model (DM) parameters per new concept (see Table 1).

Table 1. **Competing methods.** Our method is the first to suggest a solution for discovering new visual concepts using only linguistic descriptions. Our method also benefits from storage cost efficiency. In the table, we compare the storage cost for every 1∼4 concepts.

Method	Multi-concept	Single image	Crop/Mask free	Token only	Storage cost
Textual Inversion	✗	✗	-	✓	<0.1MB
Dreambooth	✗	✗	-	✗	3.3GB
Custom Diffusion	✓	✗	✗	✗	72MB
Cones	✓	✓	✗	✗	1∼10MB
Break-A-Scene	✓	✓	✗	✗	4.9GB
MCPL (Ours)	✓	✓	✓	✓	<0.1MB

In this work, we explore learning object-level concepts using only natural language descriptions and only updating and storing the textural embedding (token). We start with a motivational study that confirms, without updating DM parameters, while applying masking or cropping yields distinct embeddings, object-level learning and editing relying solely on linguistic descriptions remains challenging. Motivated by this finding, we introduce *Multi-Concept Prompt Learning (MCPL)* Figure 3 (Top) for **mask-free text-guided learning of multiple prompts from one scene**.

However, without further assumptions on the embedding relationships, jointly learning multiple prompts is problematic. The model may disregard the semantic associations and instead prioritise optimising multiple embedding vectors for optimal image-level reconstruction. To enhance the accuracy of prompt-object level correlation, we propose the following regularisation techniques: 1) To ensure a concentrated correlation between each prompt-concept pair, we propose *Attention Masking (AttnMask)*, restricting prompt learning to relevant regions defined by a cross-attention-guided mask. 2) Recognising that multiple objects within a scene are semantically distinct, we introduce *Prompts Contrastive Loss (PromptCL)* to facilitate the disentanglement of prompt embeddings associated with multiple concepts. 3) To further enable accurate control of each learned embedding, we bind each learnable prompt with a related descriptive adjective word, referred to as *Bind adj.*, that we empirically observe has a strong regional correlation. The middle and bottom row of Figure 3 illustrates the proposed regularisation techniques. Our evaluation shows that our framework improves precision in learning object-level concepts and facilitates explainable hypothesis generation via local editing. This is achieved without requiring explicit image annotations, as exemplified in Figure 1 (ours).

In this work we implement our proposed method based on Textural Inversion by (Gal et al., 2022), which only learns and stores textural embeddings, and is *complementary to the more expensive generation-focused approaches that update DM parameters* (see Table 1). To our knowledge, our technique is the first to learn multiple object-level concepts without using a crop or mask. To evaluate this new task, we generate in-distribution natural images and collect out-of-distribution biomedical images, each featuring 2 to 5 concepts along with object-level masks. This results in a dataset comprising 25 concepts and 1,000 sentence-image pairs. We run around 3500 GPU hours experiments to compare with competitive methods, with each run taking approximately one hour. We assess concept disentanglement using t-SNE and evaluate object embedding similarity against masked ground truth in pre-trained BERT, CLIP, DINOv1, and DINOv2 embedding spaces. Our results show that our method can identify multiple concepts in an image and supports the discovery of new concepts using only linguistic descriptions.

2. Related Works

Language-driven vision concept discovery. In many scientific fields, discovery often begins with visual observation and then progresses by exploring the existing knowledge base to pinpoint unfamiliar object-level concepts. These concepts are subsequently defined using new terms, facilitating the development of hypotheses (Schickore, 2022). The emergence of artificial intelligence, particularly large pre-trained Vision-Language Models (VLM), has laid the groundwork for automating this discovery process (Wang et al., 2023). Language-driven local editing in VLMs shows promise for helping scientists to generate hypotheses and create designs (Hertz et al., 2022; Tumanyan et al., 2023; Patashnik et al., 2023). However, a key challenge remains: relying solely on linguistic descriptions, current methods may not always map words to their corresponding object-level concepts precisely.

Prompt learning for Diffusion Model. In text-guided image synthesis, prompt learning links the appearance and style of an unseen image to a learnable prompt, enabling transfer to new images. This is achieved either by learning and storing textural embeddings, as in Textural Inversion (Gal et al., 2022), or by optimising the entire diffusion model to reconstruct a given example image, as demonstrated in DreamBooth (Ruiz et al., 2022).

Multiple concept learning and composing. Recent advancements focus on efficiently composing multiple con-

cepts learned separately from single object images or crops (Custom Diffusion (Kumari et al., 2023), Cones (Liu et al., 2023), SVDiff (Han et al., 2023), Perfusion (Tewel et al., 2023). ELITE (Wei et al., 2023) and Break-A-Scene (Avrahami et al., 2023) adopt masks for improved object-level concept learning, with Break-A-Scene specifically aiming for multi-concept learning and integrating within single images, aligning closely with our objectives. Our approach differs from Break-A-Scene in two key aspects: 1) we aim to *eliminate the need for labour-intensive image annotations*, and 2) we explore the limits of multi-concept learning *without updating or storing the DM parameters*.

3. Methods

In this section, we outline the preliminaries in Section 3.1 and present a motivational study in Section 3.2. These tests validate the presence of object-level embeddings in the pretrained textural embedding space, highlighting the challenges in learning multiple concepts without image annotations. Inspired by these results, we introduce the *Multi-Concept Prompt Learning (MCPL)* in Section 3.3. To address the multi-object optimisation challenge in tandem with a single image-level reconstruction goal, we propose several regularisation techniques in Section 3.4.

3.1. Preliminaries

Text-guided diffusion models are probabilistic generative models that approximate the data distribution (specifically, images in our work) by progressively denoising Gaussian random noise, conditioned on text embeddings. Specifically, we are interested in a denoising network ϵ_θ being pre-trained such that, given an initial Gaussian random noise map $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, conditioned to text embeddings v, generates an image $\tilde{x} = \epsilon_\theta(\boldsymbol{\epsilon}, v)$ closely resembling a given example image x. Here, $v = c_\phi(p)$, where c_ϕ is a pre-trained text encoder with parameters ϕ and p is the text. During training, ϕ and θ are jointly optimised to denoise a noised image embedding z_t to minimise the loss:

$$L_{DM} = L_{DM}(x, \tilde{x}) := E_{z,v,\epsilon,t}\|\epsilon - \epsilon_\theta(z_t, v)\|^2. \quad (1)$$

Here, $z_t := \alpha_t z + \sigma_t \boldsymbol{\epsilon}$ is the noised version of the initial image embedding z at time $t \sim \text{Uniform}(1, T)$, α_t, σ_t are noise scheduler terms, and $z = \mathcal{E}(x)$, where $\mathcal{E}$ is the encoder of a pretrained autoencoder $D(\mathcal{E}(x)) \approx x$, following Latent Diffusion Models (LDMs) (Rombach et al., 2022) for computational efficiency. During inference, the pre-trained model iteratively eliminates noise from a new random noise map to generate a new image.

Textural Inversion (Gal et al., 2022) is aimed at identifying the text embedding v^* for a new prompt p^* with pre-trained $\{\epsilon_\theta, c_\phi\}$. Given a few (3-5) example images representing a specific subject or concept, the method optimises v^* in the frozen latent space of text encoder c_ϕ. The objective is to generate images via the denoising network ϵ_θ that closely resembles the example images (after decoding) when conditioned on v^*. The optimisation is guided by the diffusion model loss defined in equation 1, updating only v^* while keeping c_ϕ and ϵ_θ frozen. During training, the generation is conditioned on prompts combining randomly selected text templates y (e.g., "A photo of", "A sketch of" from CLIP (Radford et al., 2021)) with the new prompt p^*, resulting in phrases like "A photo of p^*" and "A sketch of p^*".

Cross-attention layers play a pivotal role in directing the text-guided diffusion process. Within the denoising network, ϵ_θ, at each time step t the textual embedding, $v = c_\phi(p)$, interacts with the image embedding, z_t, via the cross-attention layer. Here, $Q = f_Q(z_t)$, $K = f_K(v)$, and $V = f_V(v)$ are acquired using learned linear layers f_Q, f_K, f_V. As (Hertz et al., 2022) highlighted, the per-prompt cross-attention maps, $M = \text{Softmax}(QK^T/\sqrt{d})$, correlate to the similarity between Q and K. Therefore the average of the cross-attention maps over all time steps reflects the crucial regions corresponding to each prompt word, as depicted in Figure 3. In this study, the per-prompt attention map is a key metric for evaluating the prompt-concept correlation. Our results will show that without adequate constraints, the attention maps for newly learned prompts often lack consistent disentanglement and precise prompt-concept correlation.

3.2. Motivational study

To understand the possibility of learning multiple concepts within a frozen textural embedding space, we explored whether *Textural Inversion* can discern semantically distinct concepts from both masked and cropped images, each highlighting a single concept. Figure 2 (two examples on the left) gives a highlight of our result, with a full version in Appendix A.5, confirms that: 1) multiple unique embeddings can be derived from a single multi-concept image, albeit with human intervention, and 2) despite having well-learned individual concepts, synthesising them into a unified multi-concept scene remains challenging. To address these issues, we introduce the Multi-Concept Prompt Learning (MCPL) framework. MCPL modifies Textural Inversion to enable simultaneous learning of multiple prompts within the same string.

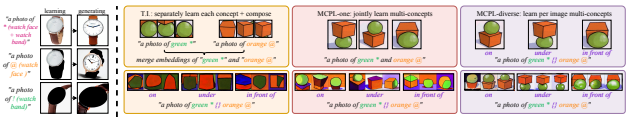

Figure 2. **Motivational study and preliminary MCPL results.** We use *Textural Inversion (T.I.)* to learn concepts from both masked (left-first) or cropped (left-second) images; *MCPL-one*, learning both concepts jointly from the full image with a single string; and *MCPL-diverse* accounting for per-image specific relationships.

3.3. Multi-Concept Prompt Learning (MCPL)

For example image(s) x, MCPL learn a list of embeddings $\mathcal{V} = [v^*, \dots, v^{\&}]$ corresponds to multiple new prompts $\mathcal{P} = [p^*, \dots, p^{\&}]$ within the descriptive sentence as shown in Figure 3. The optimisation is guided by the image-level L_{DM} defined in equation 1, but now updating $\{v^*, \dots, v^{\&}\}$ while keeping c_ϕ and ϵ_θ frozen. The MCPL algorithm is outlined in Algorithm 1.

Algorithm 1 MCPL

Input: example image(s) x, pre-trained $\{c_\theta, \epsilon_\theta\}$.
Output: a list of embeddings $\mathcal{V} = [v^*, \dots, v^{\&}]$ corresponds to multiple new prompts $\mathcal{P} = [p^*, \dots, p^{\&}]$.
`# optimising` $\{v^*, \dots, v^{\&}\}$ `with` L_{DM}
for *step* $= 1, \dots, S$ **do**
  Encode example image(s) $z = \mathcal{E}(x)$ and randomly sample neutral texts y to make string $[y, p^*, \dots, p^{\&}]$
  Compute $\mathcal{V}_\dagger = [v^y, v^*, \dots, v^{\&}] = [c_\theta(p^y), c_\theta(p^*), \dots, c_\theta(p^{\&})]$
  for $t = T, T-1, \dots, 1$ **do**
    $\mathcal{V} := \arg\min_{\mathcal{V}} E_{z,\mathcal{V},\epsilon,t} \|\epsilon - \epsilon_\theta(z_t, \mathcal{V}_\dagger)\|^2$
  end
end
Return $(\mathcal{P}, \mathcal{V})$

Training strategies and preliminary results. Recognising the complexity of learning multiple embeddings with a single image-generation goal, we propose three training strategies: 1) *MCPL-all*, a naive approach that learns embeddings for all prompts in the string (including adjectives, prepositions and nouns. etc.); 2) *MCPL-one*, which simplifies the objective by learning single prompt (nouns) per concept; 3) *MCPL-diverse*, where different strings are learned per image to observe variances among examples. Preliminary evaluations of *MCPL-one* and *MCPL-diverse* methods on the "ball" and "box" multi-concept task are shown in Figure 2. Our findings indicate that *MCPL-one* enhance the joint learning of multiple concepts within the same scene over separate learning. Meanwhile, *MCPL-diverse* goes further by facilitating the learning of intricate relationships between multiple concepts.

Limitations of plain MCPL. We aim to discover new visual concepts using only linguistic descriptions and then enable accurate local editing. It requires accurate object-level prompt-concept correlation, to evaluate, we visualise the average cross-attention maps for each prompt. depicted in Figure 4 (top), plain *MCPL* inadequately capture this correlation, especially for the target concept. These results suggest that *naively extending image-level prompt learning techniques (Gal et al., 2022) to object-level multi-concept learning poses optimisation challenges*, notwithstanding the problem reformulation efforts discussed in Section 3.3. Specifically, optimising multiple object-level prompts based on a single image-level objective proves to be non-trivial.

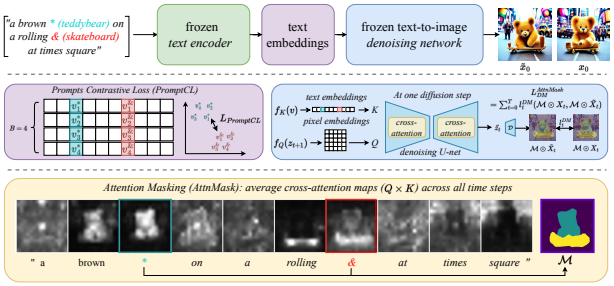

Figure 3. **Method overview.** *MCPL* takes a sentence (top-left) and a sample image x_0 (top-right) as input, feeding them into a pre-trained text-guided diffusion model comprising a text encoder c_ϕ and a denoising network ϵ_θ. The string's multiple prompts are encoded into a sequence of embeddings which guide the network to generate images $\tilde{x}_0$ close to the target one x_0. MCPL focuses on learning multiple learnable prompts (coloured texts), updating only the embeddings v^* and $v^{\&}$ of the learnable prompts while keeping c_ϕ and ϵ_θ frozen. We introduce *Prompts Contrastive Loss (PromptCL)* to help separate multiple concepts within learnable embeddings. We also apply *Attention Masking (AttnMask)*, using masks based on the average cross-attention of prompts, to refine prompt learning on images. Optionally we associate each learnable prompt with an adjective (e.g., "brown" and "rolling") to improve control over each learned concept, referred to as *Bind adj.*

Given the image generation loss equation 1, prompt embeddings may converge to trivial solutions that prioritize image-level reconstruction at the expense of semantic prompt-object correlations, thereby contradicting our objectives. In the next section, we introduce multiple regularisation terms to overcome this challenge.

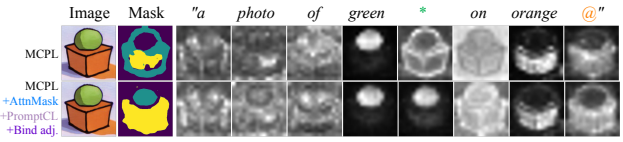

Figure 4. **Enhancing object-level prompt-concept correlation in MCPL** using the proposed regularisations: *AttnMask*, *PromptCL* and *Bind adj.*. We compare *MCPL-one* applying all regularisation terms against the plain *MCPL-one*, using a "Ball and Box" example. We use the average cross-attention maps and the *AttnMask* to assess the accuracy of correlation. Full ablation results in Appendix A.6

3.4. Regularising the multi-concept prompts learning

Encouraging focused prompt-concept correlation with Attention Masking (*AttnMask*). Previous results show plain *MCPL* may learn prompts focused on irrelevant areas. To correct this, we apply masks to both generated and target images over all the denoising steps (Figure 3, middle-right). These masks, derived from the average cross-attention of selected learnable prompts (Figure 3, bottom-row), con-

strain the image generation loss (equation 1) to focus on pertinent areas, thereby improving prompt-concept correlation. To calculate the mask, we compute for each selected learnable prompt $p \in \mathcal{P}$ the average attention map over all time steps $\overline{M}^p = 1/T \sum_{t=1}^{T} M_t^p$. We then apply a threshold to produce binary maps for each learnable prompt, where $B(M^p) := \{1 \text{ if } M^p > k, 0 \text{ otherwise}\}$ and $k = 0.5$ throughout all our experiments. For multiple prompt learning objectives, the final mask $\mathcal{M}$ is a union of multiple binary masks of all learnable prompts $\mathcal{M} = \bigcup_{p \in \mathcal{P}} B(M^p)$. We compute the Hadamard product of $\mathcal{M}$ with x and $\tilde{x}$ to derive our masked loss $L_{DM}^{AttnMask}$ as equation 2. Our *AttnMask* is inspired by (Hertz et al., 2022), but a reverse of the same idea, where the *AttnMask* is applied over the pixel-level loss equation 1 to constrain the prompt learning to only related regions.

$$L_{DM}^{AttnMask} = L_{DM}\left(\mathcal{M} \odot x, \mathcal{M} \odot \tilde{x}\right), \tag{2}$$

Encouraging semantically disentangled multi-concepts with Prompts Contrastive Loss (*PromptCL*). *AttnMask* focuses the learning of multiple prompts on the joint area of target objects, eliminating the influence of irrelevant regions like the background. However, it doesn't inherently promote separation between the embeddings of different target concepts. Leveraging the mutual exclusivity of multiple objects in an image, we introduce a contrastive loss in the latent space where embeddings are optimised. Specifically, we employ an InfoNCE loss (Oord et al., 2018), a standard in contrastive and representation learning, to encourage disentanglement between groups of embeddings corresponding to distinct learnable concepts (Figure 3, middle-left).

Concretely, at each learning step as described in Algorithm 1, a mini-batch B minor augmented (e.g. with random flip) example images are sampled, with N learnable prompts for each image, yields a set of BN embeddings, $\{v_b^n\}_{b=1,n=1}^{B,N}$. Then, the similarity between every pair v_i and v_j of the BN samples is computed using cosine similarity:

$$sim(v_i, v_j) = v_i^T . v_j / ||v_i|| ||v_j||. \tag{3}$$

Given our goal is to differentiate the embeddings corresponding to each prompt, we consider the embeddings of the same concept as positive samples while the others as negative. Next, the contrastive loss $l_{i,j \in B}^{\eta}$ for a positive pair v_i^{η} and v_j^{η} of each concept $\eta \in N$ (two augmented views of the example image) is shown in the equation 4, where τ is a temperature parameter following (Chen et al., 2020). The contrastive loss is computed for BN views of each of the N learnable concepts. The total contrastive loss $L_{PromptCL}$ is shown in equation 5.

$$l_{i,j \in B}^{\eta} = -\log\left(\frac{\exp(sim(v_i^{\eta}, v_j^{\eta}))/\tau}{\sum_{\eta=1}^{N}\sum_{j=1,j\neq i}^{B}\exp(sim(v_i^{\eta}, v_j^{\eta})/\tau)}\right) \tag{4}$$

$$L_{PromptCL} = \frac{1}{N}\frac{1}{B}\sum_{\eta=1}^{N}\sum_{i=1}^{B} l_{i,j \in B}^{\eta} \tag{5}$$

Enhance prompt-concept correlation by binding learnable prompt with the adjective word (*Bind adj.*). An additional observation from the misaligned results in Figure 4 (top) reveals that adjective words often correlate strongly with specific regions. This suggests that the pre-trained model is already adept at recognising descriptive concepts like colour or the term "fluffy" (see full results in Figure 23). To leverage this innate understanding, we propose to optionally associate one adjective word for each learnable prompt as one positive group during the contrastive loss calculation. In particular, consider M adjective words associated with N learnable prompts. Then the positive pair v_i^{η} and v_j^{η} of each concept is sampled from $\eta \in MN$ instead of N. Therefore the contrastive loss is now computed for BNM views of each of the N learnable concepts. The resulting total contrastive loss $L_{PromptCL}^{adj}$ is detailed in equation 6. We scale $L_{PromptCL}^{adj}$ with a scaling term γ and add with $L_{DM}^{AttnMask}$ (equation 2), for them to have comparable magnitudes, resulting our final loss in equation 7.

$$L_{PromptCL}^{adj} = \frac{1}{NM}\frac{1}{B}\sum_{\eta=1}^{NM}\sum_{i=1}^{B} l_{i,j \in B}^{\eta} \tag{6}$$

$$L = L_{DM}^{AttnMask} + \gamma L_{PromptCL}^{adj}, \tag{7}$$

Assessing regularisation terms with cross-attention. We assess our proposed regularisation terms on improving the accuracy of semantic correlations between prompts and concepts. We visualise the cross-attention and segmentation masks, as shown in Figure 4. Our visual results suggest that incorporating all of the proposed regularisation terms enhances concept disentanglement, whereas applying them in isolation yields suboptimal outcomes (refer to full ablation results in Appendix A.6). Moreover, the results demonstrate that *MCPL-one* is a more effective learning strategy than *MCPL-all*, highlighting the importance of excluding irrelevant prompts to maintain a focused learning objective.

4. Experiments

4.1. Experiment and Implementation Details

Multi-concept dataset. We generate in-distribution natural images and collect out-of-distribution biomedical images, each featuring 2 to 5 concepts along with object-level masks. This results in a dataset comprising 25 concepts and 1,000 sentence-image pairs. For natural images, we generate multi-concept images using prior local editing (Patashnik et al., 2023) and multi-concept composing method (Avrahami et al., 2023), both support generating or predicting ob-

ject masks. We generate each image using simple prompts, comprising one adjective and one noun for every relevant concept. For biomedical images, we request a human or a machine, such as GPT-4, to similarly describe each image using one adjective and one noun for each pertinent concept. For more details, a full list of prompts used and examples, please read Appendix A.3.

Competing Methods. We compare three baseline methods: 1) *Textural Inversion (TI-m)* applied to each masked object serving as our best estimate for the unknown disentangled "ground truth" object-embedding. 2) *Break-A-Scene (BAS)*, the state-of-the-art (SoTA) mask-based multi-concept learning method, serves as a performance upper bound, though it's not directly comparable. 3) *MCPL-all* as our naive adaptation of the *Textural Inversion* method to achieve the multi-concepts learning goal. For our method, we compare two training strategies of our method: MCPL-all and MCPL-one. For each, we examine three variations to scrutinise the impact of the regularisation terms discussed in Section 3.4. **All MCPL learnings are performed on unmasked images without updating DM parameters.** To evaluate the robustness and concept disentanglement capability of each learning method, we repeatedly learn each multi-concept pair around 10 times, randomly sampling four images each time. We generated a total of 560 masked objects for each MCPL variant and 320 for the BAS baseline.

However, it's important to note that preparing *the BAS as the object-level embedding upper bound is costly*, as it requires an additional pre-trained segmentation model, and occasionally human-in-the-loop, to obtain masks during both the learning and evaluation phases (see Appendix A.4 for details). In contrast, our method utilises its own *AttnMask* to *generate masked concepts during image generation with no extra cost*.

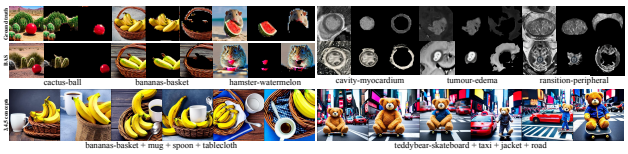

Figure 5. Visualisation of the prepared ground truth examples (first-row), the generated images with Break-A-Scene (second-row) and examples of images containing more than two concepts (more examples available in Appendix A.3).

Implementation details. We use the same prompts collected during the data preparation, substituting nouns as learnable prompts, which are merged with CLIP prompts from Section 3.1. This process creates phrases such as *"A photo of brown * on a rolling @ at times square"*. Unless otherwise noted, we retain the original hyper-parameter choices of LDM (Rombach et al., 2022). Our experiments were conducted using a single V100 GPU with a batch size of 4. The base learning rate was set to 0.005. Following LDM, we further scale the base learning rate by the number of GPUs and the batch size, for an effective rate of 0.02. On calculating $L_{PromptCL}$, we apply the temperature and scaling term (τ, γ) of $(0.2, 0.0005)$ when *AttnMask* is not applied, and $(0.3, 0.00075)$ when *AttnMask* is applied. All results were produced using 6100 optimisation steps. We find that these parameters work well for most cases. The experiments were executed on a single V100 GPU, with each run taking approximately one hour, resulting in a total computational cost of around 3500 GPU hours (or 150 days on a single GPU). We employed various metrics to evaluate the method.

4.2. Quantitative Evaluations

Investigate the concepts disentanglement with t-SNE. To assess disentanglement, we begin by calculating and visualising the t-SNE projection of the learned features (Van der Maaten & Hinton, 2008). The results, depicted in Figure 6, encompass both natural and biomedical datasets. They illustrate that our *MCPL-one* combined with all regularisation terms can effectively distinguish all learned concepts compared to all baselines. It's noteworthy that the learned embeddings from both the mask-based "ground truth" and BAS show less disentanglement compared to ours, attributable to their lack of a specific disentanglement objective, such as the *PromptCL* loss in MCPL. This finding confirms the necessity of our proposed method.

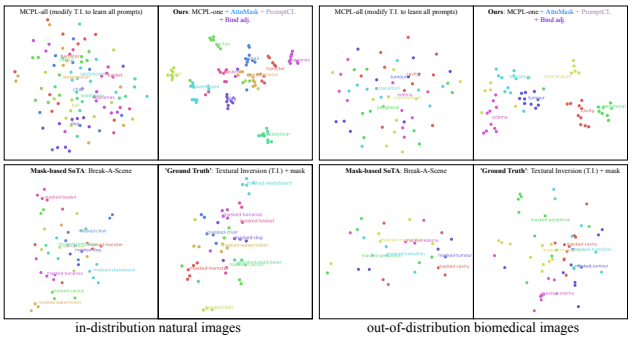

Figure 6. **The t-SNE projection of the learned embeddings**. Our method can effectively distinguish all learned concepts (about 10 embeddings each concept) compared to Textural Inversion (MCPL-all), the SoTA mask-based learning method, Break-A-Scene, and the masked "ground truth" (see full results in Appendix A.1).

Embedding similarity relative to the "ground truth". To assess the preservation of per-concept semantic and textural details, we calculate both prompt and image fidelity. This evaluation follows prior research by (Gal et al., 2022) and (Ruiz et al., 2022), but differently, we perform the calculations at the object level. We compared the masked "ground truth" and generated masked-objects across four embedding

spaces. Prompt fidelity is determined by measuring the average pairwise cosine similarity between the embeddings learned from the estimated "ground truth" and the generated masked images, in the pre-trained embedding space of BERT (Devlin et al., 2018). For image fidelity, we compare the average pairwise cosine similarity in the pre-trained embedding spaces of CLIP (Radford et al., 2021), DINOv1 (Caron et al., 2021) and DINOv2 (Oquab et al., 2023), all based on the ViT-S backbone.

The results in Figure 7 show our method combined with all the proposed regularisation terms can improve both prompt and image fidelity consistently. Our fully regularised version (*MCPL-one+CL+Mask*) achieved competitive performance compared to the SoTA mask-based method (BAS) on the natural dataset. In the OOD medical dataset, BAS outperformed our method significantly in the DINOv1 embedding space, although the performance was comparable in other spaces. The discrepancy stems from our method's attention masks yielding less accurate object masks compared to *BAS whose masks are obtained through a specialised human-in-the-loop segmentation protocol*, as detailed in Appendix A.4. This difference is depicted in Figures 5 and 11.

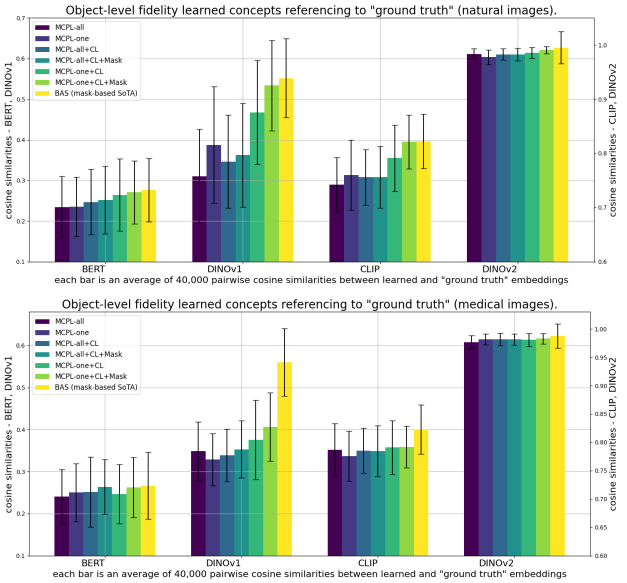

Figure 7. **Embedding similarity in learned object-level concepts compared to masked "ground truth" (two concepts per image).** We compare Textural Inversion (MCPL-all) and the SoTA mask-based learning method, BAS, against our regularised versions. The analysis is conducted in both pre-trained text (BERT) and image encoder spaces (CLIP, DINOv1, and DINOv2), with each bar representing an *average of 40k* pairwise cosine similarities.

Learning more than two concepts. To validate our method's robustness, we expanded the evaluation to learning tasks with more than two concepts per image, specifically natural images containing 3, 4, and 5 concepts. Table 2 presents the average similarities of object-embedding learnt from images with 2 to 5 concepts. These results align with previous findings, showing that our fully regularised version surpasses all baselines and closely approaches the mask-based performance upper limit set by BAS.

Method names	MCPL -all	MCPL -one	MCPL-all +Reg.	MCPL-one +Reg.	BAS
DINOv1	0.334	0.405	0.377	0.503	0.529
CLIP	0.289	0.324	0.312	0.331	0.354

Table 2. **Average object-level embedding similarity of all natural images including 2 to 5 concepts per image.** Each number is an average of *360k pairwise* cosine similarities. *Reg.* represents including all the proposed regularisation terms.

We group the learned embeddings by the number of concepts per image to evaluate their impact on learning efficiency. The results in Figure 8 reveal that: 1) learning efficiency diminishes with the increase in the number of concepts per image, a trend also evident in the mask-based BAS approach; 2) although our fully regularised version continues to outperform under-regularised versions, the performance gap to BAS widens, highlighting the heightened challenge of mask-free multi-concept learning in more complex scenes.

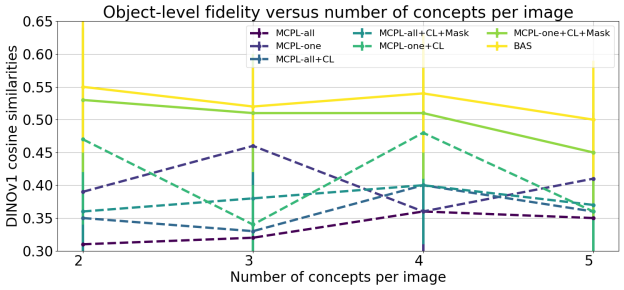

Figure 8. **Evaluate the learning as the number of concepts per image increases.** Here each data point represents an *average of 20∼40k* pairwise cosine similarities measured by DINOv1.

Object-level evaluations. As our main goal is to accurately learn object-level embeddings, we perform an analysis of embedding similarity at the object level as highlighted in Figure 9. Notably, our method sometimes surpasses the mask-based method, BAS, at the object level. *This underscores the potential of our mask-free, language-driven approach to learning multiple concepts from a single image.*

4.3. Qualitative Evaluation

Visualise concepts disentanglement and learning. To evaluate disentanglement and prompt-to-concept correlation, we visualise attention and attention masks for learnable prompts. Figures 11 and 12 display results for both natural and medical images. The visual outcomes align with

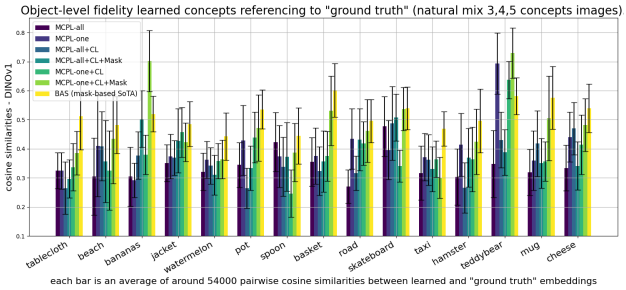

Figure 9. **Object-level embedding similarities with DINOv1 (mix of 3 to 5 concepts per image).** Each bar representing an *average of 160k* pairwise cosine similarities. A comprehensive object-level comparison is available in the Appendix (Section A.2).

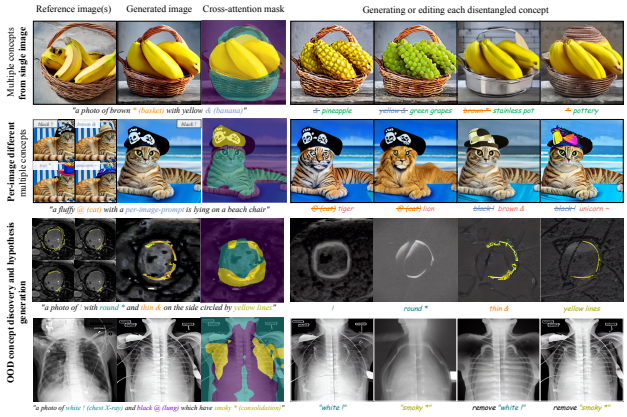

Figure 10. **MCPL learning and editing capabilities.** Top-row: learning and editing multiple concepts with a single string. Second-row: discovering per-image different concepts with per-image specified string. Third-row: learning to disentangle multiple unseen concepts from cardiac MRI images. Bottom-row: learning to disentangle multiple unseen concepts from chest X-ray images.

earlier quantitative findings, affirming the effectiveness of our proposed MCPL method and regularisation terms.

Image editing over disentangled concepts. Hypothesis generation in scientific fields can be accelerated by integrating new concepts into existing observations, a process that benefits from local image editing. We demonstrate our method enables *mask-free object-level learning, editing and quantification* (Figure 10 top-row). The framework also has the flexibility to *handle per-image specified string to learn the different concepts within each image*, as shown in the second-row example of Figure 10. Furthermore, our method can also learn unknown concepts from challenging out-of-distribution images (Figure 10 third and bottom rows), *opening an avenue of knowledge mining from pairs of textbook figures and captions*. It is worth noting that, compared to BAS, **our method does not rely on a separate segmentation model and mask to achieve local editing**. Our method optimises the disentanglement of multiple concepts, leading to accurate word-concept correlation in the cross-attention hence supporting mask-free local editing method (e.g. P2P (Hertz et al., 2022)) directly.

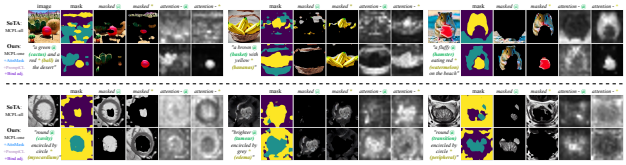

Figure 11. **Visualisation of generated concepts with the "SoTA" and our method (two concepts).** Masks are derived from cross-attentions. Full ablation results are presented in the Appendix A.6

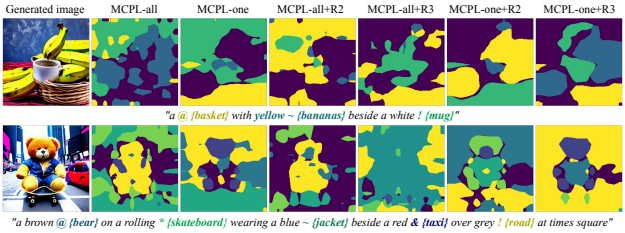

Figure 12. **Visualisation of generated concepts with the "SoTA" and our method (3 or 5 concepts).** Masks are derived from cross-attentions. Full ablation results are presented in the Appendix A.6

4.4. Ablation studies.

We also conduct a set of ablation studies to assess various components and capabilities of our method, with details in the Appendix. They are: 1) The *MCPL-diverse* training strategy has demonstrated potential in learning tasks with varying concepts per image. Therefore, we performed further experiments to assess its effectiveness, with findings detailed in Section A.7 confirming its efficacy. 2) Our language-driven approach benefits from the proposed adjective binding mechanism. To better understand its role, we conducted an ablation study detailed in Section A.8, which confirmed its significance. 3) For a comprehensive evaluation, we visually compare our tuning-free method, which is not specifically designed for composing complex scenes when prompt interactions change, with SoTA composing-focused methods in complex scene tasks. This comparison is detailed in Section A.9 and we achieved promising results.

5. Conclusions

In conclusion, we introduced MCPL to tackle the novel challenge of mask-free learning of multiple concepts using images and natural language descriptions. This approach is expected to assist the discovery of new concepts through natural language-driven human-machine interaction, potentially revolutionising task hypothesis generation and local image editing without requiring explicit knowledge of the new vision concept.

Broader impact

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., et al. The medical segmentation decathlon. *Nature communications*, 13(1): 4128, 2022.

Avrahami, O., Aberman, K., Fried, O., Cohen-Or, D., and Lischinski, D. Break-a-scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2305.16311*, 2023.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Cheng, B., Schwing, A., and Kirillov, A. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34: 17864–17875, 2021.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. URL https://arxiv.org/abs/2208.01618.

Han, L., Li, Y., Zhang, H., Milanfar, P., Metaxas, D., and Yang, F. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. 2022.

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.

Kumari, N., Zhang, B., Zhang, R., Shechtman, E., and Zhu, J.-Y. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931–1941, 2023.

Lalande, A., Chen, Z., Decourselle, T., Qayyum, A., Pommier, T., Lorgis, L., de La Rosa, E., Cochet, A., Cottin, Y., Ginhac, D., et al. Emidec: a database usable for the automatic evaluation of myocardial infarction from delayed-enhancement cardiac mri. *Data*, 5(4):89, 2020.

Liu, Z., Feng, R., Zhu, K., Zhang, Y., Zheng, K., Liu, Y., Zhao, D., Zhou, J., and Cao, Y. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*, 2023.

Ma, J. and Wang, B. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Patashnik, O., Garibi, D., Azuri, I., Averbuch-Elor, H., and Cohen-Or, D. Localizing object-level shape variations with text-to-image diffusion models, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022.

Schickore, J. Scientific Discovery. In Zalta, E. N. and Nodelman, U. (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition, 2022.

Tewel, Y., Gal, R., Chechik, G., and Atzmon, Y. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–11, 2023.

Tumanyan, N., Geyer, M., Bagon, S., and Dekel, T. 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*, pp. 1921–1930, 2023.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., and Zuo, W. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023.

Wu, Z., Lischinski, D., and Shechtman, E. Stylespace analysis: Disentangled controls for stylegan image generation, 2020.

A. Appendix

A.1. Full t-SNE results

To assess disentanglement, we calculate and visualise the t-SNE projections. We approximate the "ground truth" using features learned through *Textural Inversion* on per-concept masked images. For benchmarking, we compare our method with the SoTA mask-based multi-concept learning method, *Break-A-Scene (BAS)*, which acts as a performance benchmark but isn't directly comparable. Additionally, we assess variants integrating our proposed regularization techniques to align with the learning objectives.

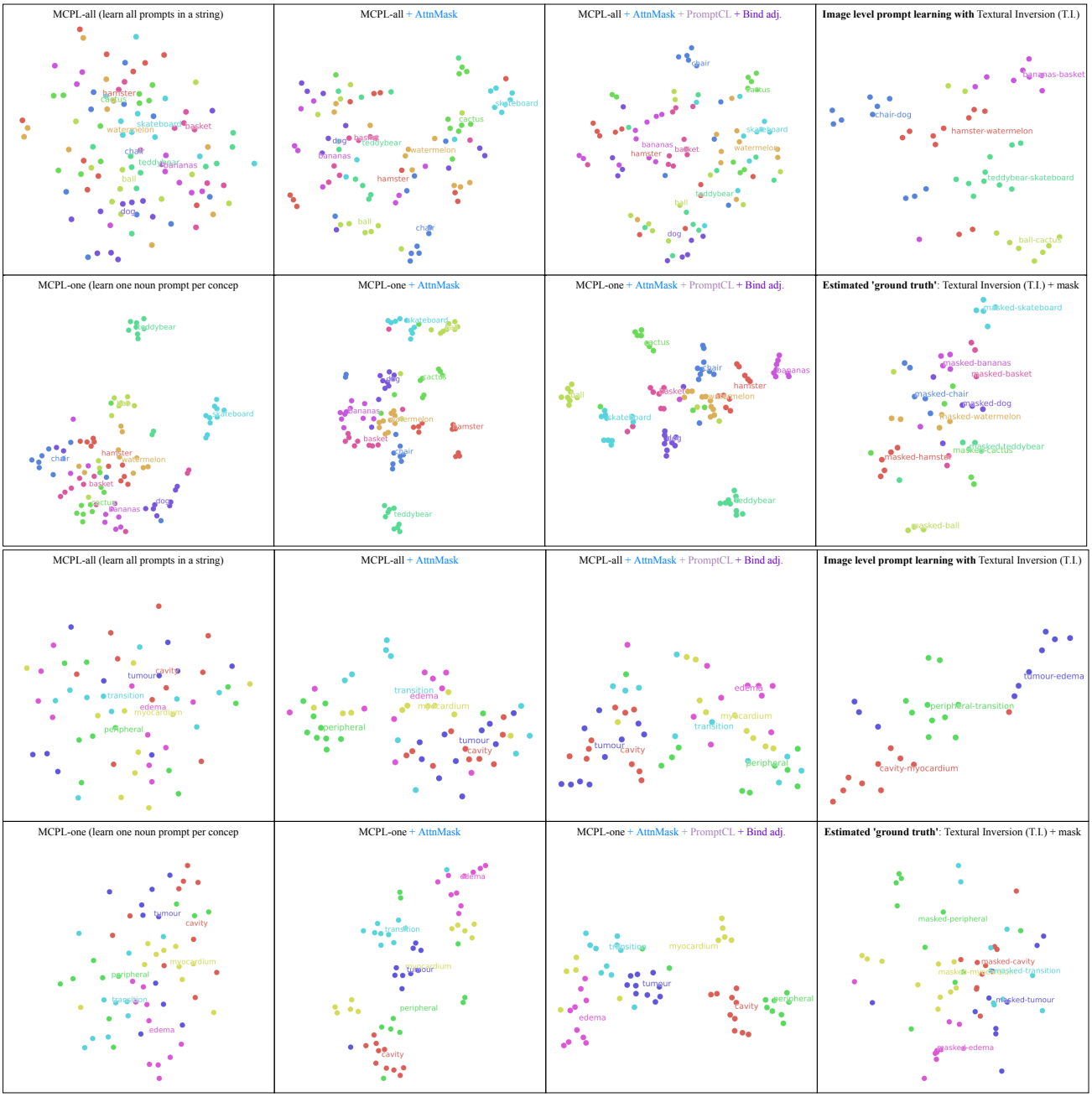

Figure 13. The t-SNE visualisations of learned prompt-concept features (comparing all variants) on the in-distribution natural (top) and out-distribution medical (bottom) dataset. The results confirmed **our *MCPL-one* combined with all regularisation terms can effectively distinguish all learned concepts** compared to all baselines.

A.2. All object-level embedding similarity of the learned concept relative to the estimated "ground truth".

To assess how well our method preserves object-level semantic and textural details, we evaluate both prompt and image fidelity. Our experiments, depicted in Figures 14 and 15, involve learning from dual-concept natural and medical images, and from 3 to 5 natural images. Results **consistently demonstrate that our method adding all proposed regularisation terms accurately learns object-level embeddings**, in comparison with masked ground truth in pre-trained embedding spaces such as BERT, CLIP, DINOv1, and DINOv2. Notably, our method sometimes surpasses the state-of-the-art mask-based technique, Break-A-Scene, at the object level. This underscores the effectiveness and potential of our mask-free, language-driven approach to learning multiple concepts from a single image.

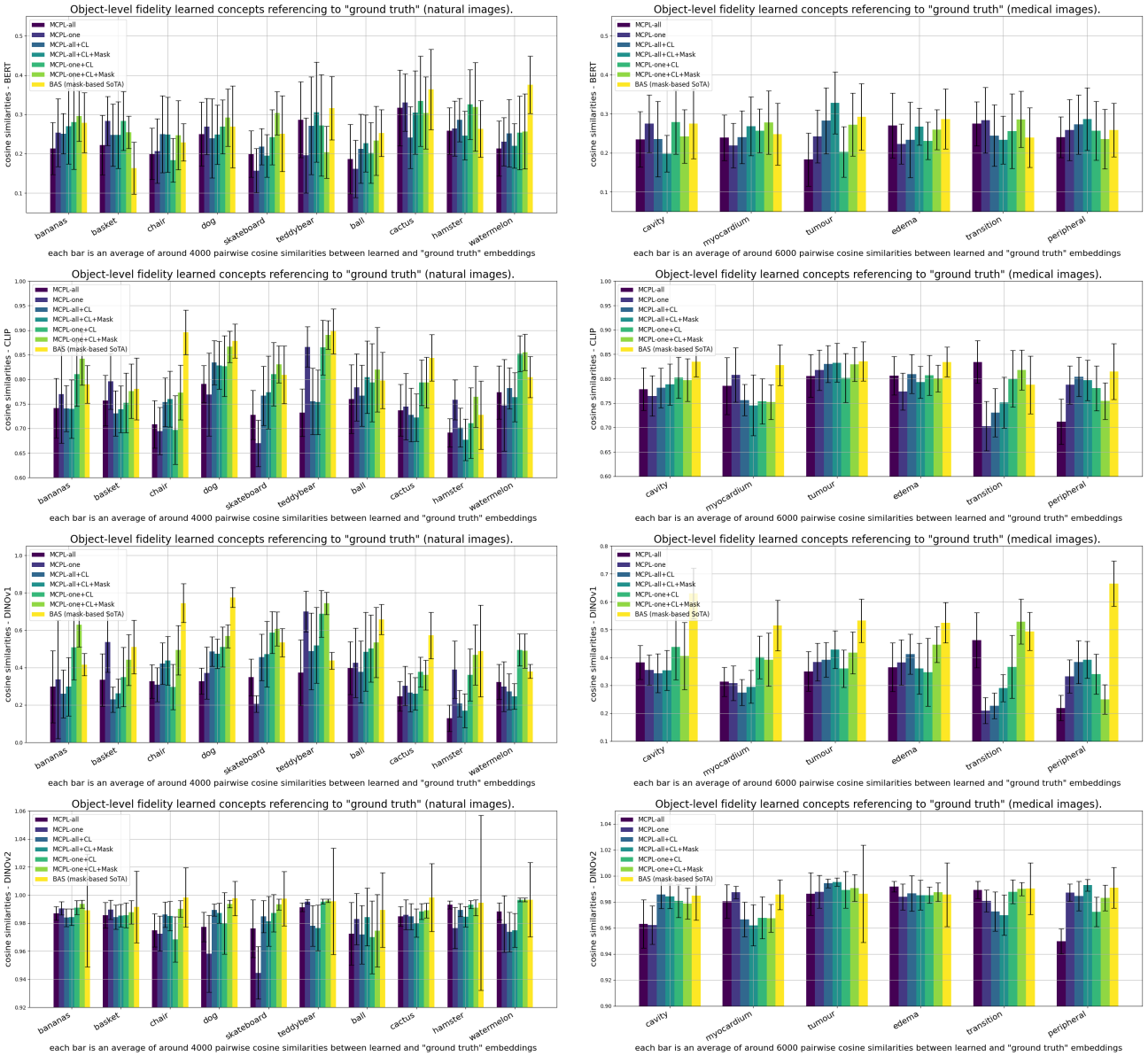

Figure 14. Two-concepts natural (left column) and medical (right column) per-object embedding similarity between the learned concept relative to the masked "ground truth". Each plot computes either the textural or image embeddings in one of the four embedding spaces (BERT, CLIP, DINOv1 and DINOv2). We compare our base version adding variations of our proposed regularisation terms. We also compare against the state-of-the-art (SoTA) mask-based learning method, Break-A-Scene (BAS) (Avrahami et al., 2023).

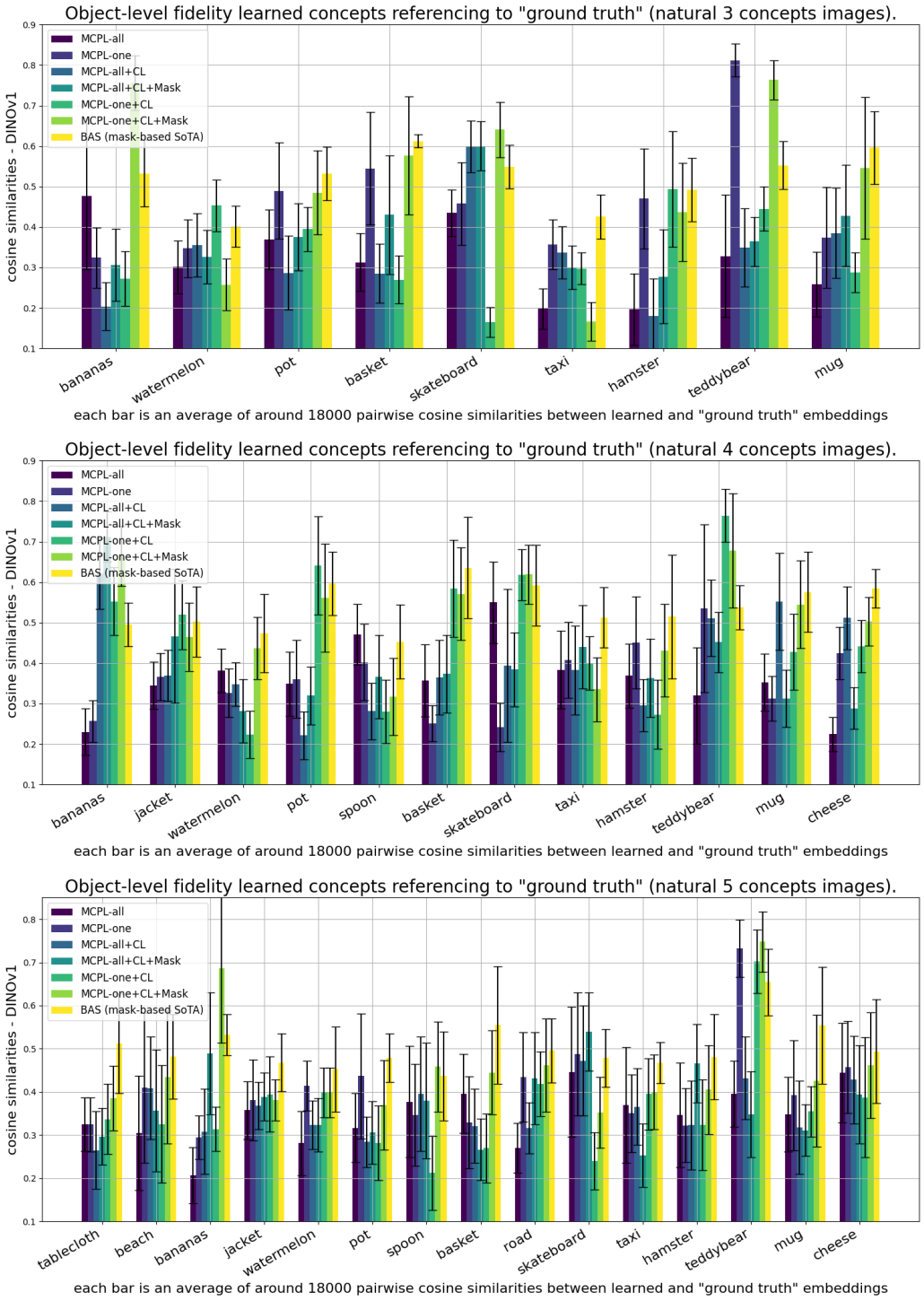

Figure 15. Natural images (each containing 3,4 or 5 objects) per-object embedding similarity between the learned concept relative to the masked "ground truth". Each plot computes image embeddings in the embedding spaces of DINOv1. We compare our base version adding variations of our proposed regularisation terms. **Notably, our method sometimes surpasses the SoTA mask-based technique, Break-A-Scene (Avrahami et al., 2023), at the object level such as bananas and teddybear.**

A.3. Dataset preparation.

For the in-distribution natural images dataset, we first generate variations of two-concept images using local text-driven editing, as proposed by (Patashnik et al., 2023). This minimizes the influence of irrelevant elements like background. This approach also produces per-text local masks based on attention maps, assisting us in getting our best approximation for the "ground truth" of disentangled embeddings. We generate five sets of natural images containing 10 object-level concepts. We generate each image using simple prompts, comprising one adjective and one noun for every relevant concept. For three to five concept images, we use break-a-scene (Avrahami et al., 2023) to generate the more complex composed images. We generate nine sets containing 9 more object-level concepts. We then use separate pre-trained segmentation models—MaskFormer (Cheng et al., 2021) to create masked objects, refer to Appendix A.4 for details of this process.

For the out-of-distribution bio-medical image dataset, we assemble three sets of radiological images featuring 6 organ/lesion concepts. These images are sourced from three public MRI segmentation datasets: heart myocardial infarction (Lalande et al., 2020), prostate segmentation (Antonelli et al., 2022), and Brain Tumor Segmentation (BraTS) (Menze et al., 2014). Each dataset includes per-concept masks. For biomedical images, we request a human or a machine, such as GPT-4, to similarly describe each image using one adjective and one noun for each pertinent concept. For both natural and biomedical datasets, we collected 40 images for each concept. Figure 16 and Figure 17 gives some examples of the prepared datasets.

Two-concepts (natural images)

- "a brown {bear/ tokens1} on a rolling {skateboard/ tokens2} at times square"
- "a fluffy {hamster/ tokens1} eating red {watermelon/ tokens2} on the beach"
- "a green {cactus/ tokens1} and a red {ball/ tokens2} in the desert"
- "a brown {basket/ tokens1} with yellow {bananas/ tokens2}"
- "a white {chair tokens1} with a black {dog/ tokens2} on it"

Two-concepts (medical images)

- "a {scan/ tokens1} with brighter {cavity/ tokens2} encircled by grey {myocardium/ tokens3}"
- "a {scan/ tokens1} with round {transition/ tokens2} encircled by circle {peripheral/ tokens3}"
- "a {scan/ tokens1} with round {tumour/ tokens2} encircled by circle {edema/ tokens3}"

Three-concepts (natural images)

- "a brown {bear/ tokens1} on a rolling {skateboard/ tokens2} beside a red {taxi/ tokens3} at times square"
- "a green {pot/ tokens1} and a {hamster/ tokens1} eating red {watermelon/ tokens2} on the beach"
- "a {basket/ tokens1} with yellow {bananas/ tokens2} beside a white {mug/ tokens3}"

Four-concepts (natural images)

- "a brown {bear/ tokens1} on a rolling {skateboard/ tokens2} wearing a blue {jacket/ tokens3} beside a red {taxi/ tokens4} at times square"
- "a green {pot/ tokens1} and a {hamster/ tokens1} eating red {watermelon/ tokens2} beside a yellow {cheese/ tokens4} on the beach"
- "a {basket/ tokens1} with yellow {bananas/ tokens2} beside a white {mug/ tokens3} and silver {spoon/ tokens4}"

Five-concepts (natural images)

- "a brown {bear/ tokens1} on a rolling {skateboard/ tokens2} wearing a blue {jacket/ tokens3} beside a red {taxi/ tokens4} over grey {road/ tokens5} at times square"
- "a green {pot/ tokens1} and a {hamster/ tokens1} eating red {watermelon/ tokens2} beside a yellow {cheese/ tokens4} on the sandy {beach/ tokens5}"
- "a {basket/ tokens1} with yellow {bananas/ tokens2} beside a white {mug/ tokens3} and silver {spoon/ tokens4} over blue {tablecloth/ tokens5}"

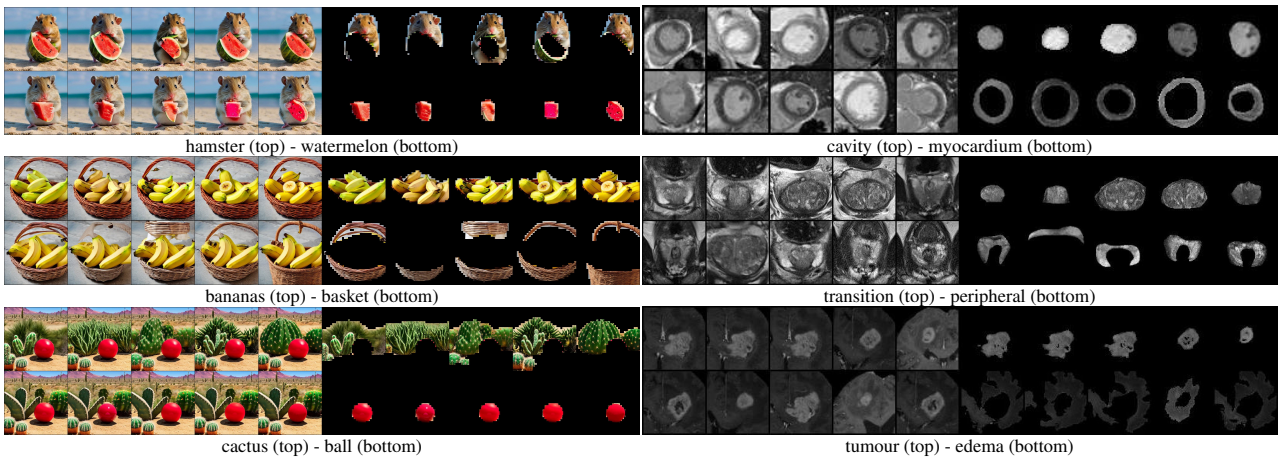

Figure 16. **Evaluation dataset (two concepts).** We prepared five sets of in-distribution natural images and three sets of out-of-distribution biomedical images, each containing two concepts resulting in a total of 16 concepts.

Figure 17. **Evaluation dataset (three to five concepts).** We generate nine sets containing 9 more object-level concepts.

A.4. Break-A-Scene experiments setup

Break-A-Scene (BAS) (Avrahami et al., 2023) learns multiple concepts from images paired with object-level masks. It augments input images with masks to highlight target concepts and updates both textural embeddings and model weights accordingly. BAS introduces 'union sampling', a training strategy that randomly selects subsets of multi-concepts in each iteration to enhance the combination of multiple concepts in generated images, see Figure 28 for an illustration. During inference, BAS employs a pre-trained segmentation model to obtain masked objects, facilitating localised editing.

To fit BAS (Avrahami et al., 2023) into our evaluation protocol, we first learned object-level concepts and then generated masked objects for evaluation, including the following steps:

1. BAS Learning: For each concept pair, we randomly selected 20 images with ground truth segmentations from our dataset for BAS learning, resulting in 20 BAS embeddings per concept.
2. BAS Generation: We then generated 20 images for each concept pair, producing a total of 100 BAS-generated natural images and 60 medical images.
3. Segmentation: For masked object production with BAS, we used different pre-trained segmentation models. MaskFormer (Cheng et al., 2021) was effective for natural images, but segmenting medical images posed challenges due to their out-of-distribution characteristics.

4. Quantitative Evaluation: With the obtained masked objects (20 per concept), we applied the embedding similarity evaluation protocol from Section 4.2 to assess the preservation of semantic and textural details per concept in four embedding spaces.

For segmenting medical images, given the diversity of classes in our dataset, we utilised MedSAM (Ma & Wang, 2023), a state-of-the-art foundation model adapted from SAM (Kirillov et al., 2023) for the medical domain. MedSAM requires a bounding box for input, making it a multi-step, human-in-the-loop process. We initially assessed segmentation quality from several (up to five) bounding box proposals, as exemplified in Figure 19. MedSAM, despite having a bounding box, cannot fully automate segmentation for all classes.

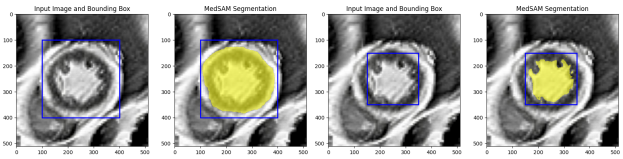

Figure 19. This demonstration shows how MedSAM is used to segment medical images generated by BAS. On the left, MedSAM segmentation with a large bounding box prompt can identify the combined area of the cavity-myocardium classes, but it does not distinguish between the two. On the right, using a smaller bounding box prompt, MedSAM successfully segments the central cavity class. We calculate the difference to get the segmentation of the missing myocardium class (outer ring-like pattern).

Figure 18. Visualisation of the Break-A-Scene results of generated and masked natural images.

Thus, we employed an additional post-processing step to discern the segmentation of both classes by calculating the difference between the two segmentations.

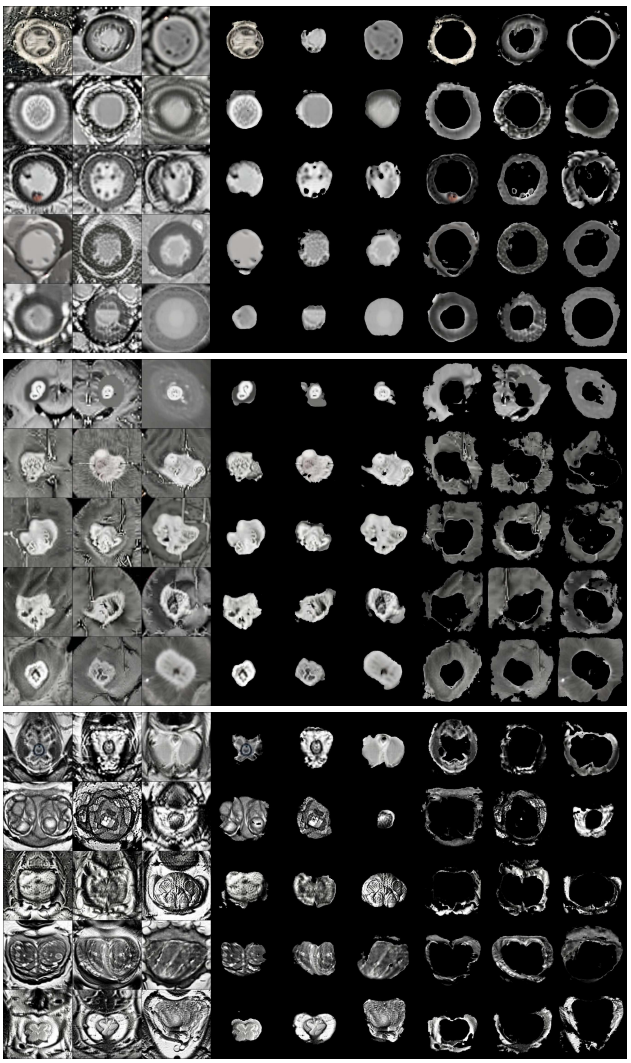

Figure 20. Visualisation of the Break-A-Scene results of generated and masked medical images.

A.5. Full motivational experiment results

Do multiple distinct embeddings arise from the same image? To understand the possibility of learning multiple concepts within a frozen textural embedding space, we explored whether *Textural Inversion* can discern semantically distinct concepts from processed images, each highlighting a single concept. Following (Wu et al., 2020), we used images with manual masks to isolate concepts, as seen in Figure 21. We applied *Textural Inversion* to these images to learn embeddings for the unmasked or masked images. *Our findings indicate that when focusing on isolated concepts, Textural Inversion can successfully learn distinct embeddings, as validated by the generated representations of each concept.*

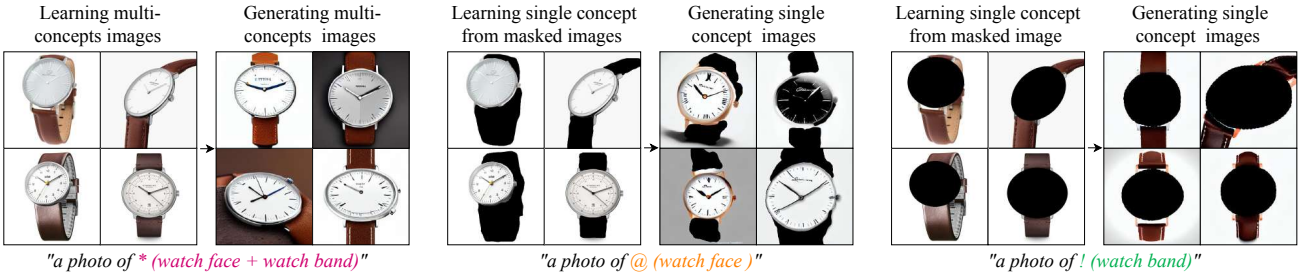

Figure 21. **Motivational study with watch images.** We learn embeddings using Textural Inversion on both unmasked multi-concept images ("watch face" and "watch band") and masked single-concept images ("watch face" or "watch band").

Is separate learning of concepts sufficient for multi-object image generation? While separate learning with carefully sampled or masked images in a multi-object scene deviates from our objective, it is valuable to evaluate its effectiveness. Specifically, we use Textural Inversion to separately learn concepts like "ball" and "box" from carefully cropped images, as shown in Figure 22. We then attempt to compose images using strings that combine these concepts, such as "a photo of a green ball on orange box." *Our results indicate that the accurate composition of multi-object images remains challenging, even when individual concepts are well-learned.*

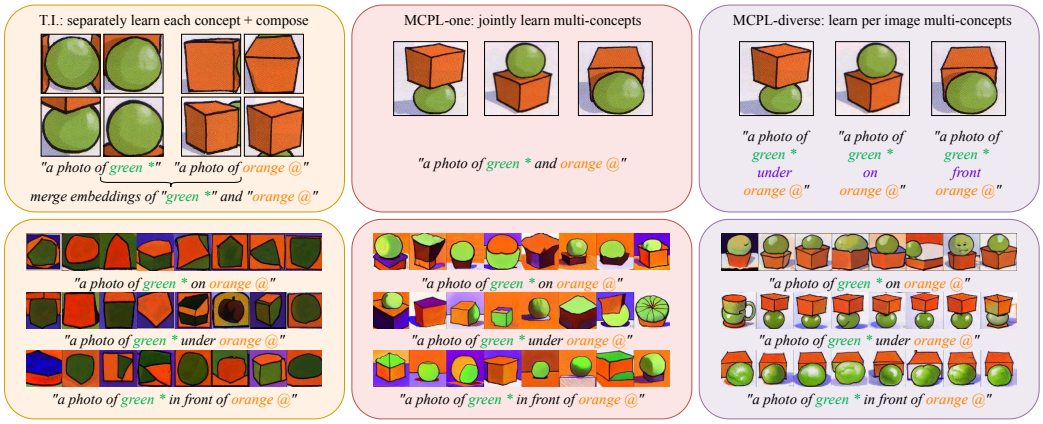

Figure 22. **Learning and Composing "ball" and "box"**. We learned the concepts of "ball" and "box" using different methods (top row) and composed them into unified scenes (bottom row). We compare three learning methods: *Textural Inversion* (Gal et al., 2022), which learns each concept separately from isolated images (left); *MCPL-one*, which jointly learns both concepts from uncropped examples using a single prompt string (middle); and *MCPL-diverse*, which advances this by learning both concepts with per-image specific relationships (right).

A.6. Full ablation results of assessing regularisation terms with cross-attention

We present in this section the full results of assessing our proposed regularisation terms in Section 3.4. The results presented in Figure 23 indicate that plain *MCPL* may not accurately capture semantic correlations between prompts and objects. While adding incorporating the proposed regularisation terms enhances concept disentanglement. We assess the efficacy of these terms in disentangling learned concepts by visualising attention and segmentation masks, as shown in Figure 23. Figure 24 and Figure 25 present the same visualisation of the scaled quantitative experiments involving 2 to 5 concepts.

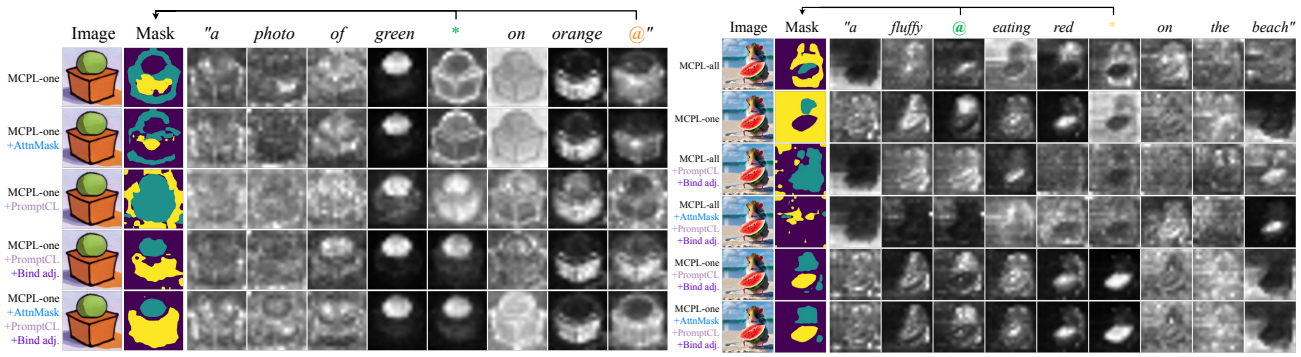

Figure 23. Enhancing object-level prompt-concept correlation in MCPL using proposed *AttnMask*, *PromptCL* and *Bind adj.* regularisation techniques. We use average cross-attention maps to quantify the correlation of each prompt with its corresponding object-level concept. Additionally, we construct attention-based masks from multiple selected prompts for the concepts of interest. The visual results confirm that **incorporating all of the proposed regularisation terms enhances concept disentanglement, whereas applying them in isolation yields suboptimal outcomes.**

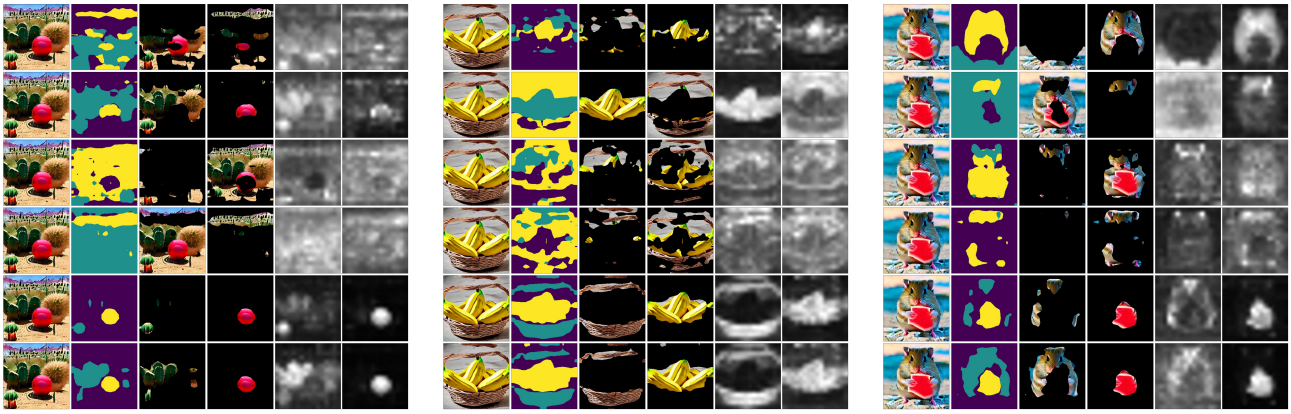

Figure 24. Visualisation of concepts (two concepts). We compare all baseline methods across each row (from top to bottom): 1) MCPL-all, 2) MCPL-one, 3) MCPL-all+*PromptCL+Bind adj.*, 4) MCPL-all+*AttnMask+PromptCL+Bind adj.*, 5) MCPL-one+*PromptCL+Bind adj.*, 6) MCPL-one+*AttnMask+PromptCL+Bind adj.*. **The results on both the natural and medical images confirmed our conclusion** — the inclusion of all proposed regularisation terms (toward the bottom row) consistently demonstrated their effectiveness in enhancing the accuracy of prompt-concept correlation.

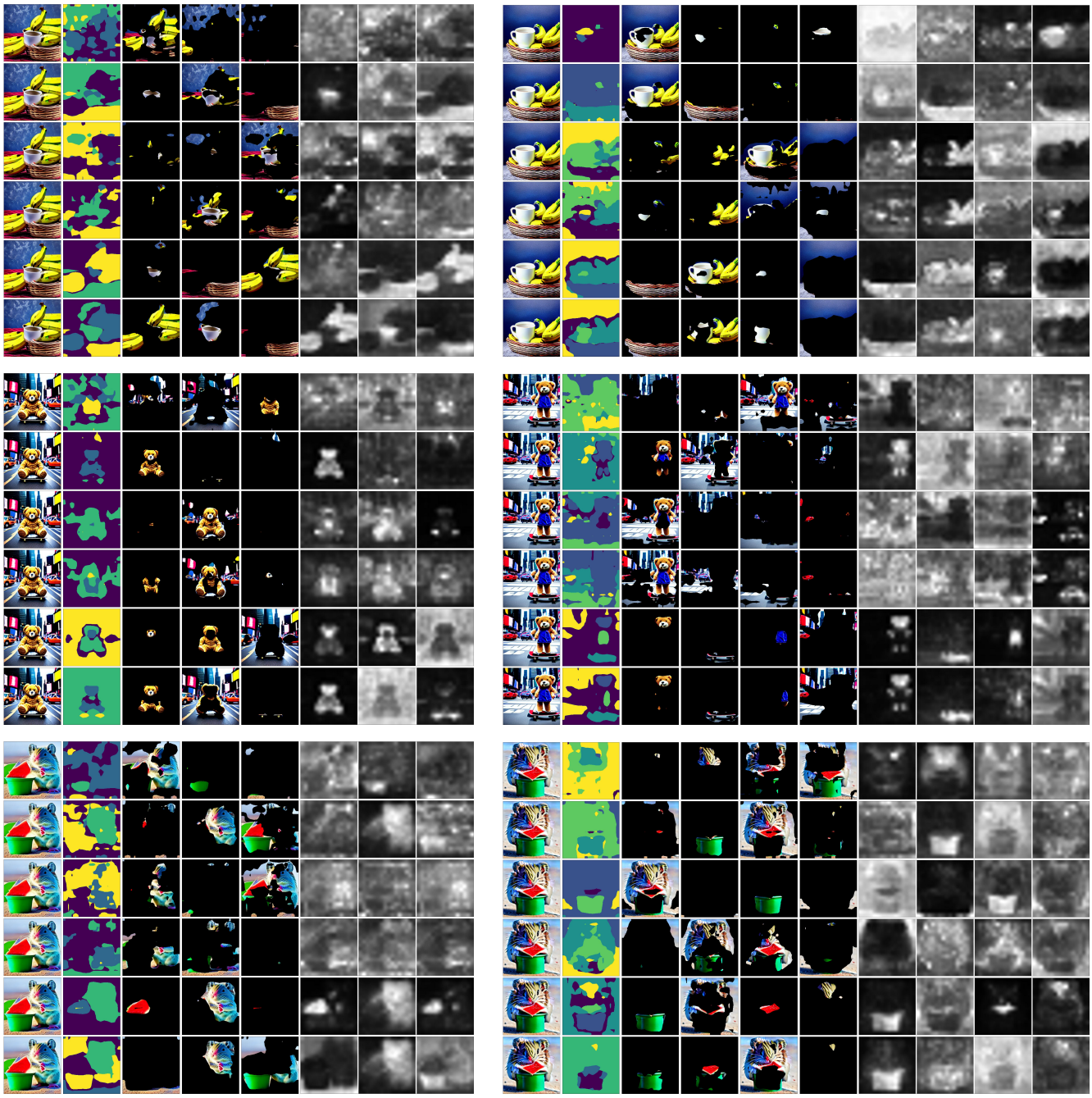

Figure 25. Visualisation of concepts (3 to 5 concepts). We compare all baseline methods across each row: 1) MCPL-all, 2) MCPL-one, 3) MCPL-all+*PromptCL+Bind adj.*, 4) MCPL-all+*AttnMask+PromptCL+Bind adj.*, 5) MCPL-one+*PromptCL+Bind adj.*, 6) MCPL-one+*AttnMask+PromptCL+Bind adj.*. Our findings indicate an increased challenge in learning multiple concepts as the number of objects in an image rises, particularly with a mask-free, language-driven approach like ours. Despite this, we **consistently observe improved accuracy in prompt-concept correlation when incorporating all proposed regularization terms (toward the bottom row)**.

A.7. Ablation study comparing MCPL-diverse versus MCPL-one in learning per-image different concept tasks

The *MCPL-diverse* training strategy has demonstrated potential in learning tasks with varying concepts per image. Therefore, we performed further experiments as detailed in Figure 26 confirming its efficacy.

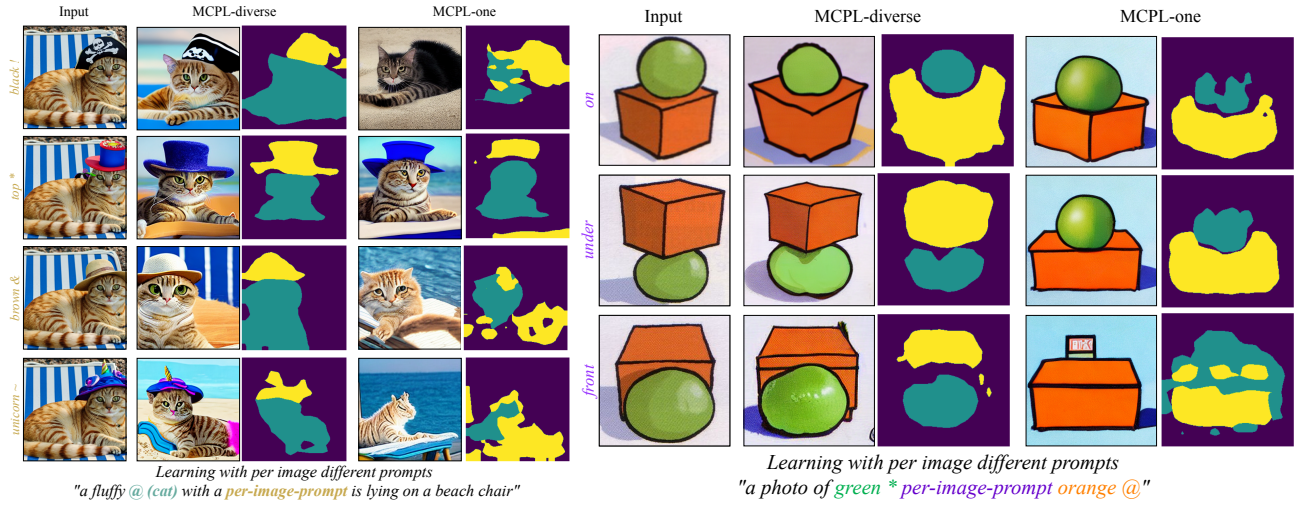

Figure 26. Visual comparison of MCPL-diverse versus MCPL-one in learning per-image different concept tasks: Left - cat with different hat example. Right - ball and box relationships example. **As MCPL-diverse are specially designed for such tasks, it outperforms MCPL-one, which fails to capture per image different relationships.**

A.8. Ablation study on the effect of Adjective words.

Our language-driven approach benefits from the proposed adjective binding mechanism. To better understand its role, we performed experiments removing adjective words, as shown in Figure 27, which confirmed its significance.

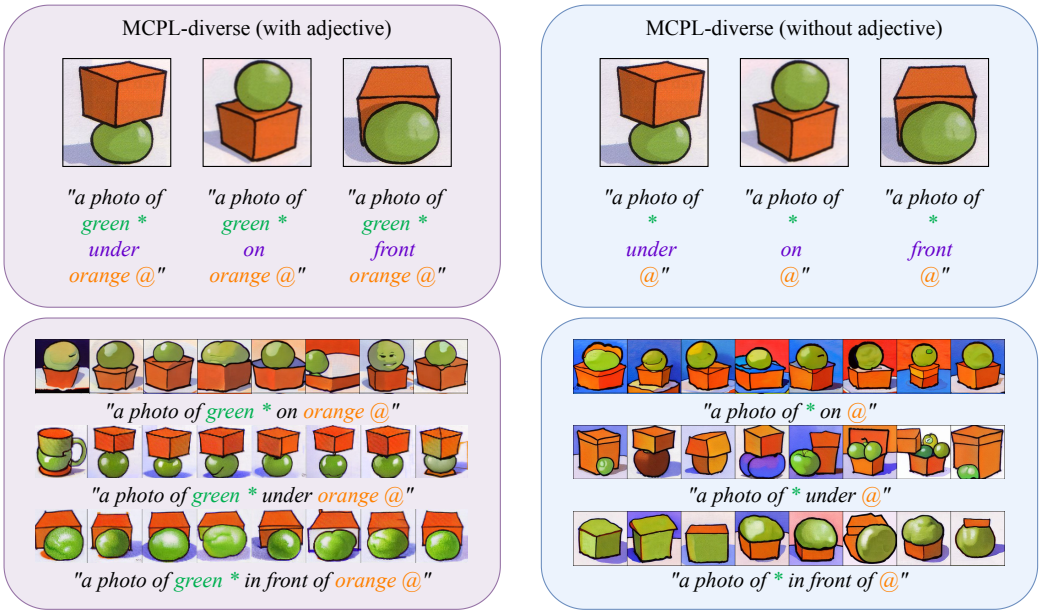

Figure 27. Visual comparison of MCPL-diverse with adjective word versus without adjective word. **Adjective words are crucial in linking each prompt to the correct region; without them, the model may struggle for regional guidance and we observe reduced performance.**

A.9. Visual comparison with BAS in composing complex scenes.

In tasks learning more than two concepts from a single image, we compare MCPL with Break-A-Scene (BAS). *Unlike BAS, MCPL neither uses segmentation masks as input nor updates model parameters*. To level the playing field, we adopted BAS's 'union sampling' training strategy, which randomly selects subsets of multi-concepts in each iteration. We manually prepared a set of cropped images of each individual concept and randomly selected subsets to combine. This approach, termed *'random crop,'* serves as our equivalent training strategy, see Figure 28 for an illustration. Given that each cropped image has a different number of concepts, we utilised our *MCPL-diverse*, designed to learn varying concepts per image. In Figure 28 and Figure 29 we showcase examples of such tasks against a set of competitive baselines.

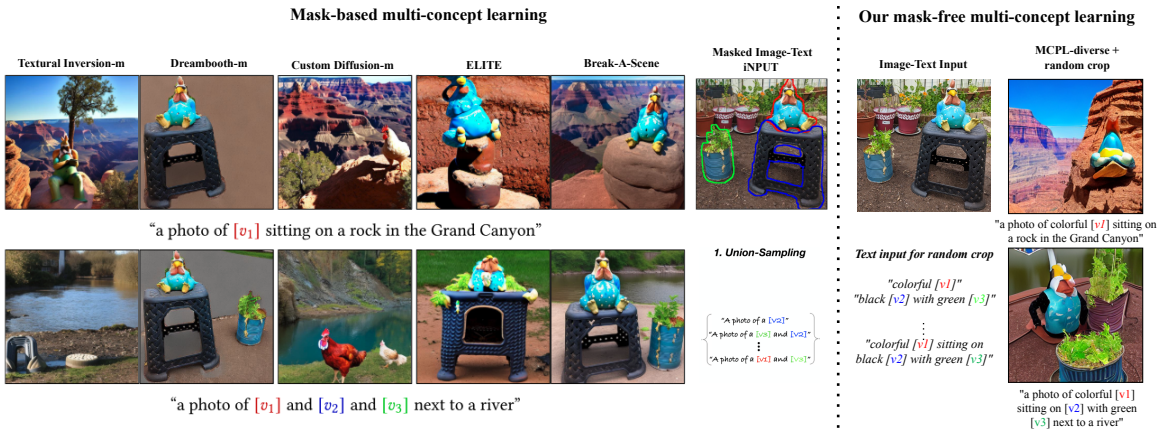

Figure 28. A qualitative comparison between our method (*MCPL-diverse*) and mask-based approaches: BAS, Textural Inversion (Gal et al., 2022) (masked), DreamBooth (Ruiz et al., 2022) (masked), Custom Diffusion (Kumari et al., 2023) (masked) and ELITE (Wei et al., 2023). *We stress our focus is not on optimising Diffusion Model (DM) parameters for enhanced composing performance*, unlike other compared methods. **Our approach delivers commendable results by learning from textural tokens (less than 0.1 MB) instead of updating the Diffusion Model of 4.9 GB as done in BAS, and it does so without depending on visual annotations such as masks.** *For tasks prioritising composition, our method can serve as an initial mask proposal to identify pertinent tokens, subsequently integrating finetuning-based techniques for refinement.* Images modified from BAS (Avrahami et al., 2023).

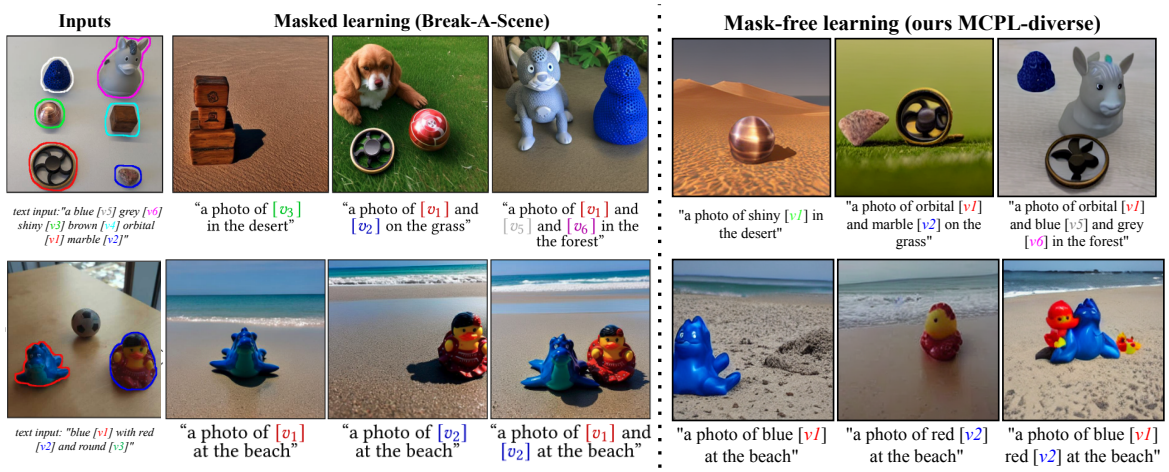

Figure 29. A qualitative comparison between BAS and our method (MCPL-one and MCPL-diverse). **Top**, learning six concepts from a single image and then composing a subset, is particularly challenging, which BAS has acknowledged. **Similar to BAS, our method also faces challenges with a high number of concepts, but it shows promising and competitive results. Bottom** learns three concepts from a single image. In this example, BAS performed better. Our MCPL-diverse, which neither uses mask inputs nor updates model parameters, showed decent results and was closer to BAS. Images modified from BAS (Avrahami et al., 2023).