ANCHORDIFF: TRAINING-FREE CONCEPT GROUNDING FOR DITS VIA ANCHOR-BASED GRAPH PROPAGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-Modal Diffusion Transformers (MM-DiTs) encode rich representations for training-free concept grounding, but existing attention-based methods often produce overlapping activations on visually confusable concepts—a failure mode we call concept leakage—where target responses spill over to non-target objects. To address this issue, we propose **AnchorDiff**, a training-free grounding method that decouples semantic localization from structural refinement. AnchorDiff selects a high-confidence anchor from concept-to-image attention map and propagates it as a one-hot seed over a hybrid graph derived from image-to-image self-attention. The graph uses output-space similarity for dense within-object propagation and a row-wise attention gate to suppress cross-object connections. Additionally, we introduce the **Multi-Concept Confusion Dataset**, which contains images with multiple visually similar concepts and separate masks, enabling explicit evaluation of concept leakage. Experiments show that AnchorDiff achieves strong grounding performance on ImageNet-Segmentation and PascalVOC, while substantially reducing concept leakage on our Multi-Concept Confusion Dataset. Code and dataset are available at this anonymous repository.

1 INTRODUCTION

Diffusion models have become powerful text-to-image generators (Ho et al., 2020; Rombach et al., 2022), yet extracting spatially grounded semantic information from their internal representations remains challenging. Prior interpretability work has mainly studied UNet-based diffusion models, where cross-attention maps provide useful localization cues (Tang et al., 2022; Hertz et al., 2022). However, the interpretability of more recent diffusion transformers remains less explored, despite their growing adoption in state-of-the-art generative models (Peebles & Xie, 2023).

Recent work has shown that DiT representations can support dense prediction and zero-shot segmentation (Helbling et al., 2025; Kim et al., 2025), suggesting that their internal attention structures encode rich semantic and spatial information (Esser et al., 2024). In particular, ConceptAttention (Helbling et al., 2025) computes output-space similarities between concept and image features and achieves strong zero-shot segmentation performance, in some cases surpassing CLIP-based methods (Radford et al., 2021; Chefer et al., 2021; Selvaraju et al., 2017; Gandelsman et al., 2023; Sun et al., 2024).

Dense concept responses, obtained from concept-image similarity or concept-to-image attention, are often not sufficiently object-specific in scenes with visually similar concepts. In visually confusable scenes, dense responses overlap across multiple similar objects, causing target activation to spill over to non-target objects. We call this failure mode *"concept leakage"*. As shown in Figure 1, existing training-free MM-DiT methods such as raw cross-attention maps, Seg4Diff (Kim et al., 2025), and ConceptAttention (Helbling et al., 2025) all exhibit varying degrees of concept leakage. ConceptAttention partially mitigates this by applying a pixelwise softmax over queried concepts, which improves contrast when distinct concepts are queried jointly, but the leakage among visually similar objects persists even under such competition. Moreover, with a single concept query, ConceptAttention relies on scene-dependent auxiliary background concepts such as "grass" or "sky" to produce a clear saliency map.

Our key observation is that image self-attention (Vaswani et al., 2017) in Multi-modal Diffusion Transformers (MM-DiTs) provides structural cues that can separate visually confusable objects. Row-wise patterns of image-to-image attention $A_{I\to I}$ reflect object boundaries: patches from different

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

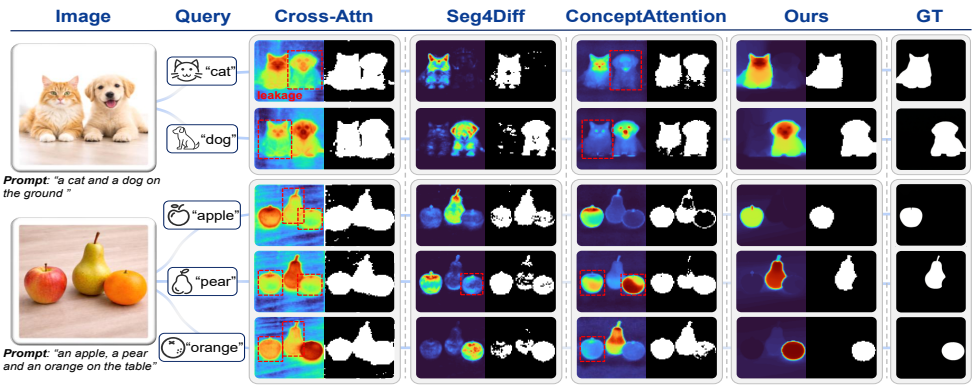

Figure 1: **Concept leakage in semantic grounding.** On visually similar concepts, existing methods suffer from concept leakage, where target responses spill over to non-target objects. AnchorDiff mitigates this by anchoring the target response and propagating it over an object-aware affinity graph.

regions tend to attend differently. In contrast, output features $O_{I\to I} = A_{I\to I}V$ provide dense within-object connectivity, but can also connect visually similar objects. We therefore combine these two signals by using row-wise attention similarity as a structural gate on output-space affinity, yielding a graph that preserves within-object connectivity while suppressing cross-object propagation.

Based on this observation, we propose **AnchorDiff**, a training-free semantic grounding method for MM-DiT models. Instead of using dense concept responses directly, AnchorDiff first selects the maximum-response location of concept-to-image attention $A_{C\to I}$ as a high-confidence semantic anchor. It then initializes a one-hot seed at this anchor and propagates it over the hybrid self-attention graph. This decouples semantic localization from structural refinement: $A_{C\to I}$ identifies where the target concept is likely to appear, while the $A_{I\to I}$-derived graph determines how activation should propagate within object boundaries. To explicitly evaluate concept leakage, we introduce the **Multi-Concept Confusion Dataset**, a benchmark of images with two or three visually similar foreground concepts and separate per-concept masks. Unlike standard foreground segmentation datasets, it evaluates whether a method can localize the queried concept while suppressing activation on similar non-target objects.

In summary, our contributions are:

- We identify concept leakage as a key limitation of attention-based semantic grounding in visually confusable scenes. To address it, we propose AnchorDiff, a training-free framework that combines $A_{C\to I}$-based sparse semantic anchoring with $A_{I\to I}$-based graph propagation for object-aware refinement.
- We show that MM-DiT image self-attention provides complementary cues for graph construction: row-wise attention similarity suppresses cross-object connections, while output-space affinity derived from self-attention outputs provides dense within-object connectivity.
- We introduce the Multi-Concept Confusion Dataset, a dedicated benchmark for evaluating concept leakage, containing images with multiple visually similar concepts and separate per-concept masks.
- We demonstrate that AnchorDiff achieves state-of-the-art training-free grounding performance on ImageNet-Segmentation, PascalVOC, and our Multi-Concept Confusion Dataset, improving localization quality while substantially reducing non-target activation.

2 Related Work

UNet-based diffusion grounding. Cross-attention maps in UNet-based diffusion models provide spatial localization of text concepts. DAAM (Tang et al., 2022) aggregates cross-attention across timesteps to produce word-level heatmaps. DiffSegmenter (Wang et al., 2025) segments Stable Diffusion outputs by multiplying a dense cross-attention score map with a raw self-attention map in a single step. OVAM (Marcos-Manchón et al., 2024) introduces token-optimized attention maps for open-vocabulary segmentation in SDXL (Podell et al.). These methods operate on UNet-based diffusion models with separate image and text processing streams, whereas AnchorDiff targets MM-

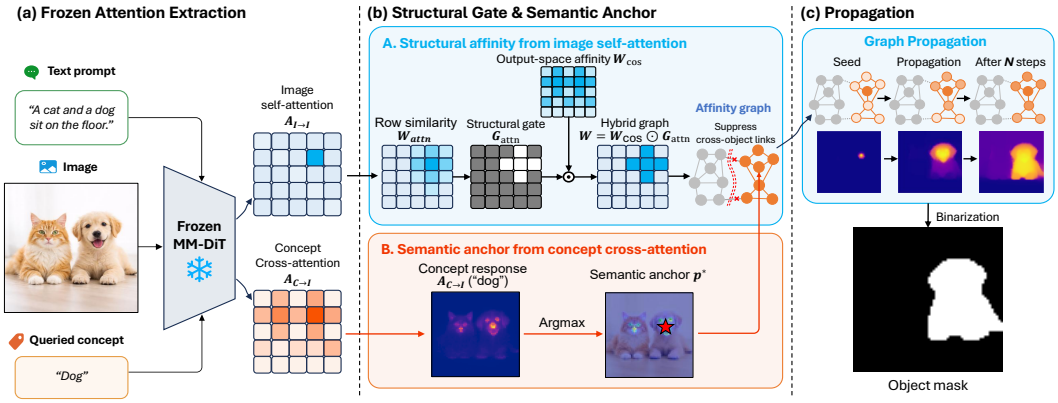

Figure 2: **Overview of AnchorDiff.** AnchorDiff extracts concept-to-image attention for anchor selection and image-to-image self-attention for graph construction. A one-hot seed at the semantic anchor is propagated over a hybrid graph combining output-space affinity with a row-wise attention gate, producing object-confined masks with reduced concept leakage.

DiT models with joint image-text attention, enabling the extraction of cross-modal and intra-modal attention signals.

DiT-based semantic grounding. Recent work has begun to repurpose DiT representations for segmentation and semantic grounding. ConceptAttention (Helbling et al., 2025) injects concept tokens as attention probes and computes output-space concept-image similarities with a pixelwise softmax over concepts. Seg4Diff (Kim et al., 2025) analyzes MM-DiT attention and identifies expert layers whose cross-modal attention aligns text tokens with spatially coherent image regions. While these methods reveal strong semantic localization signals in DiT representations, they obtain masks from dense output-space concept responses or dense cross-modal attention maps. Such dense responses can overlap in visually confusable scenes, leading to concept leakage. AnchorDiff instead uses concept-to-image attention only to select a sparse semantic anchor, and then propagates it through an object-aware graph derived from image self-attention.

3 Preliminaries

Before introducing AnchorDiff, we analyze the image-to-image self-attention of MM-DiT. Raw attention weights are too local for direct propagation, but row-wise attention patterns capture object boundaries, while self-attention output-space features provide within-object connectivity.

3.1 Locality of raw image-to-image attention

Let $A_{I\to I} \in \mathbb{R}^{N_{\text{img}}\times N_{\text{img}}}$ denote the image-to-image self-attention matrix extracted from MM-DiT. Although $A_{I\to I}$ captures interactions among image tokens, its weights are concentrated around local neighborhoods. On the Multi-Concept Confusion Dataset, the mean attention weight drops from 2.18×10^{-2} for neighboring patches to 8.6×10^{-5} for patches more than 16 tokens apart, as shown in Figure 3. Thus, raw $A_{I\to I}$ alone provides insufficient long-range connectivity for object propagation.

3.2 Object structure from row-wise attention similarity

Although individual entries of $A_{I\to I}$ are dominated by local neighbors, each row $A_{I\to I}[i,:]$ summarizes how patch i relates to all image tokens. Patches from the same object tend to have similar row patterns, while patches across object boundaries differ. As shown in Figure 4, PCA projections and clustering of attention rows form spatially coherent object regions without using concept labels. We therefore define row-wise structural similarity as

$$W_{\text{attn}}[i,j] = \cos(A_{I\to I}[i,:], A_{I\to I}[j,:])\,. \tag{1}$$

On the Multi-Concept Confusion Dataset, W_{attn} is higher for same-object pairs (21.83%) than for confusable different-object pairs (4.65%) or foreground-background pairs (1.13%) (Table 1), indicating that it provides a concept-agnostic cue for suppressing cross-object connections.

3.3 Output-space affinity for dense propagation

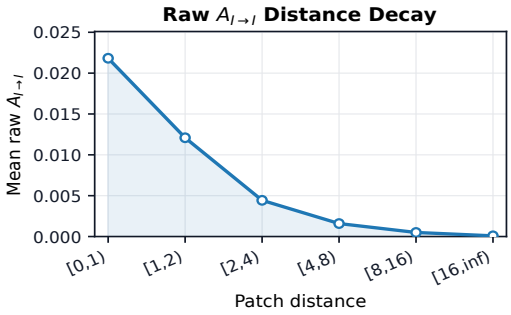

Figure 3: **Raw image-to-image attention is spatially local.** Mean raw attention weights are measured over patch-pair distance bins and decay rapidly on the 64×64 grid.

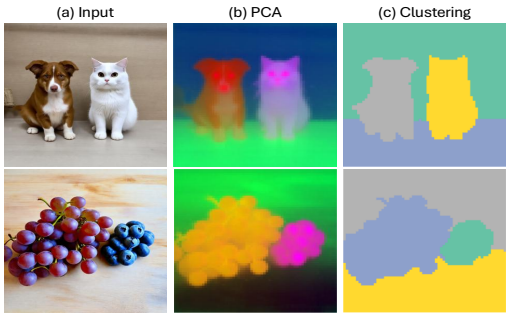

Figure 4: **Row-wise consistency in image self-attention.** PCA projections and clustering of $A_{I\to I}$ rows form spatially coherent object regions, without using concept labels.

MM-DiT self-attention also produces output-space features $O_{I\to I} = A_{I\to I}V$, where V denotes the Value projection of image tokens. We define the output-space affinity as:

$$W_{\cos}[i,j] = \cos(O_{I\to I}[i,:], O_{I\to I}[j,:]) \,. \quad (2)$$

Compared with raw attention weights or row-wise attention similarity alone, $W_{\cos}$ provides denser long-range connectivity among the same-object patches, making it suitable for propagating activation across the entire object. As shown in Figure 5, output-space similarity produces dense responses over same-object regions but also activates visually similar non-target objects, causing cross-object leakage. The statistics in Table 1 further quantify this trade-off. $W_{\cos}$ assigns high affinity to same-object pairs (79.40%), but also to confusable different-object pairs (67.27%) and foreground-background pairs (41.80%). In contrast, gating with G_{attn} reduces cross-object and foreground-background affinities to 0.04% and 0.03%. We therefore use $W_{\cos}$ as the propagation weight and thresholded W_{attn} as the structural gate to obtain G_{attn}.

Table 1: **Affinity statistics.** Mean affinities (%) for same-object, confusable different-object, and foreground-background patch pairs.

Affinity	Same obj.	Conf. diff.	FG-BG
W_{attn}	21.83	4.65	1.13
$W_{\cos}$	79.40	67.27	41.80
$W_{\cos} \odot G_{\text{attn}}$	10.14	0.04	0.03

Table 2: **Anchor localization accuracy.** The selected anchor is correct if it lies inside the ground-truth mask of the queried concept.

Model	Correct	Total	Rate (%)
SD3	2532	2842	89.09
SD3.5	2565	2842	90.25

4 METHOD

We propose AnchorDiff, a training-free grounding method for MM-DiT models. As shown in Figure 2, AnchorDiff decouples semantic localization from structural refinement: it selects a high-confidence anchor from concept-to-image attention map $A_{C\to I}$ and propagates a one-hot seed over a hybrid affinity graph constructed from image self-attention signals $A_{I\to I}$ and $O_{I\to I}$.

4.1 ATTENTION EXTRACTION

Given a text prompt and target concepts $C = \{c_1, \ldots, c_K\}$, we extract image-to-image self-attention for structural propagation and concept-to-image attention for semantic anchor selection. From the standard image-text denoising process, we record

$$A^{(\ell)}_{I\to I} = \text{softmax}\left(\frac{Q^{(\ell)}_{\text{img}}\left(K^{(\ell)}_{\text{img}}\right)^{\top}}{\sqrt{d_h}}\right) \in \mathbb{R}^{N_{\text{img}}\times N_{\text{img}}}, \quad (3)$$

where ℓ indexes the transformer layer, N_{img} is the number of image tokens, and d_h is the attention head dimension. We also record the corresponding self-attention output features

$$O^{(\ell)}_{I\to I} = A^{(\ell)}_{I\to I}V^{(\ell)}_{\text{img}} \in \mathbb{R}^{N_{\text{img}}\times d_h}. \quad (4)$$

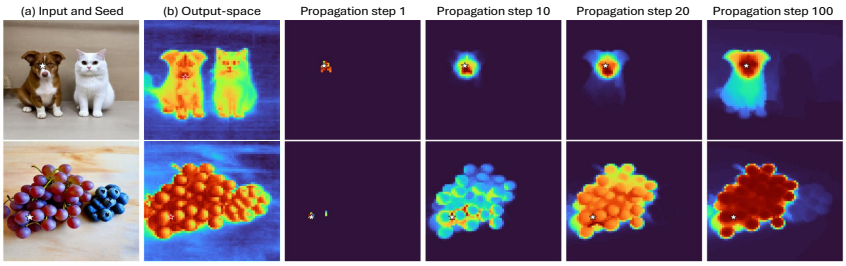

Figure 5: **Output-space propagation with structural gating.** Output-space affinity yields dense same-object responses but also activates similar non-target objects. After gating with G_{attn}, propagation from a one-hot anchor recovers the target object while reducing cross-object leakage.

We then perform a second forward pass with the target concept tokens injected following (Helbling et al., 2025) and extract concept-to-image attention

$$A^{(\ell)}_{C\to I} = \text{softmax}\left(\frac{Q^{(\ell)}_{\text{concept}}\left(K^{(\ell)}_{\text{img}}\right)^{\top}}{\sqrt{d_h}}\right) \in \mathbb{R}^{K\times N_{\text{img}}}, \tag{5}$$

where each row gives a coarse response for one queried concept. Finally, we aggregate the extracted attention maps and output features over selected layers for anchor selection and graph construction:

$$A_{C\to I} = \frac{1}{|L|}\sum_{\ell\in L} A^{(\ell)}_{C\to I}, \quad A_{I\to I} = \frac{1}{|L|}\sum_{\ell\in L} A^{(\ell)}_{I\to I}, \quad O_{I\to I} = \frac{1}{|L|}\sum_{\ell\in L} O^{(\ell)}_{I\to I}. \tag{6}$$

4.2 Semantic Anchor Selection

Dense concept responses usually overlap in visually confusable scenes. Instead of using the full response map from $A_{C\to I}$, AnchorDiff retains only the maximum-response location for each concept. For concept c_k, we define

$$h_{c_k} = A_{C\to I}[k,:], \quad p^*_k = \arg\max_i h_{c_k}[i]. \tag{7}$$

We use p^*_k as a semantic anchor and initialize a one-hot seed at this location for graph propagation. To verify the reliability of this anchor, we measure whether p^*_k falls inside the ground-truth mask of the queried concept c_k. On the Multi-Concept Confusion Dataset, the anchor localization accuracy reaches 89.09% for SD3 and 90.25% for SD3.5-Large over 2842 image-concept pairs, as shown in Table 2. This indicates that the peak of $A_{C\to I}$ usually provides a reliable seed for graph propagation.

4.3 Hybrid Affinity Graph Construction

We construct a hybrid affinity graph from $A_{I\to I}$ and $O_{I\to I}$. We first compute row-wise attention similarity:

$$W_{\text{attn}}[i,j] = \frac{A_{I\to I}[i,:]\cdot A_{I\to I}[j,:]}{\|A_{I\to I}[i,:]\|\,\|A_{I\to I}[j,:]\|}. \tag{8}$$

We binarize W_{attn} into a structural gate by retaining only the top 2% entries:

$$G_{\text{attn}}[i,j] = \begin{cases} 1, & W_{\text{attn}}[i,j] > \tau_W, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

where τ_W is the 98th percentile of W_{attn} entries. We then compute output-space affinity from $O_{I\to I}$:

$$W_{\cos}[i,j] = \frac{O_{I\to I}[i,:]\cdot O_{I\to I}[j,:]}{\|O_{I\to I}[i,:]\|\,\|O_{I\to I}[j,:]\|}. \tag{10}$$

The final graph uses $W_{\cos}$ as propagation weights and G_{attn} as the structural gate:

$$W[i,j] = W_{\cos}[i,j]\cdot G_{\text{attn}}[i,j], \quad \hat{W}[i,:] = \frac{W[i,:]}{\sum_j W[i,j]}. \tag{11}$$



Figure 6: Representative samples from our Multi-Concept Confusion Dataset. Each image contains two or three visually similar concepts, with per-concept mask overlaps shown in different colors.

4.4 One-hot anchor propagation

We initialize a one-hot seed $s^{(0)} \in \mathbb{R}^{N_{\text{img}}}$ at the anchor p_k^*, with $s^{(0)}_{p_k^*} = 1$ and all other entries set to 0. We then propagate it over the normalized affinity graph:

$$s^{(n+1)} = \hat{W}^T s^{(n)}, \quad n = 0, \ldots, N_{\text{diff}} - 1. \tag{12}$$

The structural gate restricts propagation to structurally consistent high-affinity regions. After N_{diff} steps, we obtain

$$\hat{h}_{c_k} = \text{minmax}\big(s^{(N_{\text{diff}})}\big), \tag{13}$$

where $\text{minmax}(\cdot)$ linearly normalizes the response to $[0, 1]$. We obtain the final binary mask by thresholding $\hat{h}_{c_k}$ at its mean value.

5 Experiments

We evaluate AnchorDiff in two settings: single-concept grounding on standard benchmarks, and multi-concept grounding under visually similar distractors using our Multi-Concept Confusion Dataset.

5.1 Experimental setup

Implementation details. We apply AnchorDiff to two frozen MM-DiT backbones, SD3 Medium and SD3.5 Large (Esser et al., 2024). All experiments are training-free. All images are resized to 1024×1024, corresponding to a 64×64 image-token grid with $N_{\text{img}} = 4096$ image tokens. AnchorDiff extracts masks from MM-DiT attention at the midpoint of the generation trajectory. For existing image benchmarks, we apply the same midpoint extraction along the inversion trajectory. This single-timestep extraction avoids full-trajectory attention aggregation and keeps inference efficient. For semantic anchoring, we aggregate $A_{C \to I}$ over all transformer layers to choose the target-concept anchor, since anchor selection only requires the $\arg\max$ location and benefits from broad semantic averaging. For graph construction, we use image-to-image attention layers $\{9, 18\}$ for SD3 and $\{23, 31\}$ for SD3.5, as graph propagation is sensitive to layer-specific structural quality (Figure 9). In experiments, we retain the top 2% of affinities ($\rho = 0.02$), with τ_W set to the 98th percentile of W_{attn} entries, and use $N_{\text{diff}} = 160$ propagation steps for both backbones. All experiments are run on NVIDIA GeForce RTX 4090 GPUs.

Datasets. For single-concept grounding, we evaluate on ImageNet-Segmentation (Guillaumin et al., 2014) and PascalVOC 2012 Single-Class (Everingham et al., 2015), where the latter is constructed by selecting PascalVOC 2012 validation images with exactly one annotated foreground class. To evaluate concept leakage, we construct a Multi-Concept Confusion Dataset with 1,272 images containing two or three visually similar foreground concepts, each with a separate mask. The dataset includes synthetic images from FLUX.1-dev and Seedream 3.0/4.0 (Gao et al., 2025; Seedream et al., 2025), as well as real web images. We use Qwen2.5-VL (Bai et al., 2025) to localize each concept and SAM2 (Ravi et al.) to segment the corresponding masks, followed by manual inspection. Representative examples are shown in Figure 6.

Table 3: Main results on standard zero-shot segmentation benchmarks. We report Acc, mIoU, and mAP on ImageNet-Segmentation and PascalVOC Single-Class.

Method	ImageNet-Segmentation			PascalVOC Single-Class		
	Acc ↑	mIoU ↑	mAP ↑	Acc ↑	mIoU ↑	mAP ↑
LRP (Binder et al., 2016)	50.55	33.82	55.92	48.02	30.71	53.47
Partial-LRP (Binder et al., 2016)	75.23	60.29	83.21	71.50	51.38	84.37
Rollout (Abnar & Zuidema, 2020)	70.51	54.46	81.73	69.81	51.25	85.33
ViT Attention (Dosovitskiy et al., 2020)	70.62	54.58	78.86	68.51	44.81	83.63
GradCAM (Selvaraju et al., 2017)	68.44	52.02	70.68	70.50	44.97	76.92
TextSpan (Gandelsman et al., 2023)	75.81	61.05	80.86	75.00	56.24	84.81
TransInterp (Chefer et al., 2021)	78.43	64.51	84.36	76.98	57.24	86.77
CLIPasRNN (Sun et al., 2024)	83.05	71.01	86.59	83.44	70.79	90.98
DINO SA (Caron et al., 2021)	81.76	69.14	86.05	79.51	62.62	88.53
DINOv2 SA (Oquab et al., 2024)	83.20	71.24	85.19	81.72	64.76	87.94
DINOv2 Reg SA (Darcet et al., 2024)	84.21	73.24	86.62	83.36	67.27	88.97
iBOT SA (Zhou et al., 2021)	83.18	71.21	86.51	80.91	64.95	89.07
DAAM SDXL (Tang et al., 2022)	82.33	69.97	90.48	74.63	58.18	88.79
DAAM SD2 (Tang et al., 2022)	83.78	72.09	88.95	78.80	63.29	89.55
OVAM (Marcos-Manchón et al., 2024)	83.88	72.24	89.65	80.40	65.71	91.92
Cross Attention SD3	80.90	67.92	86.79	82.92	68.77	92.16
Cross Attention SD3.5	76.52	61.97	89.37	74.23	58.01	92.44
Seg4Diff SD3 (Kim et al., 2025)	82.13	69.68	83.84	83.84	69.96	87.28
Seg4Diff SD3.5 (Kim et al., 2025)	79.53	66.01	83.29	82.97	67.88	87.36
ConceptAttention SD3 (Helbling et al., 2025)	81.48	68.75	85.52	85.77	72.04	88.48
ConceptAttention SD3.5 (Helbling et al., 2025)	83.75	72.04	88.48	85.11	71.56	92.24
Ours SD3	83.76	75.09	90.06	87.30	79.74	93.15
Ours SD3.5	**85.46**	**77.29**	**91.20**	**88.50**	**81.53**	**94.05**

Metrics. For ImageNet-Segmentation and PascalVOC Single-Class, we follow the standard binary foreground segmentation protocol and report Acc, mIoU, and mAP over the whole image, treating the queried class as foreground and all remaining pixels as background. For the Multi-Concept Confusion Dataset, we evaluate each image once per target concept. Given a target mask M_c, we define M_{other} as the union of all non-target concept masks, excluding pixels overlapping with M_c. Unlike the standard metrics, mIoU_{fg}, mAP_{fg}, and Acc_{fg} are computed only on the annotated foreground region $M_c \cup M_{\text{other}}$, measuring whether the queried target concept is separated from visually similar non-target concepts without involving background pixels. We also report the Non-target Activation Ratio (NAR):

$$\text{NAR} = \frac{\sum_{p \in M_{\text{other}}} H_c(p)}{\sum_{p \in M_c \cup M_{\text{other}}} H_c(p)}, \tag{14}$$

where H_c is the response map for concept c. Lower NAR indicates less activation leakage to visually similar non-target objects. All multi-concept metrics are averaged over image-concept pairs.

Baselines. We compare with three groups of baselines: CLIP/ViT interpretability methods, including LRP (Binder et al., 2016), Partial-LRP (Binder et al., 2016), Rollout (Abnar & Zuidema, 2020), GradCAM (Selvaraju et al., 2017), ViT Attention (Dosovitskiy et al., 2020), TextSpan (Gandelsman et al., 2023), TransInterp (Chefer et al., 2021), and CLIPasRNN (Sun et al., 2024); self-supervised ViT self-attention baselines, including DINO SA (Caron et al., 2021), DINOv2 SA (Oquab et al., 2024), DINOv2 Reg SA (Darcet et al., 2024), and iBOT SA (Zhou et al., 2021); and diffusion-based grounding methods, including DAAM (Tang et al., 2022), OVAM (Marcos-Manchón et al., 2024), Seg4Diff (Kim et al., 2025), ConceptAttention (Helbling et al., 2025), and raw SD3/SD3.5 cross-attention maps. For the Multi-Concept Confusion Dataset, we report only methods that can generate a separate response map for each queried concept.

5.2 Main results

Concept leakage evaluation. Table 4 evaluates whether methods can localize a queried concept while suppressing visually similar non-target concepts. Ours SD3.5 achieves 68.89% mIoU_{fg} and 90.02% Acc_{fg}, outperforming Seg4Diff SD3.5 by 22.32 mIoU points and 20.83 accuracy points. It also reduces NAR to 12.20, compared with 19.08 for Seg4Diff SD3.5 and 26.36 for ConceptAttention SD3.5, indicating substantially less activation leakage. These results show that dense semantic responses are prone to concept leakage, while AnchorDiff produces more object-confined masks by propagating a sparse semantic anchor over a structurally gated graph.

Standard single-concept grounding. Table 3 reports results on ImageNet-Segmentation and PascalVOC Single-Class. AnchorDiff consistently outperforms all compared baselines on both benchmarks. Ours SD3.5 achieves 77.29% mIoU on ImageNet-Segmentation and 81.53% mIoU

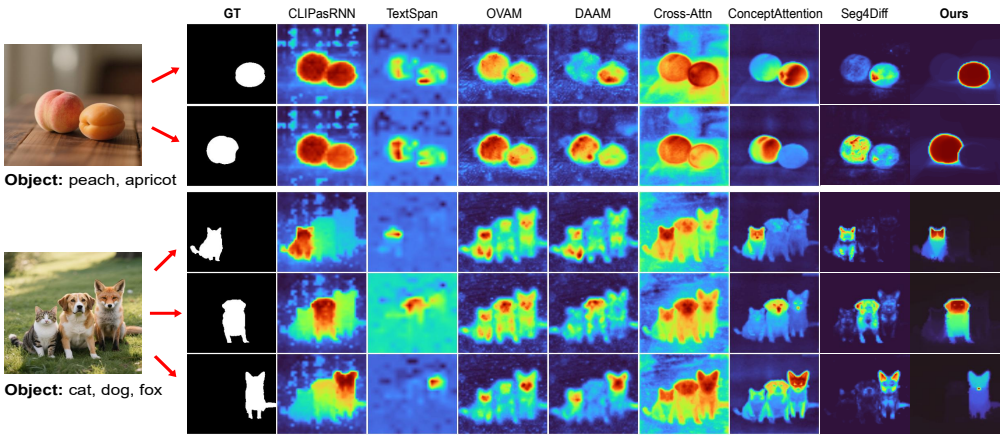

Figure 7: Qualitative results on representative samples from the Multi-Concept Confusion Dataset.

Table 4: Results on the Multi-Concept Confusion Dataset. We report mIoU_{fg}, mAP_{fg}, and Acc_{fg}, averaged over image-concept pairs. Higher is better except for NAR.

Method	$\mathrm{mIoU}_{\mathrm{fg}}\uparrow$	$\mathrm{mAP}_{\mathrm{fg}}\uparrow$	NAR $\downarrow$	$\mathrm{Acc}_{\mathrm{fg}}\uparrow$
CLIPasRNN	25.82	74.89	47.20	46.81
TextSpan	25.47	59.53	45.99	52.38
OVAM	26.28	61.76	49.58	45.76
DAAM SD2	27.10	71.03	44.08	47.02
DAAM SDXL	24.46	64.12	50.66	45.31
Cross Attention SD3	33.93	83.70	32.49	54.62
Cross Attention SD3.5	28.37	88.06	28.45	65.97
Seg4Diff SD3	44.56	82.06	22.47	65.88
Seg4Diff SD3.5	46.57	84.22	19.08	69.19
ConceptAttention SD3	38.24	80.63	32.95	55.26
ConceptAttention SD3.5	39.97	88.73	26.36	59.28
Ours SD3	64.13	87.27	16.86	86.24
Ours SD3.5	**68.89**	**89.38**	**12.20**	**90.02**

Table 5: Component ablation on the Multi-Concept Confusion Dataset. We evaluate the contribution of each component under the same inference setting. All variants are evaluated on the full dataset using the same metrics as Table 4.

Model	Settings	$\mathrm{mIoU}_{\mathrm{fg}}\uparrow$	$\mathrm{mAP}_{\mathrm{fg}}\uparrow$	NAR $\downarrow$	$\mathrm{Acc}_{\mathrm{fg}}\uparrow$
SD3	$A_{C\to I}$ only	34.61	71.66	35.51	54.91
	w/o $A_{I\to I}$ gate	7.63	9.06	55.11	50.31
	w/o $W_{\cos}$	62.65	86.87	17.72	85.71
	L9 only	55.32	80.47	13.78	84.32
	L18 only	42.28	82.06	25.23	70.92
	Ours full	64.13	87.27	16.86	86.24
SD3.5	$A_{C\to I}$ only	29.91	80.11	37.82	51.82
	w/o $A_{I\to I}$ gate	11.97	14.42	55.05	49.78
	w/o $W_{\cos}$	66.71	88.32	12.89	89.62
	L23 only	57.00	79.91	10.75	88.82
	L31 only	53.44	86.36	17.59	79.19
	Ours full	68.89	89.38	12.20	90.02

on PascalVOC Single-Class, surpassing prior diffusion-based grounding methods such as DAAM, OVAM, Seg4Diff, and ConceptAttention. These results show that the proposed anchor-based graph propagation improves localization quality not only in confusable multi-concept scenes, but also under the standard foreground segmentation protocol.

5.3 Ablation Studies

We conduct ablation studies to analyze the main components, propagation steps, and layer selection on the Multi-Concept Confusion Dataset, and the effect of auxiliary background concepts on ImageNet-Segmentation.

Main components. Table 5 evaluates the main components of our pipeline. The variant $A_{C\to I}$-only directly uses the concept-to-image attention map without graph propagation. Its substantially lower mIoU_{fg} indicates that the raw concept-to-image attention map alone is insufficient for separating confusable objects. Removing the $A_{I\to I}$ attention gate leads to the largest degradation for both backbones. This indicates that output-space similarity alone is not target-specific and can propagate activation to visually or semantically similar objects. In contrast, removing $W_{\cos}$ remains competitive but underperforms the full model in mIoU_{fg} and mAP_{fg}, suggesting that the row-wise attention gate suppresses cross-object leakage, while output-space affinity improves within-object propagation.

Propagation steps. We study the effect of the number of graph propagation steps in Figure 8. With all other parameters fixed, performance largely saturates around 160–200 steps. We choose 160 as the default propagation budget to balance target coverage, leakage suppression, and computational cost. After that, both metrics gradually decrease, while NAR continues to increase, indicating that excessive propagation can spread activation beyond the target object through weak residual connections. $\mathrm{mAP}_{\mathrm{fg}}$ also rises quickly at early steps and then remains relatively stable. These results suggest that a moderate propagation length is sufficient for object coverage, whereas overly long propagation reduces object specificity.

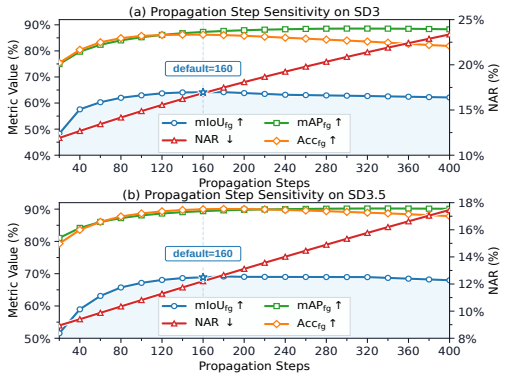

Figure 8: Propagation step sensitivity on SD3 and SD3.5.

Table 6: Single-concept grounding on ImageNet-Segmentation using SD3 and SD3.5. We compare ConceptAttention and AnchorDiff with and without auxiliary background concepts. AnchorDiff works directly with a single target concept, and auxiliary background concepts have only marginal impact across backbones.

Model	Settings	mIoU↑	Acc↑
SD3	ConceptAttention (single)	54.71	70.72
	ConceptAttention (w/o softmax)	58.87	74.11
	ConceptAttention (with BG)	68.75	81.48
	Ours (single)	75.09	83.76
	Ours (with BG)	74.88	83.73
SD3.5	ConceptAttention (single)	55.24	71.24
	ConceptAttention (w/o softmax)	58.68	73.96
	ConceptAttention (with BG)	72.04	83.75
	Ours (single)	77.29	85.46
	Ours (with BG)	77.54	85.63

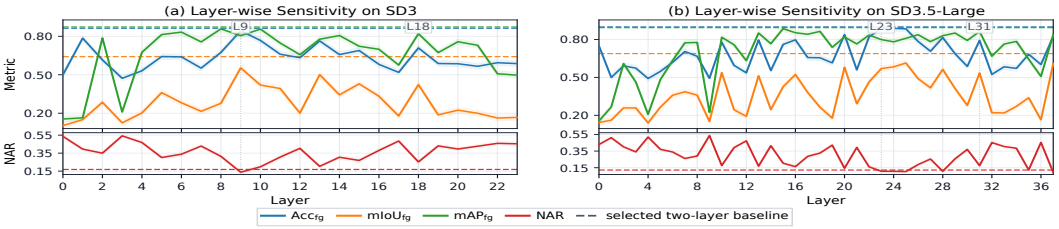

Figure 9: Layer-wise sensitivity analysis on the Multi-Concept Confusion Dataset. We evaluate single-layer image-to-image attention features across the transformer depth for SD3 and SD3.5-Large. Dashed lines indicate the selected two-layer configurations used in the full method.

Layer selection. We analyze the effect of image-to-image layer selection in Table 5 and Figure 9. The single-layer results show that different layers provide complementary cues. L9 is stronger than L18 for SD3, while L23 and L31 show different trade-offs for SD3.5. Their two-layer fusions achieve the best $\mathrm{mIoU}_{\mathrm{fg}}$ among two-layer configurations. We additionally evaluate all two-layer combinations among the top-performing single layers, as well as randomly sampled three-layer and four-layer combinations, with the full sweep on the Multi-Concept Confusion Dataset reported in Section A.3 of the appendix. The selected two-layer pairs achieve the best performance among the evaluated two-layer configurations on the Multi-Concept Confusion Dataset and are used as fixed defaults across all benchmarks. While a few multi-layer variants achieve marginally higher scores, the gains are small and we adopt the simpler two-layer configuration in the main experiments. This suggests that layer fusion benefits from complementary structural cues across depths rather than simply selecting the strongest individual layers.

Single-concept analysis. Table 6 evaluates the effect of auxiliary background concepts. ConceptAttention degrades with a single concept and benefits substantially from auxiliary background concepts, while AnchorDiff remains strong without them. Unlike methods that require background concepts to create spatial contrast, AnchorDiff uses the target response only to select a sparse anchor and obtains the mask through structural propagation. The marginal changes from adding background concepts indicate that AnchorDiff does not rely on auxiliary background concepts.

6 Conclusion

We presented AnchorDiff, a training-free semantic grounding method for MM-DiT models. AnchorDiff selects a high-confidence target-concept anchor and propagates it over a hybrid graph built from image-to-image self-attention and output-space features, decoupling semantic localization from structural refinement. Experiments show that AnchorDiff achieves strong grounding performance on ImageNet-Segmentation and PascalVOC, while substantially reducing concept leakage on our Multi-Concept Confusion Dataset. One limitation is that AnchorDiff can inherit errors from the initial semantic anchor when concept-to-image attention is unreliable, suggesting that more robust semantic initialization remains important for training-free grounding. Additionally, since AnchorDiff selects a single anchor per concept, it may fail to cover all instances when multiple same-category instances are spatially distant. Extending it to multi-anchor selection is a natural direction for future work.

AI USE STATEMENT

We used FLUX.1-dev and Seedream 3.0/4.0 to generate synthetic images for the Multi-Concept Confusion Dataset. We also used Qwen2.5-VL to localize concept instances and SAM2 to generate the corresponding segmentation masks, followed by manual inspection, as described in the paper.

During the preparation of this work, AI tools were used solely to assist with manuscript polishing and the implementation of method-related code. All AI-assisted content, including textual revisions and code suggestions, was manually reviewed and verified by the authors. The research problem, methodological design, experimental setup, evaluation protocol, analysis, and conclusions were independently developed and determined by the authors. The authors take full responsibility for the final manuscript and its technical content.

REPRODUCIBILITY STATEMENT

The method and experimental setup are described in Sections 4 and 5, with additional experimental results and analyses provided in the appendix. Code and the Multi-Concept Confusion Dataset are available through the anonymized repository: `https://anonymous.4open.science/r/AnchorDiff-7E38`.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4190–4197, 2020.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International conference on artificial neural networks*, pp. 63–71. Springer, 2016.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 782–791, 2021.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 2020.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv e-prints*, pp. arXiv–2403, 2024.

Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. *arXiv preprint arXiv:2310.05916*, 2023.

Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025.

Matthieu Guillaumin, Daniel Küttel, and Vittorio Ferrari. Imagenet auto-annotation with segmentation propagation. *International Journal of Computer Vision*, 110(3):328–348, 2014.

Alec Helbling, Tuna Meral, Benjamin Hoover, Pinar Yanardag, and Polo Chau. Conceptattention: Diffusion transformers learn highly interpretable features. In *International Conference on Machine Learning*, 2025.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Chaehyun Kim, Heeseong Shin, Eunbeen Hong, Heeji Yoon, Anurag Arnab, Paul Hongsuck Seo, Sunghwan Hong, and Seungryong Kim. Seg4diff: Unveiling open-vocabulary semantic segmentation in text-to-image diffusion transformers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Pablo Marcos-Manchón, Roberto Alcover-Couso, Juan C SanMiguel, and Jose M Martínez. Open-vocabulary attention maps with token optimization for semantic segmentation in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9242–9252, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations*.

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*, pp. 10684–10695, 2022.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. Clip as rnn: Segment countless visual concepts without training endeavor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13171–13182, 2024.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv*, 2017.

Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *IEEE Transactions on Image Processing*, 2025.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

A TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

A.1 ADDITIONAL DATASET SAMPLES

Figure 10 provides more examples from the Multi-Concept Confusion Dataset. Each image contains two or three visually confusable concepts with separate per-concept masks, shown as semi-transparent overlays. These samples highlight the fine-grained visual ambiguities used to evaluate concept leakage. The dataset can be downloaded from this anonymous repository.

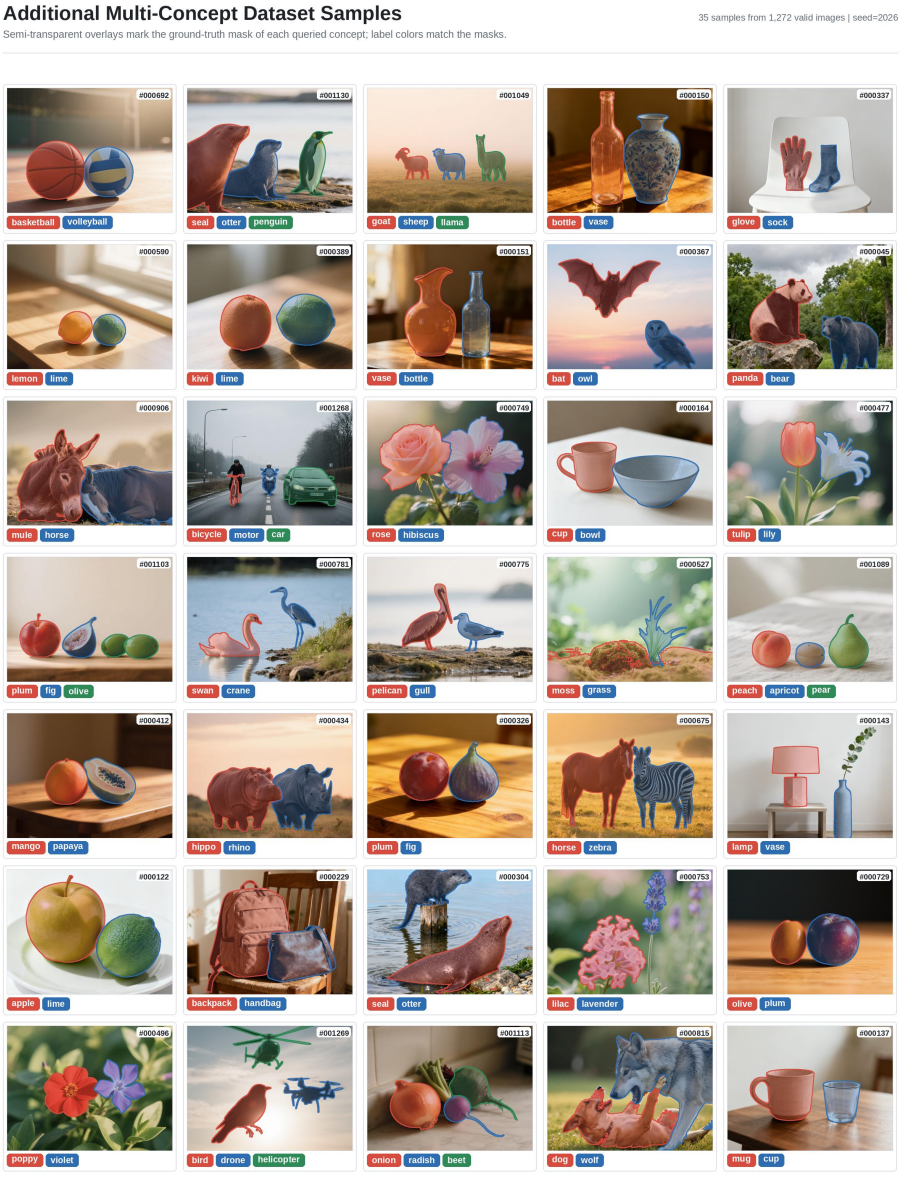

Figure 10: Additional samples from the Multi-Concept Confusion Dataset. Each image contains two or three visually confusable concepts with one binary mask per concept.

A.2 ADDITIONAL QUALITATIVE COMPARISONS

Two-concept qualitative comparisons. We provide additional qualitative comparisons on two-concept confusion examples. Each example contains two visually similar concepts, and each concept is evaluated independently with both heatmap and thresholded-mask visualizations. AnchorDiff produces more target-confined responses and reduces activation leakage to visually similar non-target concepts.

Three-concept qualitative comparisons. We provide additional qualitative comparisons on three-concept confusion examples. These examples are more challenging because the queried concept must be separated from two visually or semantically similar non-target concepts. AnchorDiff remains more object-confined across categories, while other methods often activate multiple similar objects.

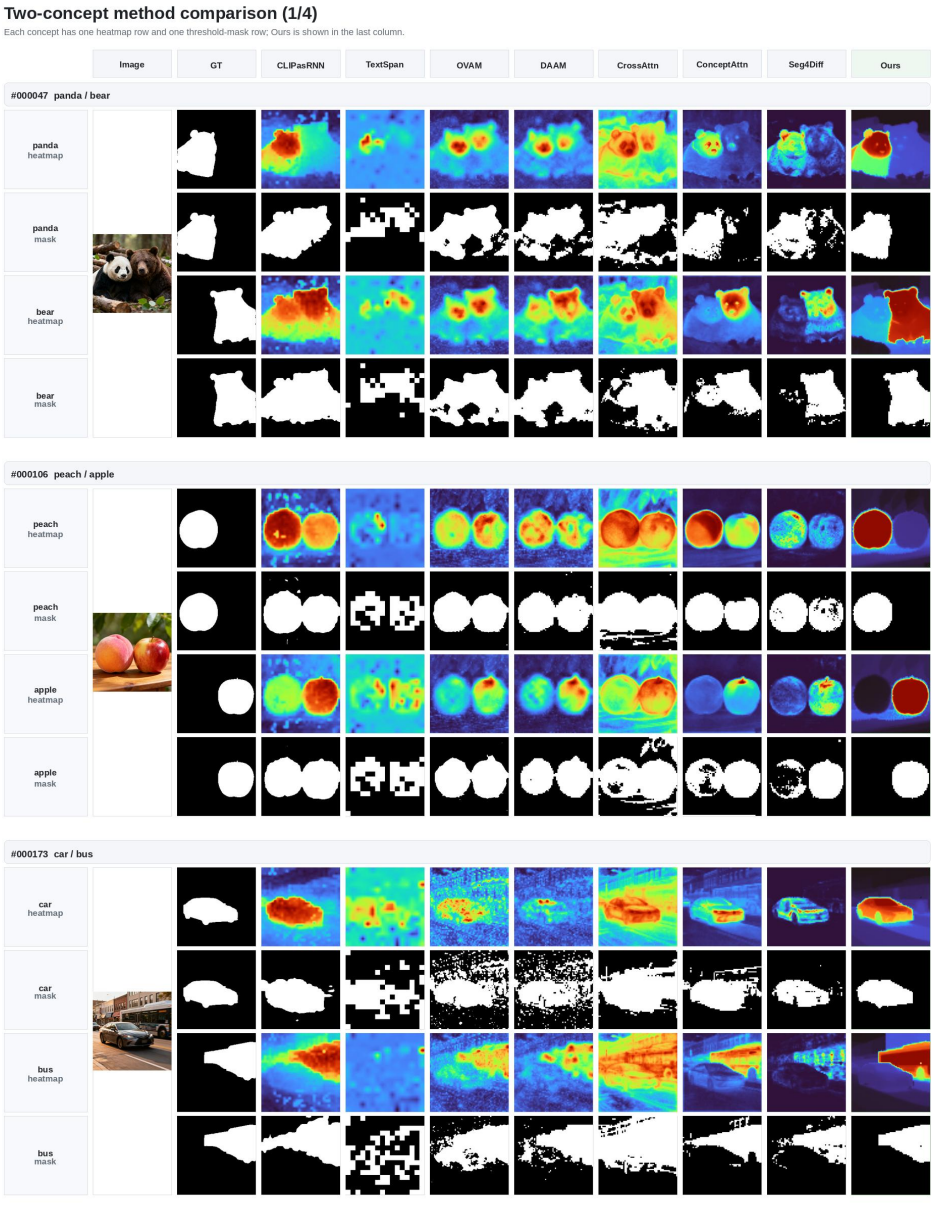

Figure 11: Additional qualitative comparisons on two-concept samples from the Multi-Concept Confusion Dataset. Page 1 of 4.

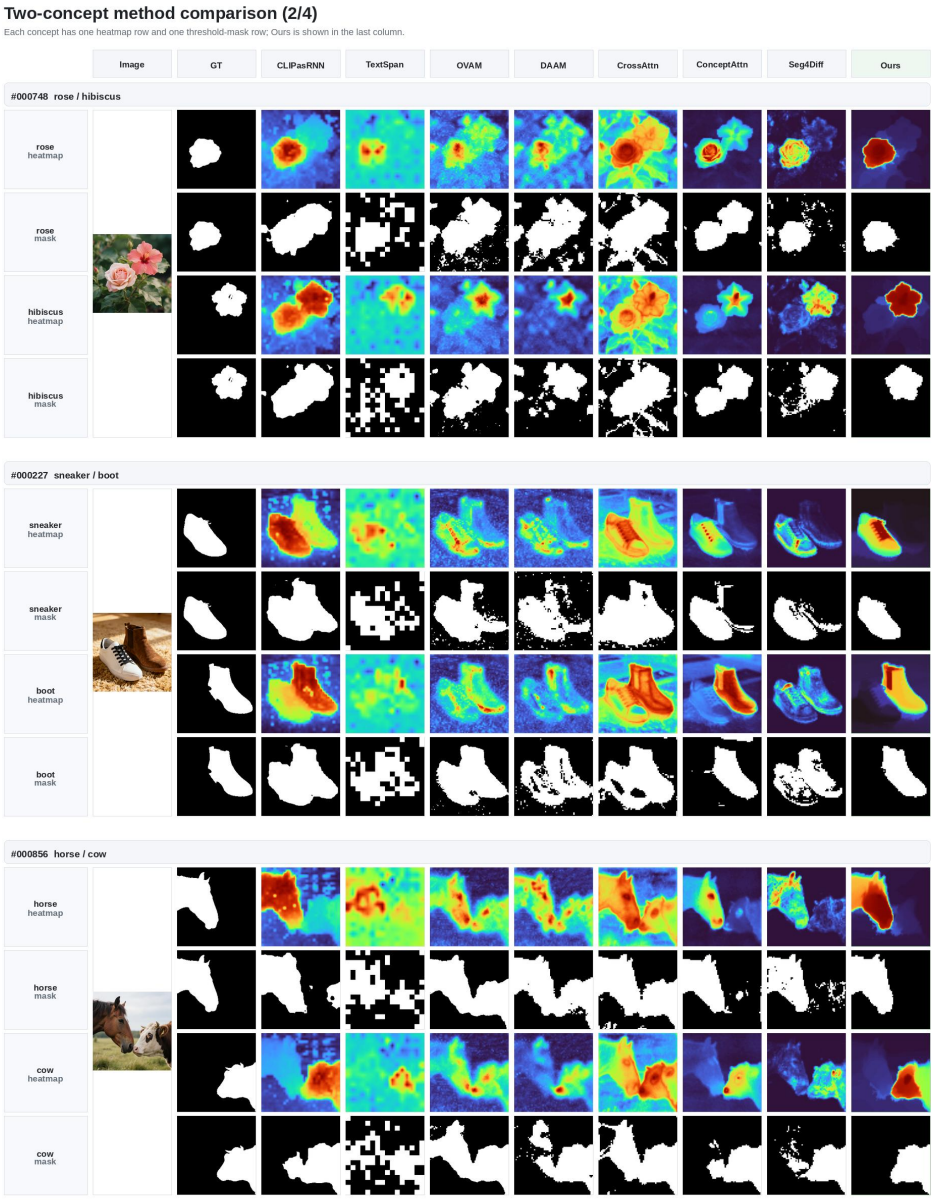

Figure 12: Additional qualitative comparisons on two-concept samples from the Multi-Concept Confusion Dataset. Page 2 of 4.

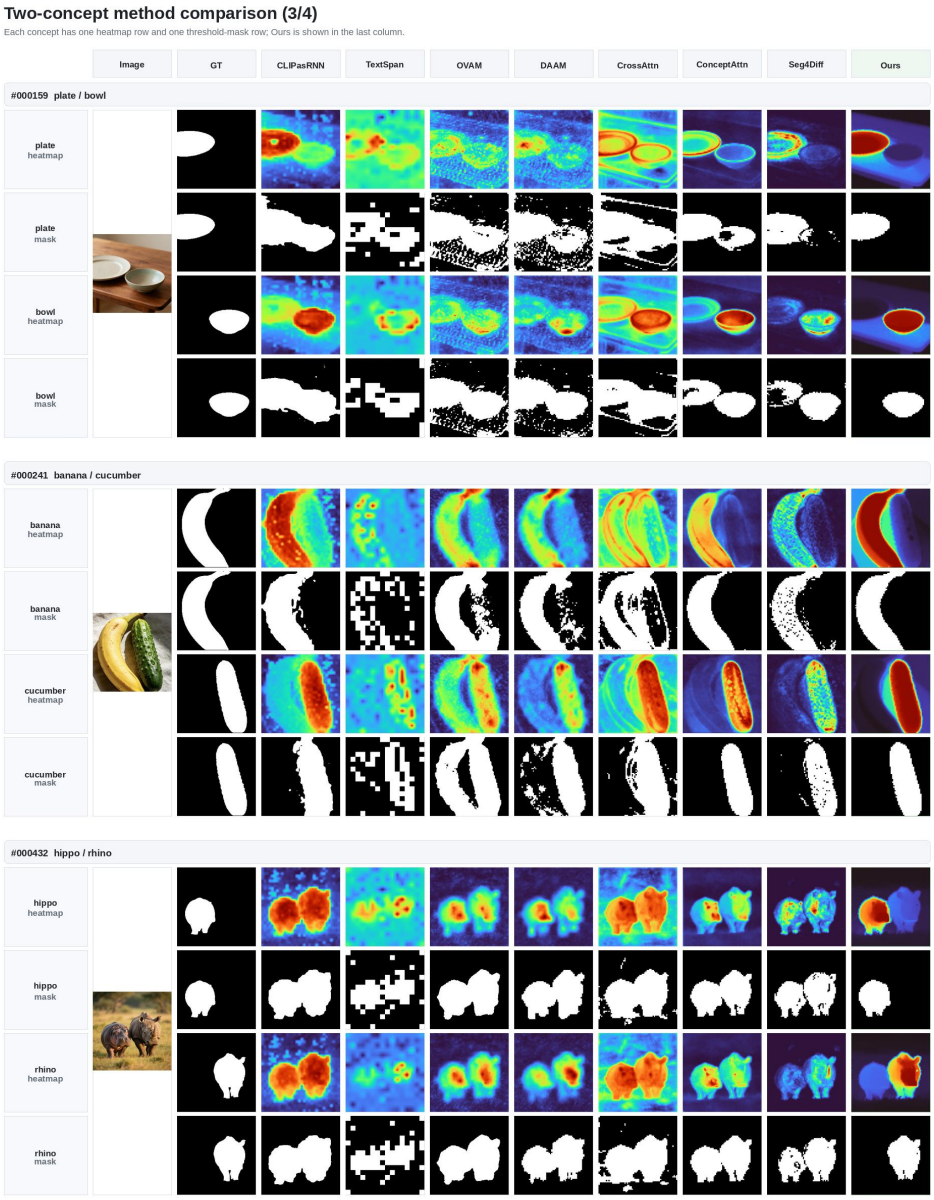

Figure 13: Additional qualitative comparisons on two-concept samples from the Multi-Concept Confusion Dataset. Page 3 of 4.

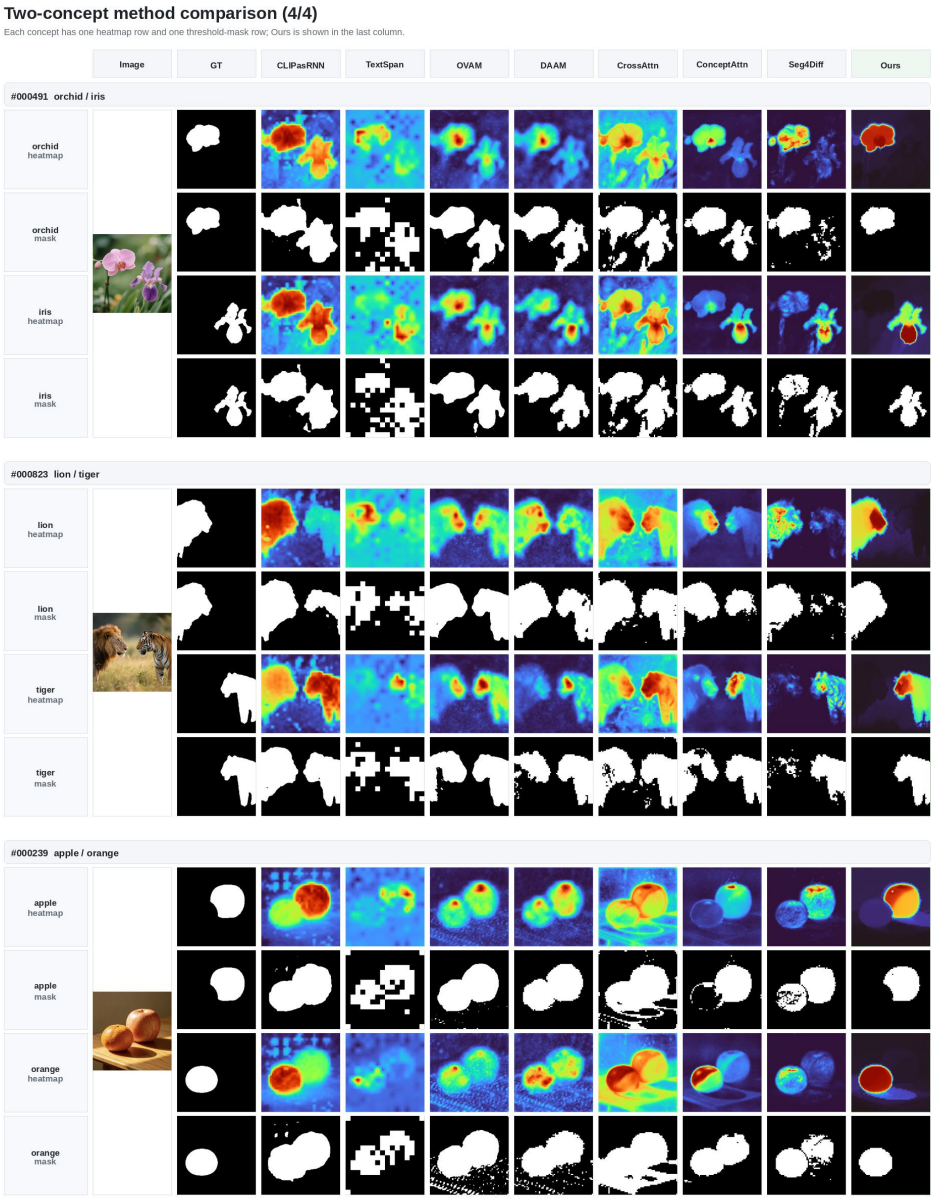

Figure 14: Additional qualitative comparisons on two-concept samples from the Multi-Concept Confusion Dataset. Page 4 of 4.

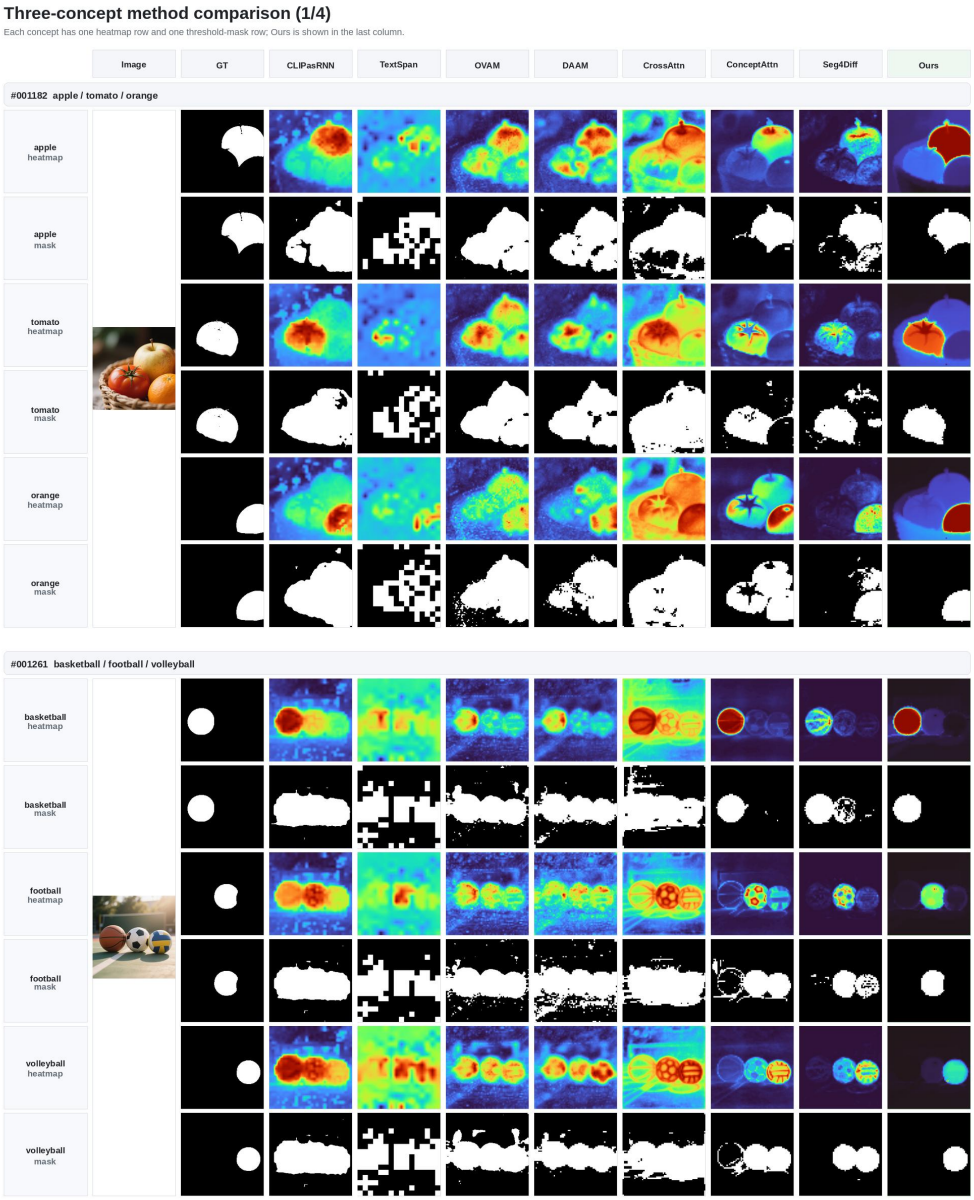

Figure 15: Additional qualitative comparisons on three-concept samples from the Multi-Concept Confusion Dataset. Page 1 of 4.

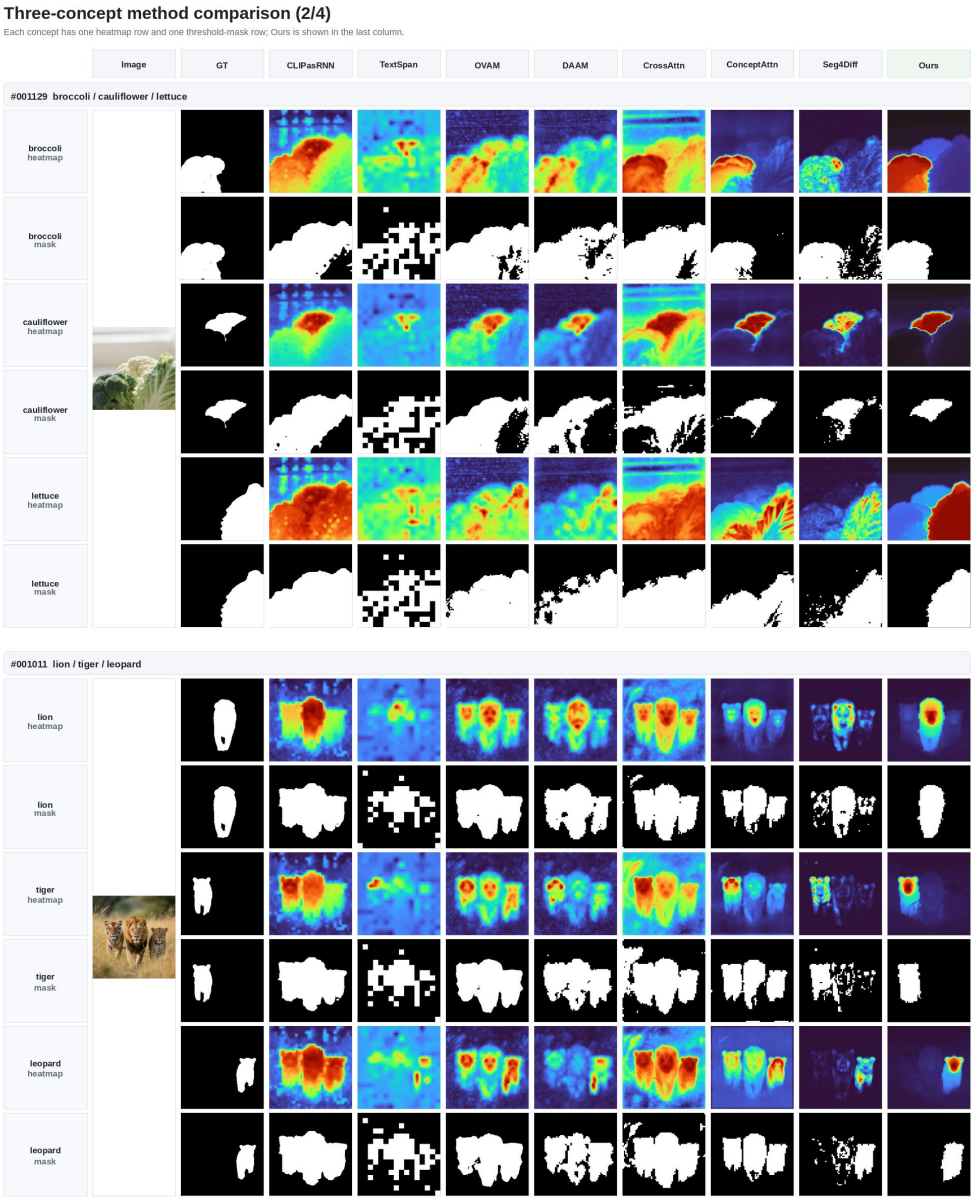

Figure 16: Additional qualitative comparisons on three-concept samples from the Multi-Concept Confusion Dataset. Page 2 of 4.

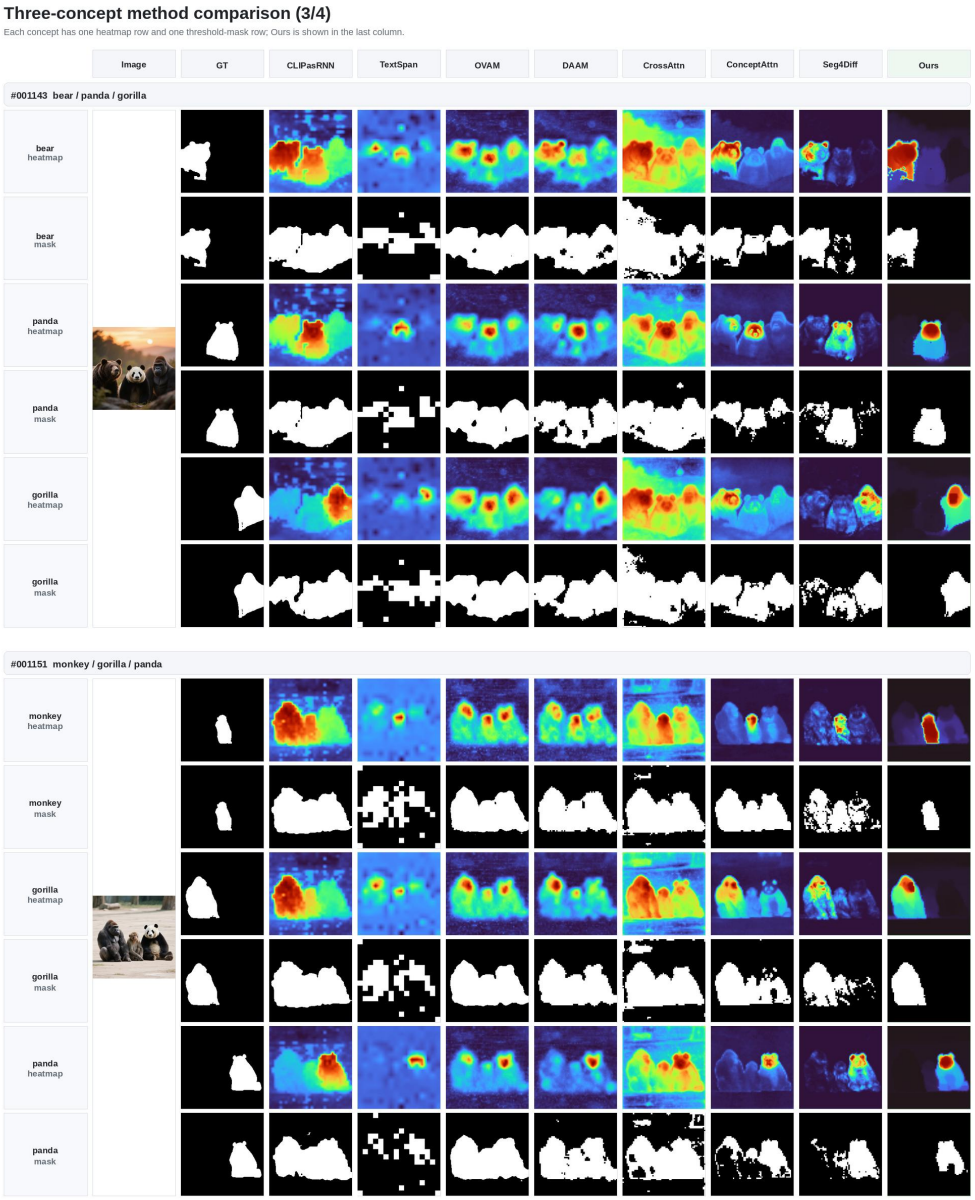

Figure 17: Additional qualitative comparisons on three-concept samples from the Multi-Concept Confusion Dataset. Page 3 of 4.

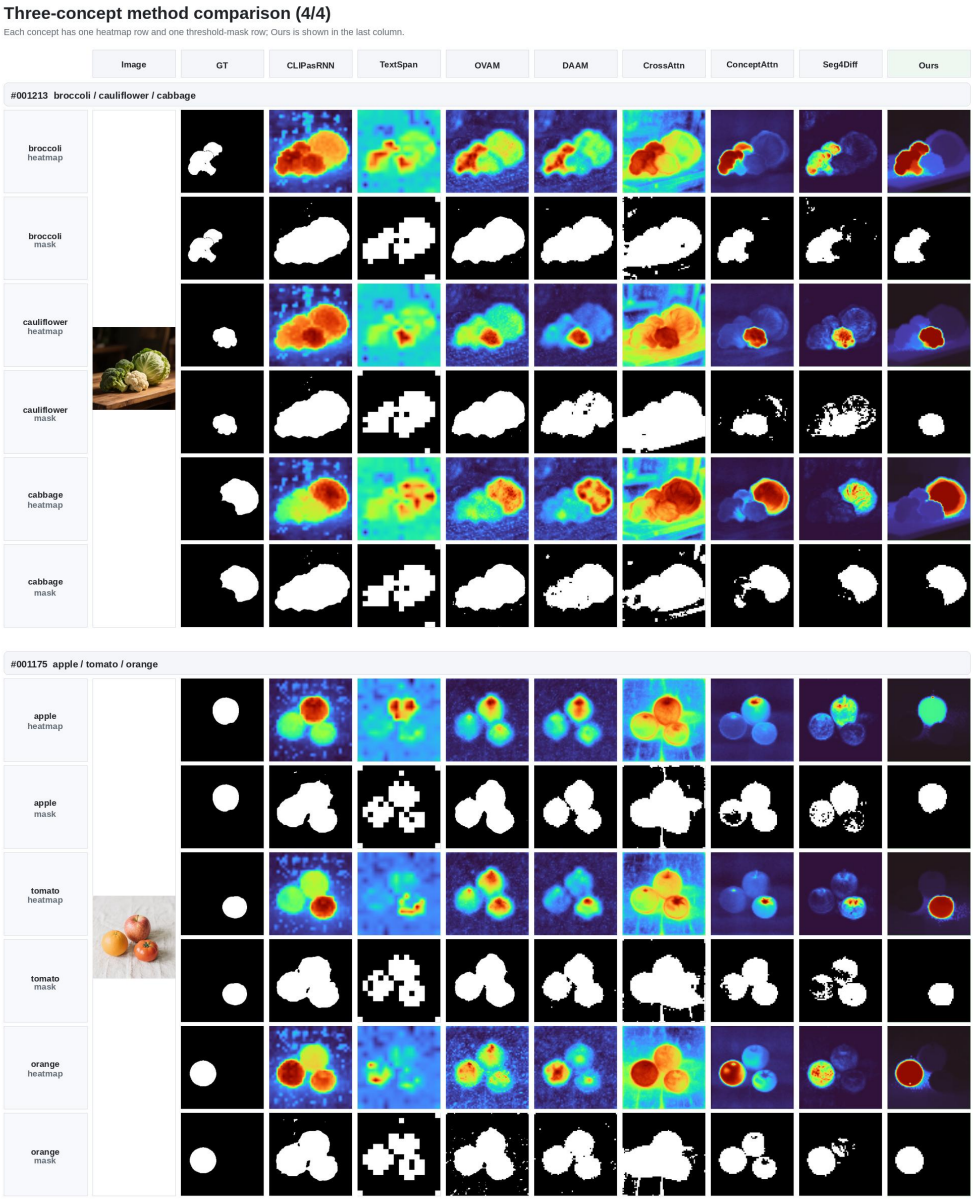

Figure 18: Additional qualitative comparisons on three-concept samples from the Multi-Concept Confusion Dataset. Page 4 of 4.

A.3 ADDITIONAL LAYER SELECTION ANALYSIS

Single-layer sensitivity. Tables 8 and 9 report numerical results when using a single image-to-image attention layer for graph propagation. Different layers show distinct trade-offs between target-object coverage and leakage suppression. For SD3, L9 is the strongest single layer, achieving the best mIoU_{fg} and Acc_{fg} while also producing the lowest NAR. Deeper layers such as L18 provide complementary behavior, with strong mAP_{fg} but weaker foreground accuracy, motivating the use of layer fusion.

For SD3.5-Large, strong single-layer performance appears in later layers, especially L20, L23–L25, L31, and L37. L25 achieves the highest single-layer mIoU_{fg}, while L37 gives the lowest NAR, indicating stronger suppression of non-target activation but not necessarily the best target coverage. This confirms that optimizing a single metric is insufficient for layer selection. The selected two-layer configurations in the main experiments therefore aim to combine complementary structural cues across depths, improving object coverage while controlling concept leakage.

Table 7: Layer-pair selection ablation on the Multi-Concept Confusion Dataset. For each backbone, we evaluate all $\binom{8}{2} = 28$ two-layer fusion pairs formed from the top-8 single-layer candidates identified by the layer-wise sensitivity analysis. This table supports the layer choices used in the main experiments by comparing different two-layer structural propagation configurations. Rows are sorted by mIoU_{fg} within each backbone. Higher is better for mIoU_{fg}, mAP_{fg}, and Acc_{fg}, while lower is better for NAR. The best value of each metric within each backbone is shown in bold.

SD3.5					**SD3**				
Layers	$\text{mIoU}_{\text{fg}} \uparrow$	$\text{mAP}_{\text{fg}} \uparrow$	NAR $\downarrow$	$\text{Acc}_{\text{fg}} \uparrow$	Layers	$\text{mIoU}_{\text{fg}} \uparrow$	$\text{mAP}_{\text{fg}} \uparrow$	NAR $\downarrow$	$\text{Acc}_{\text{fg}} \uparrow$
L23+L31	**68.89**	**89.38**	12.20	90.02	L9+L18	**64.13**	87.27	16.86	86.24
L10+L23	67.25	87.18	12.07	87.73	L9+L10	62.96	82.99	15.71	**88.13**
L24+L31	67.05	87.99	12.07	88.56	L9+L14	62.37	87.17	18.26	85.97
L25+L31	66.48	87.87	11.89	87.46	L5+L9	61.61	85.92	18.47	85.06
L20+L23	66.16	84.31	11.55	89.32	L10+L18	61.26	88.77	20.93	83.46
L10+L24	65.71	86.34	11.66	86.66	L9+L11	60.95	82.32	18.34	85.36
L20+L31	65.31	87.01	14.70	85.36	L10+L13	59.35	84.17	17.58	83.35
L23+L25	65.21	81.57	10.26	89.90	L9+L13	59.05	83.30	**15.00**	83.68
L23+L24	64.90	79.71	10.42	**90.90**	L13+L18	57.38	81.03	21.35	79.29
L20+L24	64.45	83.07	11.47	87.90	L9+L15	57.15	83.38	16.86	81.70
L10+L31	63.49	86.76	16.06	82.71	L10+L14	57.06	**89.28**	23.31	81.45
L10+L20	62.83	85.92	14.07	83.71	L5+L10	56.99	87.89	23.43	80.56
L10+L25	62.83	86.39	11.75	84.80	L5+L18	56.93	84.58	26.32	77.16
L24+L25	62.69	80.52	10.27	88.07	L14+L18	56.91	82.98	27.22	76.74
L20+L25	62.49	83.55	11.34	86.35	L13+L14	56.85	80.52	22.82	78.75
L23+L37	62.01	86.95	9.44	82.85	L10+L11	56.66	84.73	23.36	80.94
L23+L28	58.75	86.10	9.98	82.79	L11+L18	56.31	82.12	26.84	76.62
L24+L37	58.66	86.38	**9.33**	81.20	L5+L13	56.18	81.96	22.01	78.74
L31+L37	58.40	86.91	11.11	80.60	L10+L15	56.10	83.70	20.51	79.88
L20+L37	56.60	86.01	10.40	80.02	L11+L13	55.21	78.81	22.55	78.51
L28+L31	56.33	86.19	12.28	80.24	L15+L18	53.24	78.51	25.11	74.98
L24+L28	55.83	85.23	9.93	81.28	L5+L14	53.08	84.52	29.20	74.32
L25+L37	54.91	86.16	9.39	79.40	L5+L11	52.91	82.42	28.92	74.44
L20+L28	53.50	84.81	11.50	79.41	L14+L15	52.19	76.91	27.16	73.78
L25+L28	52.67	85.19	9.95	79.71	L11+L14	52.17	80.88	29.95	73.45
L10+L37	49.30	85.24	11.27	76.48	L5+L15	51.71	79.05	26.24	73.81
L10+L28	49.05	85.03	12.03	76.86	L11+L15	51.21	75.71	26.80	73.86
L28+L37	39.29	84.76	9.58	72.23	L13+L15	48.91	76.96	22.25	74.77

Layer combination selection. Table 7 reports the full two-layer sweep over the top-8 single-layer candidates ranked by mIoU_{fg} for each backbone. For SD3, L9+L18 achieves the highest foreground mIoU among all evaluated two-layer pairs, confirming the configuration used in the main experiments. For SD3.5-Large, L23+L31 achieves the best mIoU_{fg} and mAP_{fg}, supporting our use of this pair as the default SD3.5 configuration. The results also show that minimizing NAR alone is insufficient: several pairs with lower NAR under-cover the target object and obtain lower mIoU_{fg} or Acc_{fg}.

We further evaluate higher-order layer fusions by randomly sampling 20 three-layer and 20 four-layer combinations from the same top-performing single layers, with results reported in Table 10.

For SD3, a few multi-layer variants slightly exceed L9+L18 (64.13 mIoU_{fg}): the best three-layer configuration L9+L14+L18 reaches 66.64 and the best four-layer L5+L9+L10+L18 reaches 66.80. For SD3.5-Large, however, none of the sampled multi-layer combinations outperform L23+L31 (68.89), with the best three-layer reaching only 66.74 and the best four-layer reaching 61.20. The improvements are inconsistent across backbones and do not justify the additional complexity, so we adopt the simpler two-layer configuration in the main experiments. Overall, these results show that the best layer combinations are not obtained by selecting layers based on a single metric alone, nor by simply adding more layers. Instead, effective fusion requires complementary structural cues across depths that balance target-object coverage and leakage suppression.

Table 8: Single-layer sensitivity on the Multi-Concept Confusion Dataset for SD3 Medium. We evaluate graph propagation using one image-to-image attention layer at a time. Higher is better for mIoU_{fg}, mAP_{fg}, and Acc_{fg}, while lower is better for NAR. The layers used in the main two-layer configuration are highlighted in bold.

Layer	mIoU_{fg} ↑	mAP_{fg} ↑	NAR ↓	Acc_{fg} ↑	Layer	mIoU_{fg} ↑	mAP_{fg} ↑	NAR ↓	Acc_{fg} ↑
L0	10.31	15.47	53.67	49.96	L12	20.08	65.68	40.31	63.31
L1	14.97	16.36	39.49	78.53	L13	50.05	77.72	20.33	76.39
L2	28.54	78.69	34.96	61.58	L14	34.33	80.61	30.44	65.73
L3	12.47	20.94	54.21	47.23	L15	42.92	72.23	26.87	68.90
L4	20.21	67.53	46.83	53.09	L16	33.25	69.88	38.02	58.11
L5	36.06	81.33	30.14	64.22	L17	17.86	57.55	48.44	51.85
L6	28.12	83.26	33.83	63.94	**L18**	42.28	82.06	25.24	70.92
L7	21.48	75.63	42.85	55.17	L19	18.69	67.29	43.06	58.76
L8	27.72	85.78	31.42	67.44	L20	22.38	75.75	39.58	58.57
L9	**55.32**	80.47	**13.78**	**84.32**	L21	19.94	72.98	42.70	56.60
L10	41.97	85.62	19.80	76.76	L22	16.24	50.78	45.86	59.38
L11	39.26	74.75	30.62	66.00	L23	16.70	49.67	45.41	58.70

Table 9: Single-layer sensitivity on the Multi-Concept Confusion Dataset for SD3.5 Large. We evaluate graph propagation using one image-to-image attention layer at a time. Higher is better for mIoU_{fg}, mAP_{fg}, and Acc_{fg}, while lower is better for NAR. The layers used in the main two-layer configuration are highlighted in bold.

Layer	mIoU_{fg} ↑	mAP_{fg} ↑	NAR ↓	Acc_{fg} ↑	Layer	mIoU_{fg} ↑	mAP_{fg} ↑	NAR ↓	Acc_{fg} ↑
L0	14.29	15.54	42.86	74.92	L19	17.75	73.81	42.13	61.50
L1	16.13	26.58	51.06	49.96	L20	57.96	81.96	14.34	83.77
L2	25.83	60.89	40.08	59.11	L21	29.15	76.48	39.33	55.99
L3	25.84	46.35	34.00	57.35	L22	46.40	83.47	16.07	82.10
L4	13.99	20.74	51.85	49.11	**L23**	57.00	79.91	10.75	**88.82**
L5	26.43	48.22	37.09	54.74	L24	58.34	78.14	10.68	88.81
L6	35.86	63.41	33.88	61.92	L25	**61.40**	80.95	10.32	88.43
L7	38.66	77.36	26.07	70.49	L26	48.96	83.78	18.78	78.39
L8	35.95	77.65	29.06	66.70	L27	41.74	78.29	25.41	70.70
L9	15.13	22.31	53.55	49.30	L28	56.28	82.97	10.73	81.51
L10	53.71	81.78	17.63	77.84	L29	41.27	85.12	25.64	68.76
L11	24.37	75.83	39.12	59.51	L30	27.81	79.25	36.95	58.71
L12	19.20	63.30	47.24	53.60	**L31**	53.44	86.36	17.59	79.19
L13	51.25	85.10	16.87	79.56	L32	22.06	66.74	45.21	52.37
L14	24.56	76.68	41.37	55.45	L33	21.87	76.37	40.46	58.40
L15	42.61	**88.58**	20.43	76.07	L34	27.07	78.49	38.40	57.13
L16	52.25	85.34	16.29	79.74	L35	33.88	64.38	12.35	68.31
L17	38.11	84.14	28.65	65.71	L36	16.49	50.80	45.36	60.18
L18	26.41	86.42	32.59	65.54	L37	60.96	83.74	**9.40**	82.49

A.4 Visualization of the effect of $W_{\cos}$

We further provide qualitative ablations to illustrate the role of the output-space affinity $W_{\cos}$ in the proposed graph propagation process. As shown in Figures 19 to 21, we compare the propagation trajectories with and without $W_{\cos}$ at different propagation steps. These examples contain target

Table 10: Three-layer and four-layer combination ablation on the Multi-Concept Confusion Dataset. We randomly sample 20 three-layer and 20 four-layer combinations from the top-performing single layers identified in the layer-wise sensitivity analysis. Rows are sorted by $\mathrm{mIoU_{fg}}$ within each subsection.

	SD3.5					SD3			
Layers	$\mathrm{mIoU_{fg}}\uparrow$	$\mathrm{mAP_{fg}}\uparrow$	NAR↓	$\mathrm{Acc_{fg}}\uparrow$	Layers	$\mathrm{mIoU_{fg}}\uparrow$	$\mathrm{mAP_{fg}}\uparrow$	NAR↓	$\mathrm{Acc_{fg}}\uparrow$
Three-layer combinations									
L23+L24+L31	66.74	87.93	11.24	88.09	L9+L14+L18	66.64	88.16	19.79	85.88
L10+L23+L31	66.20	88.83	12.80	85.19	L5+L9+L18	65.66	87.74	19.61	85.46
L10+L20+L23	63.36	87.17	12.07	84.40	L5+L9+L10	64.82	86.45	18.51	87.27
L10+L23+L25	62.00	86.88	10.98	84.30	L9+L11+L18	64.14	86.49	19.69	85.42
L20+L25+L31	61.98	87.28	12.19	84.01	L10+L11+L14	61.31	87.54	25.03	82.66
L10+L20+L31	61.52	87.51	14.40	82.24	L10+L13+L18	61.31	85.70	19.22	82.90
L10+L24+L25	59.29	86.21	10.86	82.89	L5+L10+L13	60.85	85.64	20.08	83.00
L10+L20+L25	58.48	86.50	11.89	82.12	L9+L13+L18	60.02	85.33	16.96	82.70
L23+L31+L37	57.59	88.48	10.48	80.21	L5+L9+L13	59.77	85.13	17.57	83.01
L23+L24+L28	54.59	85.91	9.93	80.56	L5+L13+L14	59.22	83.08	23.75	79.98
L24+L28+L31	54.02	87.33	10.98	79.34	L5+L13+L18	59.02	83.24	22.48	80.05
L23+L25+L37	53.33	87.20	9.51	78.60	L10+L15+L18	58.99	85.14	21.48	80.52
L20+L24+L37	52.45	86.86	10.05	78.05	L5+L9+L15	58.32	85.05	19.42	81.38
L25+L31+L37	52.02	87.76	10.39	77.67	L5+L10+L15	58.31	85.13	22.58	80.31
L20+L25+L37	49.97	86.75	10.08	76.98	L11+L13+L14	57.98	80.88	24.24	79.46
L24+L25+L28	49.63	85.31	9.91	78.20	L5+L11+L13	57.54	81.94	23.50	79.69
L10+L20+L37	46.13	86.52	11.36	75.07	L10+L11+L15	57.13	83.45	22.67	80.23
L10+L25+L37	45.46	86.70	10.41	74.77	L11+L14+L15	54.78	78.37	27.41	75.86
L10+L25+L28	45.26	86.34	10.84	75.22	L9+L13+L15	52.22	82.78	17.19	78.88
L20+L28+L37	37.58	85.92	10.18	71.44	L13+L15+L18	51.18	79.57	22.43	76.00
Four-layer combinations									
L23+L24+L25+L31	61.20	86.89	10.81	84.45	L5+L9+L10+L18	66.80	88.15	19.25	86.46
L10+L20+L23+L24	59.03	86.83	11.26	82.21	L10+L11+L14+L18	65.73	87.79	24.58	83.86
L10+L20+L23+L25	56.80	86.84	11.26	81.24	L9+L10+L11+L18	65.13	87.15	19.24	86.40
L20+L23+L28+L31	52.02	87.70	11.43	78.24	L5+L9+L11+L18	64.89	87.09	21.30	84.67
L20+L23+L24+L37	51.00	87.41	10.02	77.36	L5+L9+L10+L11	64.41	86.06	20.19	86.49
L23+L25+L28+L31	50.36	87.66	10.62	77.66	L5+L11+L14+L18	63.31	84.63	27.53	80.44
L20+L24+L25+L28	46.47	85.53	10.41	76.42	L5+L10+L13+L18	61.44	86.32	20.83	82.38
L10+L20+L23+L37	45.18	87.54	10.80	74.65	L5+L9+L13+L14	60.72	86.07	19.48	82.65
L10+L24+L31+L37	45.05	87.91	11.04	74.35	L5+L9+L10+L13	60.15	86.00	17.54	83.34
L10+L25+L28+L31	44.58	87.49	11.55	74.58	L9+L10+L14+L15	59.98	86.23	18.96	82.69
L10+L23+L25+L37	44.42	87.50	10.19	74.29	L5+L13+L14+L18	59.92	83.75	23.81	80.17
L10+L20+L24+L37	43.64	87.20	10.67	73.95	L9+L11+L13+L14	59.39	84.75	19.65	82.42
L10+L25+L31+L37	43.50	87.75	11.10	73.70	L5+L11+L13+L14	59.03	83.00	24.71	80.00
L10+L24+L25+L28	43.43	86.54	10.44	74.39	L11+L13+L14+L18	58.97	82.25	24.18	79.75
L10+L20+L25+L28	42.89	86.53	11.17	74.07	L5+L14+L15+L18	57.51	82.24	26.14	77.75
L23+L24+L28+L37	37.61	87.07	9.50	71.44	L5+L11+L15+L18	56.51	81.84	25.85	77.55
L24+L25+L28+L37	35.31	86.63	9.49	70.53	L10+L13+L14+L15	53.81	83.71	21.07	78.65
L10+L28+L31+L37	33.98	86.79	11.27	69.57	L10+L11+L13+L15	52.38	82.78	20.91	78.30
L10+L24+L28+L37	33.21	86.68	10.11	69.35	L5+L13+L14+L15	52.01	81.20	23.84	76.46
L10+L25+L28+L37	32.30	86.56	10.19	69.01	L5+L13+L15+L18	52.00	81.54	22.93	76.53

objects and visually related non-target regions, making them suitable for analyzing both within-object propagation and cross-object leakage.

As shown in Figure 20, with $W_{\cos}$, the heat response spreads more smoothly over the entire target object as the propagation step increases. This is because $W_{\cos}$ is computed from the self-attention output features and provides dense long-range affinity between patches belonging to the same object. Consequently, patches on different parts of the target object remain strongly connected, allowing the one-hot anchor response to propagate from the initial high-confidence seed to the full object region. The resulting heatmaps are more spatially uniform inside the target object, and the corresponding binary masks recover the object shape more consistently. In contrast, removing $W_{\cos}$ weakens the dense within-object connectivity. Although the row-wise attention gate can still constrain propagation to structurally related regions, the propagation becomes less effective at transferring activation across distant parts of the same object.

As reported in Table 1, $W_{\cos}$ assigns high affinity to same-object patches, but it can also connect visually confusable different objects. In our full model, the attention gate suppresses such cross-object and foreground-background connections, thereby enlarging the relative similarity gap between the

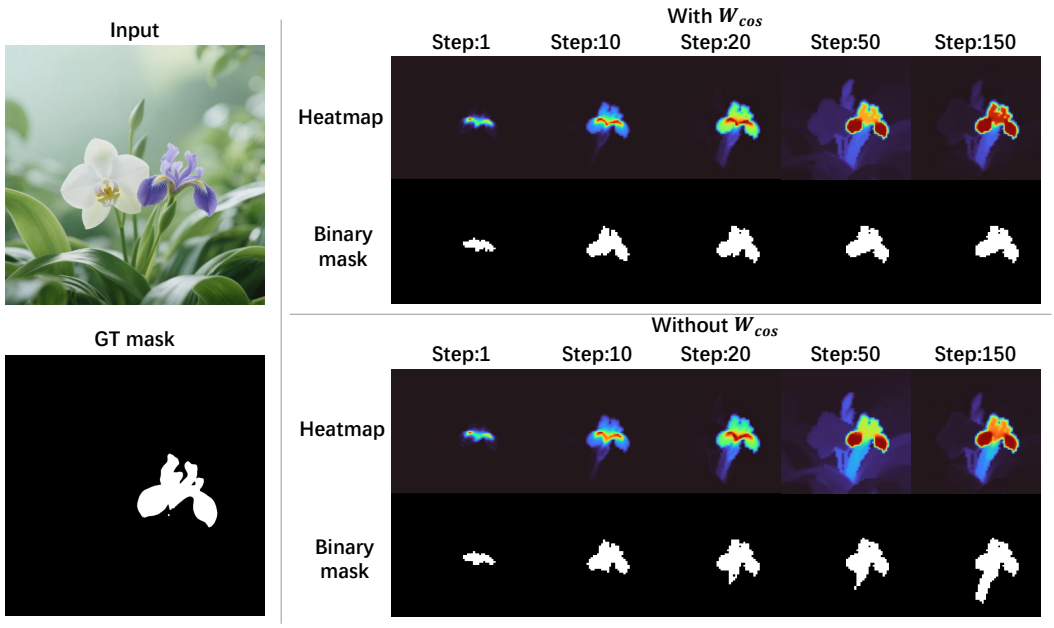

Figure 19: Visualization of graph propagation with and without $W_{\cos}$. $W_{\cos}$ improves within-object propagation, while the attention gate helps suppress leakage to nearby non-target regions. Example 1 of 3.

target object and other objects. This effect is also reflected in the visual comparisons in Figure 19 and Figure 21. With $W_{\cos}$, the target-object heatmap is brighter and more coherent inside the object, while the surrounding non-target regions remain darker. Without $W_{\cos}$, the response around the target is less well confined and is more prone to leakage. These visualizations support the quantitative ablation results and show that $W_{\cos}$ mainly improves object-level response completeness, while the attention gate prevents this dense propagation from spreading to visually similar non-target objects.

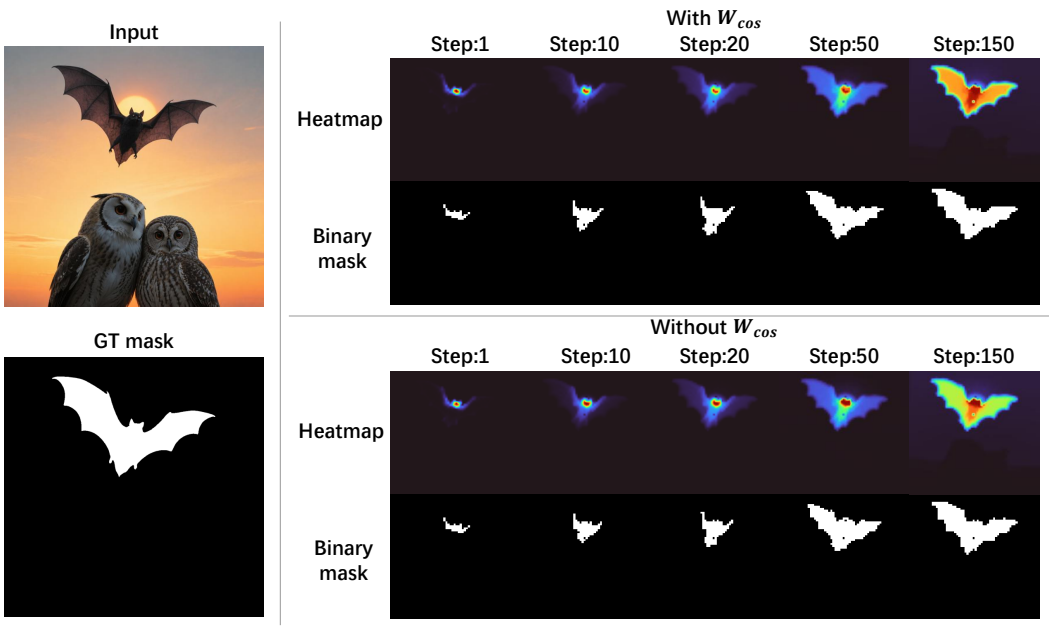

Figure 20: Visualization of graph propagation with and without $W_{\cos}$. $W_{\cos}$ improves within-object propagation, while the attention gate helps suppress leakage to nearby non-target regions. Example 2 of 3.

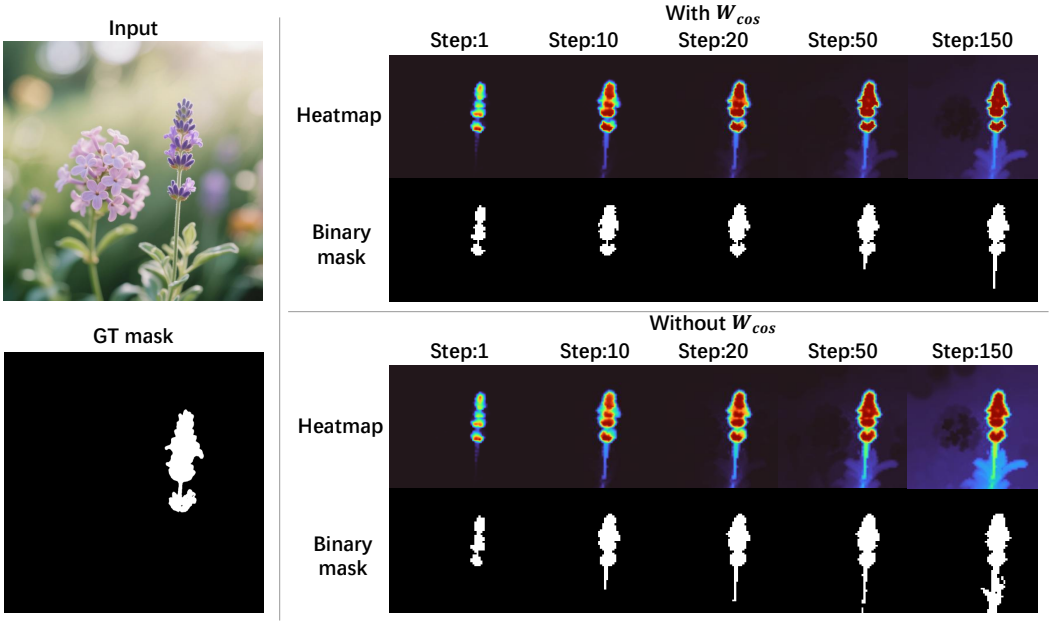

Figure 21: Visualization of graph propagation with and without $W_{\cos}$. $W_{\cos}$ improves within-object propagation, while the attention gate helps suppress leakage to nearby non-target regions. Example 3 of 3.

A.5 MULTI-CLASS SEGMENTATION

To complement the single-class benchmarks and the Multi-Concept Confusion Dataset, we evaluate all methods on the 518 PascalVOC validation images containing multiple foreground classes. This setting assesses the ability to separate different object classes within the same image. Acc is computed over the full image, while mIoU is averaged over all foreground classes and the background class.

Table 11: Multi-class segmentation on 518 PascalVOC validation images containing multiple foreground classes. All values are percentages. Higher is better; bold indicates the best result among the compared methods.

Method	Acc ↑	mIoU ↑
TextSpan	56.44	35.80
DAAM	50.54	34.20
SD3.5 Cross-Attention	62.60	37.51
ConceptAttention	78.73	39.55
Seg4Diff	77.78	46.46
AnchorDiff (ours)	**79.46**	**46.78**

As shown in Table 11, AnchorDiff achieves the highest Acc and mIoU among the compared methods. It improves over Seg4Diff by 1.68 Acc points and 0.32 mIoU points, and over ConceptAttention by 0.73 Acc points and 7.23 mIoU points. These results support the effectiveness of AnchorDiff when multiple foreground classes must be separated within the same image, extending the evidence beyond single-class benchmarks and the dedicated concept-confusion setting.

A.6 TRANSFERRING ROW-WISE STRUCTURAL INFORMATION TO EXISTING HEATMAPS

To isolate the usefulness of row-wise attention structure, we apply one structural-consistency reweighting based on W_{attn} to the original ConceptAttention and Seg4Diff heatmaps. Their heatmap-generation procedures are unchanged. This control does not introduce semantic anchor selection, $W_{\cos}$, or iterative graph propagation. The notation "+W_{attn}" denotes this single reweighting step, rather than the complete AnchorDiff pipeline.

Table 12: Transferring row-wise structural information on the Multi-Concept Confusion Dataset (1,272 images). Bold indicates the better result within each baseline/reweighted pair.

Method	mIoU_{fg} ↑	mAP_{fg} ↑	NAR ↓	Acc_{fg} ↑
ConceptAttention	39.97	88.73	26.36	59.28
ConceptAttention+ W_{attn}	**58.27**	**89.67**	**13.19**	**78.17**
Seg4Diff	46.57	84.22	19.08	69.19
Seg4Diff+ W_{attn}	**63.82**	**84.74**	**9.28**	**85.57**

Table 13: The same structural transfer control on ImageNet-Segmentation (4,276 images). Bold indicates the better result within each baseline/reweighted pair.

Method	Pixel Acc ↑	mIoU ↑	mAP ↑
ConceptAttention	83.74	72.03	**88.48**
ConceptAttention+ W_{attn}	**83.93**	**72.31**	87.29
Seg4Diff	**79.52**	**66.00**	83.28
Seg4Diff+ W_{attn}	77.62	63.42	**83.77**

On the Multi-Concept Confusion Dataset, structural reweighting improves ConceptAttention by 18.12 mIoU_{fg} points and reduces NAR by 12.90 points; the corresponding changes for Seg4Diff are +17.37 and -9.92 points (Table 12). This supports the usefulness of row-wise attention relations for separating visually confusable foreground objects, even when applied outside AnchorDiff.

The benefit is not uniform across datasets or metrics. On ImageNet-Segmentation (Table 13), ConceptAttention improves by 0.28 mIoU points, but its mAP decreases from 88.48 to 87.29. Seg4Diff loses 2.58 mIoU points and 1.90 Pixel Acc points, despite a small mAP increase. A single structural reweighting is therefore not a universally beneficial replacement for the original heatmaps. Both reweighted baselines remain below AnchorDiff's 77.29 mIoU on this benchmark.

A.7 PROPAGATION GRAPH ABLATIONS AND CONVERGENCE

We compare five propagation graphs using SD3.5. Within each dataset in the graph comparison, the backbone, selected layers, timestep, semantic anchor, propagation length, layer fusion, and mask

threshold are held fixed; only the propagation graph is changed. The propagation budget is 160 steps. The raw attention matrix $A_{I\to I}$ is distinct from the structural gate G_{attn}: the former contains direct attention weights, whereas the latter retains token pairs according to similarity between their complete attention rows.

Table 14: Propagation graph comparison across three datasets. The reported metric is mIoU for ImageNet-Segmentation and PascalVOC Single-Class, and mIoU_{fg} for the Multi-Concept Confusion Dataset. Higher is better; bold indicates the best graph within each dataset.

Propagation graph	ImageNet-Seg mIoU	PascalVOC mIoU	Multi-Concept mIoU_{fg}
Raw $A_{I\to I}$	67.90	67.18	19.73
$W_{\cos}$	41.23	43.88	11.89
G_{attn}	77.25	81.34	66.71
$A_{I\to I} \odot G_{\text{attn}}$	75.56	79.67	59.07
$W_{\cos} \odot G_{\text{attn}}$ (full)	**77.29**	**81.53**	**68.89**

Table 15: Propagation-length sweep on the Multi-Concept Confusion Dataset, reporting mIoU_{fg}. Bold marks the best result over steps within each graph column.

Steps	G_{attn}	$W_{\cos} \odot G_{\text{attn}}$
0	1.42	1.42
1	23.96	32.50
2	27.65	36.38
5	30.81	43.82
10	34.58	48.75
20	39.27	51.54
30	43.57	54.89
40	47.59	58.96
60	51.72	63.11
80	55.31	65.73
100	58.79	67.15
120	61.72	68.06
140	64.37	68.63
160	66.71	68.89
180	67.75	**69.14**
200	67.83	69.02
240	**67.87**	68.99
320	66.19	68.94
400	65.15	67.97

Table 16: Affinity-retention sensitivity across datasets. Multi-Concept results use foreground-only mIoU, mAP, and Acc, as in the main paper; the standard datasets use their standard metrics. NAR is reported only for Multi-Concept. Bold metric values indicate the best result within each dataset; bold 2% marks the default retained ratio.

Dataset	ρ	mIoU ↑	mAP ↑	NAR ↓	Acc ↑
Multi-Concept Confusion	1%	60.70	85.71	**9.38**	84.53
	2%	**68.89**	89.38	12.20	**90.02**
	3%	68.53	**90.66**	15.23	89.42
	5%	66.62	90.47	21.56	87.19
ImageNet-Segmentation	1%	73.35	88.29	–	82.88
	2%	77.29	91.20	–	85.46
	3%	**77.72**	**91.29**	–	**85.71**
	5%	75.55	89.99	–	84.33
PascalVOC Single-Class	1%	78.60	91.87	–	86.69
	2%	**81.53**	**94.05**	–	**88.50**
	3%	81.53	94.03	–	88.42
	5%	80.28	93.30	–	87.75

Table 14 identifies the structural gate as the main contributor to the strong performance. Using G_{attn} alone obtains 77.25, 81.34, and 66.71 on the three datasets, respectively. The full graph improves these scores by 0.04, 0.20, and 2.18 points.

Raw attention performs substantially worse, especially on Multi-Concept Confusion (19.73), while $W_{\cos}$ alone also performs poorly (11.89). Gating raw attention improves the results, but $A_{I\to I} \odot G_{\text{attn}}$ remains below the full graph by 1.73, 1.87, and 9.82 points. These comparisons distinguish edge selection from edge weighting: the gate selects structurally compatible token pairs, while output-space affinity weights the retained connections. They are consistent with the locality and affinity analyses in the main paper, but do not imply that spatial locality is intrinsically undesirable for segmentation.

Propagation-length comparison. We additionally compare the gate-only graph and the full graph over 0–400 propagation steps on the Multi-Concept Confusion Dataset (Table 15). The two graphs have the same initial score (1.42), and the full graph is better at every evaluated nonzero step count. At 120 steps it reaches 68.06, already exceeding the gate-only maximum of 67.87 at 240 steps. Complementing these quantitative results, Figures 19–21 suggest that cosine weighting helps maintain spatially concentrated target responses during propagation, with reduced over-expansion of the binarized masks at larger iteration counts. Overall, these observations suggest that $W_{\cos}$ reweights the connections retained by the structural gate to support effective target coverage in fewer propagation steps while helping keep the propagated responses concentrated within the target region.

A.8 Structural retention and target-scale sensitivity

Retained affinity. We vary the retained ratio $\rho \in \{1\%, 2\%, 3\%, 5\%\}$ while keeping all other settings fixed. Here $\rho = 2\%$ means retaining entries above the 98th percentile of W_{attn}. Table 16 consolidates the Multi-Concept sweep and the corresponding standard-benchmark sweeps.

On Multi-Concept Confusion Dataset, retaining 1% minimizes NAR (9.38) but reduces mIoU_{fg} to 60.70. Increasing retention to 2% improves mIoU_{fg} to 68.89 and Acc_{fg} to 90.02. Denser graphs further increase NAR, reaching 21.56 at 5%; the highest mAP_{fg} instead occurs at 3%. Thus, 2% offers a useful coverage–leakage trade-off. The same default gives the highest reported mIoU, mAP, and Acc on PascalVOC. On ImageNet-Segmentation, 3% performs best, improving mIoU by 0.43 points over 2%.

Target area and propagation length. We partition the 2,842 annotated targets in the Multi-Concept Confusion Dataset by ground-truth area relative to the image. Table 17 reports mIoU_{fg} at different propagation lengths. The middle bin includes targets occupying at least 10% but less than 25% of the image.

Table 17: Propagation-step sensitivity stratified by relative target area on Multi-Concept Confusion. n is the number of annotated targets in each bin; numbered columns are propagation steps and entries are mIoU_{fg}. Bold indicates the best evaluated step count within each row.

Target / image area	n	40	80	120	160	240	320	400
$< 10\%$	1,970	61.83	65.85	**66.45**	65.96	65.04	64.72	63.45
$[10\%, 25\%)$	800	53.64	66.59	72.44	75.91	77.88	**78.15**	77.75
$\geq 25\%$	72	39.65	52.79	63.60	71.20	78.15	81.98	**83.84**

Small targets peak at 120 steps (66.45), after which additional propagation lowers their score to 63.45 at 400 steps. Medium targets benefit up to 320 steps (78.15), and large targets continue improving through the largest tested budget of 400 steps (83.84). A fixed 160-step budget is therefore a compromise across target scales, rather than the optimum for every group.

A.9 Inference efficiency and shared multi-concept computation

Each evaluated image uses one standard MM-DiT forward pass and one concept-query forward pass. All queried categories in an image are processed together in the concept-query pass, and the affinity graph is constructed once and shared across concepts. The subsequent propagation operates on the stored graph and does not invoke an additional backbone forward pass.

Under this matched setting, with 160 propagation steps, AnchorDiff takes 1.0983 seconds per image on average, compared with 1.0690 seconds for ConceptAttention. This adds 0.0293 seconds per image, corresponding to approximately 2.7% overhead relative to ConceptAttention. The additional computation is the graph propagation used for mask refinement. For a fixed graph and fixed number of concepts, this work grows approximately linearly with the number of propagation steps, while the two backbone passes dominate the measured total latency.

A.10 Limitation: coverage of same-category instances

We examine a limitation of the single-anchor formulation: spatially separate instances of the same category may not be fully covered by propagation from one selected anchor. We evaluate this behavior on 623 cases constructed from the PascalVOC validation split. Each case contains at least two instances of the queried category, with 3.05 instances on average and a maximum of 19. Other foreground categories may also be present. An instance is counted as recalled when the predicted mask covers at least 50% of its ground-truth pixels. Instance Recall measures the percentage of queried-category instances satisfying this criterion, while Mask Area measures the percentage of image pixels predicted as foreground.

Table 18: Analysis of same-category instance coverage on 623 PascalVOC cases. All values are percentages. Higher is better for mIoU, mAP, Acc, and Instance Recall; Mask Area has no preferred direction. Bold indicates the best result for each performance metric.

Method	mIoU ↑	mAP ↑	Acc ↑	Instance Recall ↑	Mask Area
AnchorDiff (ours)	62.42	92.25	83.34	67.13	25.25
Cross Attention	59.37	**92.34**	80.46	84.37	30.66
ConceptAttention	61.32	91.79	82.23	80.52	27.20
Seg4Diff	**63.23**	92.21	**84.86**	74.81	20.80
DAAM	44.78	79.95	65.70	**88.15**	44.08
TextSpan	46.97	85.88	70.14	87.18	35.64

As shown in Table 18, AnchorDiff achieves the second-highest mIoU (62.42), behind Seg4Diff (63.23), but has the lowest Instance Recall (67.13) among the compared methods. This reveals a coverage limitation: relatively high mask overlap does not imply that all instances of the queried category are recovered. Compared with Cross Attention, AnchorDiff improves mIoU from 59.37 to 62.42 and reduces predicted mask area from 30.66% to 25.25%, but recalls fewer instances. Conversely, DAAM attains the highest Instance Recall (88.15), together with the largest predicted area (44.08%) and the lowest mIoU (44.78). These results show why instance coverage and overall mask quality should be assessed separately. The single-anchor formulation does not explicitly stop propagation after reaching one instance, and retained graph connections may still allow it to reach multiple instances. However, spatially separate instances may remain disconnected from the selected anchor, leading to incomplete coverage. Therefore, introducing additional anchors is a possible direction for improving same-category instance coverage.