PROBING AND STEERING VISION MODEL DECISIONS ALONG CONCEPT AXES IN DIFFUSION SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce Mask-supervised Orthonormal Concept Axes (MOCA), a diffusion-based framework for concept-wise probing and steering of vision model decisions. Through independent manipulation of individual visual concepts in a given image and observing the resulting model response, MOCA enables concept-wise probing underlying a vision model's decisions. By directly scaling concept strengths, MOCA formulates visual explanation as concept-wise interventions, probing feature sensitivity to reveal whether specific visual elements explain the decision or act as spurious correlations. Experiments across multiple datasets and vision models demonstrate that MOCA produces spatially localized and semantically meaningful interventions, enables identification of spurious cues, and supports downstream mitigation. We extend MOCA for visual counterfactual explanation, providing a broader range of explanations for the same prediction through interventions on different visual concepts within the image. Experiment results show that MOCA improves the overall accuracy in the mitigation task by 44.9% on HardImageNet and 3.0% on COCO relative to the corresponding baselines, achieves comparable performance in counterfactual explanation tasks, and consistently support additional explanations through non-ground-truth concepts.

1 INTRODUCTION

Vision models have proliferated but assessing their reliability and robustness remains challenging. Understanding how vision models arrive at their decisions is critical for such an assessment. For example, when classifying an object x as category y, we would like to know whether the model predicts y mainly because it has learned statistical associations between x-related patterns and y during training, or because it identifies defining visual characteristics of x, such as its shape, texture, and structure, and uses them as evidence for the prediction.

Prior work has approached the question of why a model makes a particular prediction from multiple perspectives. Saliency-based methods trace a decision backward through the model to highlight contributing image regions, providing a spatial explanation of where predictive cues reside (Zhou et al., 2016; Selvaraju et al., 2017; Jalwana et al., 2021). A complementary line of work instead probes the decision itself, examining how changes to the input move the prediction toward or across the model's decision boundary (Moayeri et al., 2022; Goyal et al., 2019). Adversarial-based approaches introduce small perturbations to the input and measure the resulting change in prediction (Ilyas et al., 2019; Kang et al., 2023; Li et al., 2024), whereas more recent visual counterfactual explanation (VCEs) methods leverage generative models to to synthesize realistic and semantically meaningful edits that flip predictions (Jeanneret et al., 2022; 2023; Augustin et al., 2022; Sobieski et al., 2025; Guo et al., 2026). We identify two major constraints in existing methods: first, these methods primarily capture what causes a model to reverse the prediction, rather than how visual features intrinsically support it. Second, most methods implicitly assume that prediction shifts are driven solely by the intended target edit or perturbation. This assumption overlooks unintended modifications to background cues and co-occurring objects, which can carry spurious correlations resulting in flawed explanations.

To address these limitations, we reposition this task as concept-level visual explanation: for any identified visual concept (target or background cue), we perform isolated interventions that operate exclusively on that concept while keeping all other visual elements unchanged. As illustrated in

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.

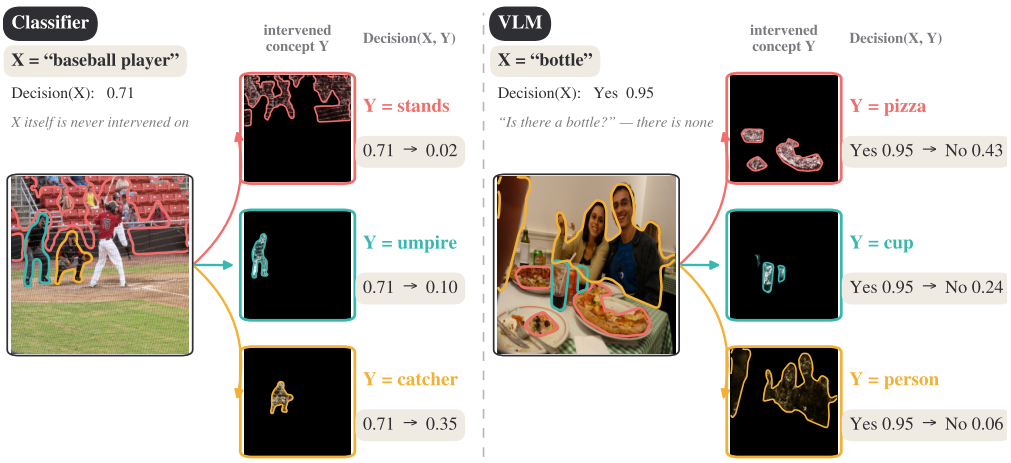

Figure 1: **Concept level explanations with MOCA.** Given a model decision on a concept of interest X, MOCA enables separate interventions on individual visual concepts in the image and measure the resulting change in the model decision. Here, we illustrate interventions on different Y while keeping X unchanged, providing multiple concept-wise explanations for the same prediction. We illustrate MOCA in both classifier (left) and VLM (right) settings.

Fig. 1, we directly answer: *"How does a model's prediction about X change when we intervene on different visual concepts (including X itself and other identifiable concepts Y) while keeping all other context intact?"* With the explanation, we can reveal not only the model's sensitivity to individual concepts, but also whether a concept acts as a true decision driver or a spurious shortcut.

We formulate visual explanation as concept-wise intervention within a diffusion framework. MOCA introduces image-specific concept handles that allow localized manipulation while minimizing unintended changes to other visual content. These handles can be directly scaled to suppress or amplify individual concepts, and can be adapted for downstream tasks such as sensitivity probing, model steering, and visual counterfactual explanation. Our main contributions and findings are as follows:

- We present a novel diffusion framework, MOCA, which learns image-specific, mask-supervised orthonormal concept axes for concept-wise probing and localized concept-wise manipulations.
- We introduce scaling-based interventions that provide image explanations from multiple concept-wise perspectives by varying concept strength rather than changing the underlying concepts. The resulting concept-wise sensitivity enables identification of spurious cues and support downstream mitigation, improving accuracy from 57.87% to 83.88% on HardImageNet and from 89.50% to 92.20% on COCO;
- We extend MOCA to steer model decisions, and demonstrate its applicability to model stress testing and visual counterfactual explanation. MOCA steering reaches an 89.87% top-1 flip rate on HardImageNet and 75.86%/63.57% directional flip rates on COCO; For VCE, MOCA provides explanations through both ground-truth and non-ground-truth objects, achieving average flip rates of 64.7% and 38.3%.

2 Related Work

Spurious Correlations and Mitigation in Vision Models. Model bias can arise when vision models rely on unintended visual cues rather than task-relevant evidence, such as the reliance on spurious correlations, where features that are correlated with class labels in the training data influence model predictions despite not being semantically essential to the target concept. Prior work has studied how to identify and expose such dependencies. Neuhaus et al. (2023) introduce a scalable framework for identifying spurious features, while Hard ImageNet (Moayeri et al., 2022) reveals model reliance on suspected cues by ablating image regions and measuring changes in classifier predictions. LANCE (Prabhu et al., 2023) uses diffusion-based counterfactual editing to modify potentially spurious visual attributes, and Fatima et al. (2025) study how object size and location affect reliance on spurious backgrounds. Mitigation methods aim to reduce such reliance once spurious

cues are identified. Group DRO (Sagawa et al., 2020) minimizes worst-group loss using group annotations, while Kirichenko et al. (2023) retrain the last layer on group-balanced data. For multimodal models, Yang et al. (2023) propose a CLIP-specific fine-tuning strategy, RAVL (Varma et al., 2024) suppresses reliance on discovered spurious regions through region-aware fine-tuning, and Alshami et al. (2025) use self-supervised masking to encourage reliance on genuine visual features.

Explainable Artificial Intelligence (XAI) in the vision domain, can be broadly divided into intrinsically interpretable models and post-hoc methods. Intrinsically interpretable models are designed to incorporate interpretable structures (Bohle et al., 2021; Espinosa Zarlenga et al., 2022) or representations (Huang & Li, 2020; Böhle et al., 2022; Zhu et al., 2025) into the model itself. In contrast, post-hoc XAI approaches explain the behavior of pretrained models by identifying global or local input features that contribute to model predictions. Saliency-based explanation (Zhou et al., 2016; Selvaraju et al., 2017; Jalwana et al., 2021) leverage target-class gradients to weight intermediate feature maps and localize image regions that contribute strongly to a prediction. Recent work has moved beyond pixel- or region-level attribution toward higher-level semantic explanations, including concept attribution (Kim et al., 2018; Wang et al., 2024) and visual counterfactual explanations.

Visual Counterfactual Explanation with Diffusion Models has emerged as a research direction. DiME (Jeanneret et al., 2022) applied diffusion models to generate counterfactual explanations for analyzing spurious correlations. ACE (Jeanneret et al., 2023) adopts adversarial attack optimization to generate counterfactual explanations, using diffusion models to suppress OOD perturbations and mask-based inpainting to preserve image content outside the edited region. DVCE (Augustin et al., 2022) employs cone regularization using an adversarially robust classifier to avoid trivial, non-semantic perturbations and generate realistic counterfactual explanations. To improve spatial controllability, recent studies explore how to generate localized counterfactual changes while preserving the semantic. RCSB (Sobieski et al., 2025) restricts edits to a predefined image region and allows users to specify where counterfactual changes occur. MaskDiME (Guo et al., 2026) introduces adaptive localized sampling for spatially precise counterfactual generation, achieving faster inference than DiME. In contrast, latent counterfactual explanation methods leverage latent diffusion and text-to-image models to improve semantic controllability and computational efficiency: LDCE (Schrodi et al., 2024) introduces consensus guidance to suppress noisy or adversarial classifier gradients. TIME (Jeanneret et al., 2024) learns context- and class-specific textual embeddings to preserve image structure while steering generation toward a target class.

3 METHOD

We present MOCA, a diffusion-based framework for probing and explaining how individual visual concepts in a query image influence a vision model's prediction. We first describe our approach for decomposing a diffusion model's conditional guidance into concept-associated components in Sec. 3.1. These signals are optimized into mask-supervised orthonormal concept axes and used to construct concept-wise interventions in Sec. 3.2, and leveraged for model steering in Sec. 3.3.

3.1 CONCEPT-WISE GUIDANCE DECOMPOSITION IN DIFFUSION SPACE

Our goal is to establish a concept-level intervention for controlled manipulation of individual visual concepts and measurement of their effects on model predictions, which we refer to as *"concept probing"*. To realize such interventions, we build on a conditional diffusion model, which provides a natural way to introduce controlled semantic changes while retaining the overall content of the query image. We reconstruct the query image through diffusion inversion, and then decompose the conditional guidance into concept-specific orthonormal directions that serve as intervention axes.

Diffusion Inversion and Image Reconstruction. To achieve concept probing, we first obtain a diffusion trajectory by performing inversion, which maps the input image into the diffusion latent space. Specifically, we obtain the latent representation z_t at each timestep t by applying DDIM inversion (Song et al., 2020; Dhariwal & Nichol, 2021) over T steps. During the denoising stage, we start from $\hat{z}_T = z_T$ and iteratively denoise the latent. At each timestep t, the standard classifier-free guidance (CFG) noise prediction is computed as:

$$\hat{\epsilon}_\theta(\hat{z}_t, t, C) = \epsilon_\theta(\hat{z}_t, t, \varnothing) + w \underbrace{(\epsilon_\theta(\hat{z}_t, t, C) - \epsilon_\theta(\hat{z}_t, t, \varnothing))}_{\text{CFG guidance signal } g_t}, \tag{1}$$

where w is the guidance scale, and $\epsilon_\theta(\hat{z}_t, t, C)$ and $\epsilon_\theta(\hat{z}_t, t, \varnothing)$ denote the conditional and unconditional noise predictions, respectively. Given the predicted noise $\hat{\epsilon}_\theta(\hat{z}_t, t, C)$, the latent is updated at each reverse denoising step using DDIM sampling as:

$$\hat{z}_{t-1} = \sqrt{\frac{\bar{\alpha}_{t-1}}{\bar{\alpha}_t}}\hat{z}_t + \big(\sqrt{1-\bar{\alpha}_{t-1}} - \sqrt{\frac{\bar{\alpha}_{t-1}}{\bar{\alpha}_t}}\sqrt{1-\bar{\alpha}_t}\big)\hat{\epsilon}_\theta(\hat{z}_t, t, C), \tag{2}$$

where $\bar{\alpha}_t$ denotes the cumulative product of the diffusion noise schedule. Iterating this reverse process from $\hat{z}_T = z_T$ yields the reconstructed latent $\hat{z}_0$, which is decoded to obtain the reconstructed image $x^{(0)}$.

Concept-Wise Guidance Decomposition. With the reconstructed diffusion trajectory established, we next decompose the conditional guidance into concept-associated components for separate manipulation. For each concept c in the text condition C, we estimate its initial contribution as:

$$d_{t,c} = \epsilon_\theta(\hat{z}_t, t, C) - \epsilon_\theta(\hat{z}_t, t, C \setminus c) \tag{3}$$

We then define the residual as the remaining guidance not captured by these concept-associated directions, i.e., $R_t = g_t - \sum_{c\in\mathcal{C}} d_{t,c}$. Accordingly, the guidance term in Eq. (1) can be written as

$$\hat{\epsilon}_\theta(\hat{z}_t, t, C) = \epsilon_\theta(\hat{z}_t, t, \varnothing) + w \underbrace{\big(\textstyle\sum_{c\in\mathcal{C}} d_{t,c} + R_t\big)}_{\text{Initial concept decomposition}} \tag{4}$$

where R_t denotes the residual guidance signal. Ideally, the decomposed concept-specific guidance signals should satisfy two properties: (1) **Concept independence:** Each concept-specific signal should be independently controllable, such that modifying one signal does not affect the contributions of the remaining signals. (2) **Spatial localization:** The spatial influence of each concept-specific signal should be concentrated within the region associated with the corresponding concept, with minimal impact on unrelated regions.

3.2 Learning Mask-Supervised Orthonormal Concept Axes

Orthonormal Concept Axes. We optimize the concept axes derived from guidance decomposition based on the visual changes induced by perturbing one concept at a time. Before intervening the optimization objective, we impose an orthonormality constraint on the decomposed concept-specific guidance signals, so that each concept-specific direction is independently controllable and contributes independently to the final intervention. First, we collect the initial directions into a direction matrix $D_t \in \mathbb{R}^{n\times N}$ at each denoising step, where each row corresponds to a concept-specific direction $d_{t,c}$. We then seek an orthonormal basis $Q_t \in \mathbb{R}^{n\times N}$ derived from D_t, where each row $q_{t,c}$ corresponds to concept c and $Q_t Q_t^\top = I_n$.

Scaling-based Intervention. We define an intervention on concept c by scaling its associated guidance component along q_c with coefficient δ. We first follow the standard reverse denoising process until reaching the beginning of the intervention window $\mathcal{T}_w$. At each subsequent $t \in \mathcal{T}_w$, we modify the noise prediction by scaling the projection of the CFG guidance signal onto the concept axis $q_{t,c}$:

$$\hat{\epsilon}_\theta^{(c,\delta)}(\hat{z}_t, t, C) = \hat{\epsilon}_\theta(\hat{z}_t, t, \varnothing) + w \underbrace{\big(\delta a_{t,c} q_{t,c} + \textstyle\sum_{c'\neq c} a_{t,c'} q_{t,c'} + R_t\big)}_{\text{Orthogonal CFG decomposition}}. \tag{5}$$

$a_{t,c} = \langle g_t, q_{t,c}\rangle$ is the projection coefficient of g_t onto the concept axis $q_{t,c}$ and $R_t = g_t - \sum_{c\in\mathcal{C}} a_{t,c} q_{t,c}$ is the residual guidance. For all remaining concepts $c' \neq c$ the guidance components are kept unchanged. After applying the modified noise prediction throughout the intervention window, we continue the reverse denoising process to obtain the final latent, which is then decoded into the intervened image $\hat{x}_{c,\delta}$. We define the induced visual change relative to the unmodified reconstruction $x^{(0)}$ as $\Delta\hat{x}_{c,\delta} = \hat{x}_{c,\delta} - x^{(0)}$. Varying δ leads to enhance or suppress the visual evidence associated with concept c while minimizing changes to other visual region.

Spatial and Semantic Constraints on Concept Axes. The orthogonality constraint alone does not guarantee spatial localization: a direction $q_{t,c}$ may still induce changes outside the region corresponding to concept c due to the spatially diffuse nature of the generative model. To address this,

we further optimize the orthonormal axes using concept-specific masks M_c with localization loss:

$$\mathcal{L}_{\text{loc}}^{(c)} = \frac{\|(\mathbf{1} - M_c) \odot \hat{x}_{c,\delta}\|_1}{\|\hat{x}_{c,\delta}\|_1 + \epsilon}. \tag{6}$$

We use concept-specific masks M_c to supervise the image changes induced by each direction, penalizing changes outside the corresponding concept region. To encourage that the intervention remains semantically meaningful rather than degenerating into arbitrary or noise-like perturbations, we further introduce a CLIP-based semantic loss to regularize the edited image.

$$\mathcal{L}_{\text{sem}}^{(c)} = 1 - \cos\big(E_I(\hat{x}_{c,\delta} \odot M_c + x^{(0)} \odot (\mathbf{1} - M_c)), E_T(c)\big). \tag{7}$$

where $E_I(\cdot)$ and $E_T(\cdot)$ denote the CLIP image and text encoders, respectively. While the orthogonality constraint and localization objective encourage independently controllable and spatially localized axes, the optimization may cause the learned axes to drift away from their initial concept-associated directions. To preserve this correspondence, we additionally introduce a alignment loss:

$$\mathcal{L}_{\text{align}}^{(c)} = 1 - \frac{1}{|\mathcal{T}_w|} \sum_{t \in \mathcal{T}_w} \langle q_c(t),\, \hat{d}_c(t) \rangle. \tag{8}$$

$\mathcal{L}_{\text{align}}^{(c)}$ encourages each optimized axis to remain close to its corresponding initial concept direction while remaining separated from the axes associated with other concepts.

Jointly, we formulate our final objective and optimize the concept axes as:

$$\underset{Q}{\text{minimize}} \ \sum_{c \in \mathcal{C}} \lambda_{\text{loc}} \mathcal{L}_{\text{loc}}^{(c)} + \lambda_{\text{sem}} \mathcal{L}_{\text{sem}}^{(c)} + \mathcal{L}_{\text{align}}^{(c)}, \ \text{ s.t. } \quad Q(t)Q(t)^\top = I_n, \quad \forall\, t \in \mathcal{T}_w. \tag{9}$$

where λ_{loc} and λ_{sem} are coefficients for localization and semantic losses, respectively.we generate a concept-specific intervention for each concept using its corresponding axis. This allows us to examine how an intervention induced by each individual concept affects the model's prediction.

3.3 Steering Model Decisions with Learnable Directions

We extend MOCA by introducing a learnable steering direction to edit the region associated with a target concept c in a query image. Given the identified orthonormal basis Q, we search for this steering direction by adjusting the concept-specific axis q_c while maintaining orthogonality to all other concept axes. We perform editing over the timestep window $\mathcal{T}_w$ during the reverse diffusion process. We parameterize the adjustments using a learnable matrix $V \in \mathbb{R}^{|\mathcal{T}_w| \times N}$, where each row v_t corresponds to an adjustment at step t. At each step, we obtain the orthogonal adjustment and the updated concept axis as:

$$\Delta q_{t,c} = v_t - (v_t Q_{t,-c}^\top) Q_{t,-c}; \quad q'_{t,c} = \frac{q_{t,c} + \Delta q_{t,c}}{\|q_{t,c} + \Delta q_{t,c}\|_2}, \tag{10}$$

where $Q_{t,-c}$ contains all concept axes except $q_{t,c}$. $\Delta q_{t,c}$ determines how the axis with respect to c is steered during optimization. With the adjusted axis $q'_{t,c}$ corresponding to concept c, we steer the reverse diffusion to edit the concept region through DDIM step in Eq. (5). We optimize V using the task-oriented objectives (below) for sensitivity assessment and visual counterfactual explanation.

Task 1: Steering for Model Sensitivity Probing. The seek to probe the model's decision boundary and provide concept-wise explanations by testing whether localized interventions on different concepts can alter the model's decision. With the orthonormal basis fixed, we keep all non-target concept axes unchanged and optimize $\Delta q_{t,c}$ to adjust the axis associated with concept c. This produces a targeted intervention on the corresponding concept-specific guidance signal, inducing localized pixel-level changes in the associated image region to alter the prediction of the probed model.

$$\mathcal{L}_{probing} = \mathcal{L}_{loc}\big(\Delta q_{t,c}\big) - \lambda_{\text{dec}} \sigma P\big(\hat{x}(\Delta q_{t,c})\big) \tag{11}$$

where P denotes the predicted probability of the GT class for image classification or $P(\text{yes})$ for VLM probing and λ_{dec} is the weight of the decision objective. $\sigma \in \{-1, +1\}$ controls the steering direction, corresponding to suppression and amplification, respectively.

Task 2: Visual Counterfactual Explanation. Unlike sensitivity probing, which examines whether intervening on a concept can alter the current model decision, VCE aims to steer the prediction

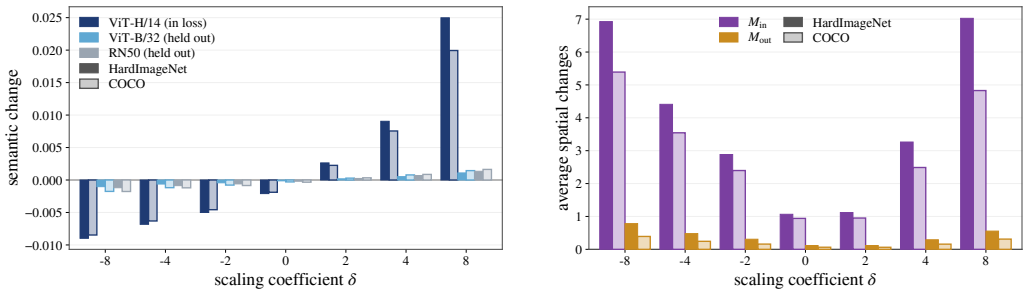

Figure 2: **Semantic response and spatial localization under scaling-based interventions.** The left panels show the mean change in CLIP cosine similarity to the target concept, evaluated using the in-loss ViT-H/14 and two held-out models (ViT-B/32 and RN50). The right panels report the mean absolute pixel-value changes inside (M_{in}) and outside (M_{out}) the target concept masks.

toward a specified target class P^{target} while preserving localized and semantically meaningful changes. Importantly, our framework allows this target prediction to be reached by steering not only the target-related concept, but also other concepts present in the image, thereby providing multiple concept-wise counterfactual explanations. Following prior work (Jeanneret et al., 2023; Guo et al., 2026), we formulate the VCE objective as:

$$\mathcal{L}_{vce} = \mathcal{L}_{loc}\big(\Delta q_{t,c}\big) - \lambda_{\text{dec}}\sigma P^{target}\big(\hat{x}(\Delta q_{t,c})\big) \tag{12}$$

4 Experiments

4.1 Evaluation of Concept Decomposition

Datasets and Implementation Details. We evaluate MOCA and its probing protocols on ResNet-50 with HardImageNet (Moayeri et al., 2022) and Qwen2.5-VL-7B-Instruct (Team, 2025) with COCO (Lin et al., 2014). We select 1,000 images from COCO val2017, which provides instance segmentation masks and captions for each image. HardImageNet consists of 15 ImageNet-1K (Deng et al., 2009) classes with strong spurious cues. For the VCE task, we follow prior work and evaluate on visually similar category pairs from ImageNet: sorrel vs. zebra, Persian cat vs. Egyptian cat, and cougar vs. cheetah. We build our framework on Stable Diffusion 2.0 Base (Rombach et al., 2022) with DDIM sampling (Song et al., 2020) and Null-Text Inversion (Mokady et al., 2023), using 50 diffusion steps for inversion and reverse denoising. The captions used to obtain the initial diffusion trajectories for HardImageNet and ImageNet are generated by Qwen2.5-VL-7B-Instruct, while for COCO we use the captions provided with the dataset. Secs. A.1 and A.2 has further details.

Evaluating the Semantic Changes and Localization of Scaling-based Interventions. We evaluate whether scaling a concept-specific direction producing semantically aligned and spatially localized changes. As shown in Fig. 2, negative coefficients ($\delta \in \{-8, -4, -2, 0\}$) decrease the similarity to the target concept, while positive coefficients ($\delta \in \{2, 4, 8\}$) increase it. As the intervention strength increases, the magnitude of semantic responses and in-mask pixel-value changes generally increases, indicating that scaling-based interventions induce direction-dependent semantic changes while maintaining spatial concentration. Moreover, our per-concept analysis shows that 90.4% and 88.2% of concept maps exhibit the expected response signs across the evaluated coefficients on HardImageNet and COCO, respectively. Detailed measurements under each δ are provided in Sec. A.3.1. Fig. 3 fur-

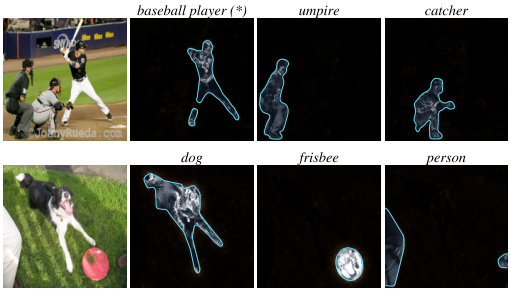

Figure 3: **Visualization of concept-specific intervention.** Blue contours indicate the boundaries of the extracted masks. Both examples illustrate that our concept axes are spatially localized and that the intervention introduced by each concept is independent. $(*)$ marks the ground-truth class for HardImageNet sample.

ther shows that different concept axes produce distinct spatial responses that are largely concentrated within their corresponding target regions.

4.2 Sensitivity Assessment through Concept-Wise Interventions

We use MOCA as a probing tool to assess the sensitivity of both unimodal and multimodal models to individual visual concepts. We first quantify how concept-wise interventions affect model decisions and analyze model sensitivity on HardimageNet. We further examine whether the identified concepts associated with spurious correlations can be used to guide mitigation. Finally, we extend the analysis to a vision-language model on COCO.

Sensitivity Assessment on HardImageNet. On HardImageNet, we report the flip rate (FR) of the top-1 prediction to quantify prediction sensitivity under input interventions, together with Fréchet Inception Distance (FID) (Heusel et al., 2017) to characterize the distributional change induced by the interventions. We compare MOCA's scaling-based interventions and steering with prior approaches used for sensitivity assessment on HardImageNet, including the Mask, BBox, and Tile interventions provided by HardImageNet and the language-guided counterfactual editing method LANCE (Prabhu et al., 2023). Tab. 1 shows that MOCA achieves high FR while inducing substantially smaller distributional changes than LANCE. We also observe that interventions on non-GT concepts lead to more decision flips than those on GT concepts, suggesting that contextual concepts can play an important role in model predictions.

Table 1: Sensitivity assessment of ResNet-50 on HardImageNet under different input interventions.

Method	FR ↑ all	FR ↑ GT	FR ↑ non-GT	FID ↓
Mask	44.27	44.27	–	63.20
BBox	45.60	<u>45.60</u>	–	84.21
Tile	<u>50.00</u>	**50.00**	–	78.87
LANCE	49.33	5.07	<u>34.80</u>	66.87
Scaling (Supp.)	28.67	8.13	20.53	17.38
Scaling (Amp.)	26.53	6.27	20.27	**16.88**
Steering	**89.87**	26.00	**63.87**	<u>16.94</u>

Spurious Correlation Discovery and Mitigation. We evaluate whether our probing tools can support spurious correlation discovery and mitigation. We select 750 images (50 images $\times$ 15 classes) from the training split of HardImageNet, taking the 50 images per class ranked highest in spurious correlation. We probe each class with scaling-based interventions, selecting for each image the non-GT concept that causes the largest drop in $P(\mathrm{GT})$ at $\delta = -8$. The most frequent concept across the class is then treated as its spurious cue. With the selected cues, we conduct mitigation experiments following the setting of Yang et al. (2023). We fine-tune the projection layer of CLIP-ResNet-50 (Radford et al., 2021) while keeping both encoders frozen,using three different existing mitigation methods: Multimodal (Yang et al., 2023), Group DRO (Sagawa et al., 2020) and Empirical Risk Mimimization (ERM). All models are evaluated with the full set of 1,000 ImageNet class names. As shown in Tab. 3 (a), the baseline performs worse when the identified spurious cue is absent (55.00%) than when it is present (62.96%). ERM and Group DRO substantially improve cue-absent accuracy to 79.39% and 77.29%, respectively, while also improving overall accuracy.

Generalization to Mutimodal Models. For VLM probing, we follow POPE (Li et al., 2023) and adapt the question template to "Is there a {category} in the image?", where {category} denotes a COCO object category. We generate questions using COCO categories and generate up to six yes-questions and no questions. Categories in yes questions correspond to objects present in the image, whereas those in no questions correspond to objects absent from the image. On COCO, we define FR as an intervention causes $P(\mathrm{yes})$ to fall below or rise above the 0.5 threshold. As shown in Tab. 2, scaling-based interventions induce moderate changes in VLM decisions, whereas steering produces substantially higher flip rates in both directions. We further show in Tab. 3 (b), using spurious cues discovered with MOCA improves overall accuracy, with gains observed for both cue-absent and cue-present batches. Further details in Sec. A.3.3.

Table 2: Sensitivity assessment of VLM on COCO under different interventions.

Method	Yes→No FR ↑	No→Yes FR ↑	FID ↓
Scaling(Supp.)	9.48	1.13	18.93
Scaling(Amp.)	11.21	1.02	19.05
Steering	**75.86**	**63.57**	**18.84**

4.3 Visual Counterfactual Explanation

We evaluate the applicability of MOCA to VCE. Since our framework operates on a DDIM-inverted latent trajectory, we compare with TIME (Jeanneret et al., 2024) as a latent-space VCE baseline.

Table 3: Mitigation results on HardImageNet and COCO. We report overall accuracy and accuracy when the identified spurious object is absent or present.

Method	HardImageNet (a)		COCO (b)	
	Avg. Acc. ↑	Obj. Absent / Present ↑	Avg. Acc. ↑	Obj. Absent / Present ↑
Baseline	0.5787	0.5500 / 0.6296	0.8950	0.9037 / 0.8406
Multimodal (Yang et al., 2023)	0.5267	0.4376 / 0.6851	0.9190	0.9304 / 0.8478
Group DRO (Sagawa et al., 2020)	0.8307	0.7729 / **0.9334**	0.9090	0.9188 / 0.8478
ERM (Yang et al., 2023)	**0.8388**	**0.7939** / 0.9186	**0.9220**	**0.9327 / 0.8551**

Table 4: Visual counterfactual explanation results on three ImageNet class pairs. The best and second-best values in each column within each block are shown in **bold** and underlined, respectively.

Method	Zebra ↔ Sorrel					Cheetah ↔ Cougar					Egyptian ↔ Persian cat				
	FR ↑	FID ↓	sFID ↓	S^3 ↑	C_{out} ↑	FR ↑	FID ↓	sFID ↓	S^3 ↑	C_{out} ↑	FR ↑	FID ↓	sFID ↓	S^3 ↑	C_{out} ↑
*Visual counterfactual explanation with the **ground-truth** object*															
ACE ℓ_1	8.0	17.33	61.05	**0.994**	-0.896	35.0	**13.86**	**61.65**	**0.991**	-0.655	45.0	**26.31**	122.89	**0.987**	-0.387
ACE ℓ_2	**44.0**	**17.31**	**60.83**	0.994	-0.694	**73.0**	14.11	61.87	0.991	-0.388	75.0	26.59	**122.73**	0.986	-0.120
TIME	3.0	48.61	84.96	0.878	-0.787	42.0	69.28	94.49	0.766	-0.096	19.0	83.85	154.32	0.839	-0.191
MaskDiME	6.0	18.29	62.79	0.991	-0.889	31.0	29.68	69.79	0.982	-0.330	62.0	43.87	128.20	0.973	0.091
MOCA (ours)	31.0	21.32	64.26	0.980	**-0.460**	71.0	32.44	64.72	0.978	**0.205**	**92.0**	60.14	130.81	0.947	**0.722**
*Visual counterfactual explanation with a **non-ground-truth** object*															
TIME	1.0	8.06	59.32	0.986	-0.959	19.0	17.03	63.81	0.933	-0.686	5.0	16.29	121.26	0.972	-0.681
MaskDiME	4.0	**4.76**	**58.31**	**0.998**	-0.949	19.0	13.33	63.81	**0.994**	-0.679	7.0	**11.47**	**118.04**	**0.994**	-0.683
MOCA (ours)	**21.0**	8.47	61.45	0.991	**-0.694**	**41.0**	**10.75**	**61.17**	0.991	**-0.398**	**53.0**	19.87	119.00	0.984	**-0.074**

We also compare with DDPM-based methods ACE ℓ_1 (Jeanneret et al., 2023), ACE ℓ_2 (Jeanneret et al., 2023), and MaskDiME (Guo et al., 2026), adapting them to our generation trajectory for comparison. Following MaskDiME, we report FR, FID, sFID as introduced in ACE, SimSiam similarity (S^3) (Chen & He, 2021) for representation-level similarity, and Counterfactual Output Transition (C_{out}) (Khorram & Fuxin, 2022) for measuring distribution shift. As shown in Tab. 4, MOCA achieves the highest FR on Egyptian cat ↔ Persian cat and the second-highest FR on the other two pairs for ground-truth object interventions, while consistently obtaining the highest C_{out}. For non-ground-truth object interventions, MOCA achieves the highest FR and C_{out} across all three pairs, demonstrating that model decisions can also be explained and steered through alternative concepts in the image. Qualitative VCE for interventions on both GT concept and non-GT concepts are provided in Figs. 4 and 5.

4.4 Ablation Study

Ablation on Steering Direction We evaluate the learned steering direction by comparing the FR under four settings: the learned direction $q_c + \Delta q$, the original concept axis q_c, a matched-size intervention along q_c, and a random steering direction.As shown in Tab. 5, the learned steering direction yields substantially higher flip rates on both HardImageNet and COCO, supporting the effectiveness of the optimized direction.

Table 5: Ablation of the learned steering direction. We report flip rate (FR) on HardImageNet and COCO.

Method	HardImageNet	COCO
q_c	24.80	1.70
q_c (matched size)	25.87	3.00
$q_c + \Delta q_{\text{rand}}$	23.33	1.30
$q_c + \Delta q$	**89.87**	**65.00**

Intervention window. We ablate the intervention window $\mathcal{T}_w$ starting from $t \in \{20, 30, 40\}$, Tab. 6 reports the flip rate (FR) for the first steering task, which measures whether the intervention alters the model's top-1 prediction, and the target success rate for VCE, together with the mean pixel change $|d|$ and the fraction of the edit falling inside the target mask. As the intervention window starts earlier, steering becomes stronger but also induces larger and less localized image changes. In particular, starting at $t = 20$ achieves the highest FR and target success rate, but results in substantially larger pixel changes and a lower in-mask fraction. In contrast, $t = 40$ produces the smallest and most localized edits, at the

Table 6: Ablation on the intervention window for MOCA. $|d|$ denotes the mean pixel change over the whole image, and In-mask denotes the fraction of the edit falling inside the target mask.

Window	FR ↑	Target ↑	$\lvert d\rvert$ ↓	In-mask ↑
20	**89.3**	**89.0**	13.19	0.677
30	75.0	73.3	5.03	0.785
40	57.3	53.0	**1.85**	**0.849**



Figure 4: **Qualitative results of visual counterfactual explanation.** MOCA effectively generates localized and semantically meaningful counterfactual explanation while preserve the overall image structure.



Figure 5: **Multi-Concept Visual Counterfactual Explanations via GT and Non-GT Concept Steering.** MOCA generates localized edits and provides explanations for a model decision from multi-concept perspectives. Green/red borders indicate successful/unsuccessful prediction flips.

cost of reduced steering effectiveness. We therefore use $t = 30$ as the default setting for VCE, as it provides a balanced trade-off between steering effectiveness and spatial localization.

5 Conclusion

We present MOCA, a diffusion-based framework that unifies concept-wise probing, intervention, and model steering. MOCA learns mask-supervised orthonormal concept axes in the diffusion guidance space, providing semantic control over multiple concepts for localized image editing. The framework enables sensitivity analysis and per-image identification of candidate spurious cues, which can further support downstream mitigation. It also steers model decisions through targeted concept interventions, providing counterfactual explanations from multiple concept perspectives.

AI use statement

In this work, we used generative AI tools for coding assistance and light language editing. We have not used generative AI tools for any other tasks. Additionally, we used generative AI tools for creating and formatting figures and tables, We have reviewed all AI-assisted work. LLM-generated

code was verified and tested for correctness by authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

We provided training details and implementation details in both Sec. 4 and Secs. A.1 and A.2. Code and other artifacts will be publicly released upon acceptance via GitHub or similar platforms.

References

Eyad Alshami, Shashank Agnihotri, Bernt Schiele, and Margret Keuper. Aim: Amending inherent interpretability via self-supervised masking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 993–1003, October 2025.

Maximilian Augustin, Valentyn Boreiko, Francesco Croce, and Matthias Hein. Diffusion visual counterfactual explanations. *Advances in Neural Information Processing Systems*, 35:364–377, 2022.

Moritz Bohle, Mario Fritz, and Bernt Schiele. Convolutional dynamic alignment networks for interpretable classifications. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10029–10038, 2021.

Moritz Böhle, Mario Fritz, and Bernt Schiele. B-cos networks: Alignment is all we need for interpretability. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10319–10328. IEEE, 2022.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15745–15753. IEEE, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept embedding models: Beyond the accuracy-explainability trade-off. *Advances in neural information processing systems*, 35:21400–21413, 2022.

Mishal Fatima, Steffen Jung, and Margret Keuper. Corner cases: How size and position of objects challenge imagenet-trained models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Yqf2BhqfyZ.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International conference on machine learning*, pp. 2376–2384. PMLR, 2019.

Changlu Guo, Anders Nymark Christensen, Anders Bjorholm Dahl, and Morten Rieger Hannemose. Maskdime: Adaptive masked diffusion for precise and efficient visual counterfactual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Zixuan Huang and Yin Li. Interpretable and accurate fine-grained recognition via region grouping. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8659–8669. IEEE, 2020.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Mohammad AAK Jalwana, Naveed Akhtar, Mohammed Bennamoun, and Ajmal Mian. Cameras: Enhanced resolution and sanity preserving class activation mapping for image saliency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16322–16331. IEEE, 2021.

Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion models for counterfactual explanations. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, December 2022.

Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Adversarial counterfactual visual explanations. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16425–16435. IEEE, 2023.

Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Text-to-image models for counterfactual explanations: a black-box approach. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4745–4755. IEEE, 2024.

Mintong Kang, Dawn Song, and Bo Li. Diffattack: Evasion attacks against diffusion-based adversarial purification. *Advances in Neural Information Processing Systems*, 36:73919–73942, 2023.

Saeed Khorram and Li Fuxin. Cycle-consistent counterfactuals by latent transformations. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10193–10202. IEEE, 2022.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *International Conference on Learning Representations*, 2023.

Jin Li, Ziqiang He, Anwei Luo, Jian-Fang Hu, Z Jane Wang, and Xiangui Kang. Advad: Exploring non-parametric diffusion for imperceptible adversarial attacks. *Advances in Neural Information Processing Systems*, 37:52323–52353, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Mazda Moayeri, Sahil Singla, and Soheil Feizi. Hard imagenet: Segmentations for objects with strong spurious cues. *Advances in Neural Information Processing Systems*, 35:10068–10077, 2022.

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

Yannic Neuhaus, Maximilian Augustin, Valentyn Boreiko, and Matthias Hein. Spurious features everywhere-large-scale detection of harmful spurious features in imagenet. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20178–20189. IEEE, 2023.

Viraj Prabhu, Sriram Yenamandra, Prithvijit Chattopadhyay, and Judy Hoffman. Lance: Stress-testing visual models by generating language-guided counterfactual images. *Advances in Neural Information Processing Systems*, 36:25165–25184, 2023.

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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

Simon Schrodi, Karim Farid, Max Argus, and Thomas Brox. Latent diffusion counterfactual explanations. In *DAGM German Conference on Pattern Recognition*, pp. 295–311. Springer, 2024.

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.

Bartlomiej Sobieski, Jakub Grzywaczewski, Bartłomiej Sadlej, Matthew Tivnan, and Przemyslaw Biecek. Rethinking visual counterfactual explanations through region constraint. In *International Conference on Learning Representations*, volume 2025, pp. 12967–13019, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020.

Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Maya Varma, Jean-Benoit Delbrouck, Zhihong Chen, Akshay Chaudhari, and Curtis Langlotz. Ravl: Discovering and mitigating spurious correlations in fine-tuned vision-language models. *Advances in Neural Information Processing Systems*, 37:82235–82264, 2024.

Hengyi Wang, Shiwei Tan, and Hao Wang. Probabilistic conceptual explainers: Trustworthy conceptual explanations for vision foundation models. In *International Conference on Machine Learning*, pp. 51502–51522. PMLR, 2024.

Yu Yang, Besmira Nushi, Hamid Palangi, and Baharan Mirzasoleiman. Mitigating spurious correlations in multi-modal models during fine-tuning. In *International Conference on Machine Learning*, 2023.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

Zhijie Zhu, Lei Fan, Maurice Pagnucco, and Yang Song. Interpretable image classification via non-parametric part prototype learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9762–9771. IEEE, 2025.

A Appendix

A.1 Supplemental Details for MOCA

MOCA is designed to provide an interface for probing each concept within an image through concept-specific interventions. To preserve reconstruction quality before intervention, we apply NTI to optimize the unconditional embeddings used during CFG along the reconstruction trajectory. **Decomposition of NTI-Adjusted Guidance.** To avoid potential concept leakage from the NTI-optimized null embedding, we decompose the NTI-adjusted CFG prediction such that the concept-specific directions are derived relative to the standard unconditional embedding $\varnothing$, while the effect of the optimized embedding $\varnothing_t$ is isolated as a reconstruction correction term. The resulting noise prediction can be equivalently written as

$$\begin{aligned}\hat{\epsilon}_\theta(\hat{z}_t, t, C) = \epsilon_\theta(\hat{z}_t, t, \varnothing_t) \\ + w\left[\epsilon_\theta(\hat{z}_t, t, C) - \epsilon_\theta(\hat{z}_t, t, \varnothing)\right] \\ + w\left[\epsilon_\theta(\hat{z}_t, t, \varnothing) - \epsilon_\theta(\hat{z}_t, t, \varnothing_t)\right].\end{aligned} \tag{13}$$

Training Details of MOCA We use Stable Diffusion 2.0 Base with a CFG scale of $w = 2.5$. We initialize the learnable matrix A_t from the concept-specific directions D_t and apply QR decomposition at each optimization iteration to obtain Q_t as described in Algorithm 1.We empirically set λ_{loc} and λ_{sem} to 1 and 5, respectively. We optimize the axes using Adam with a learning rate of 2e-3 for 40 steps. The frame Q_t is solved once per image at a fixed intervention coefficient, the same $\delta = 8$ used at inference. The mask term is the fraction of the rendered change falling outside the concept mask, which is scale-free, so the objective does not reward a particular edit magnitude. At each iteration we render the window once per supervised concept, scaling only that concept's coefficient and leaving every other coefficient and the out-of-span residual at their unedited values, so each loss term is attributable to a single axis. The per-concept gradients are accumulated and applied in one update to the shared orthonormal parametrization, so the axes are learned simultaneously rather than one after another. For steering tasks, we set $\delta = 8$ as default to induce more pronounced concept-specific changes and provide a clearer contrast between the original and intervened images. We optimize $\Delta q_{t,c}$ for 40 optimization steps using the same learning rate of 2e-3. The detailed optimization procedure for steering is in Algorithm 2, where L_{dec} denotes the task-oriented objective defined in Sec. 3.3, we set $\lambda_{dec} = 10$ for all steering experiments. For each scaling-based intervention coefficient δ, we separately optimize the concept axes.

A.2 Additional Implementation Details

Intervention Settings For scaling-based interventions, we sweep seven coefficients $\delta \in \{-8, -4, -2, 0, 2, 4, 8\}$ over three intervention windows starting at $t \in \{20, 30, 40\}$, respectively. We report the scaling and sensitivity probing results using $t = 40$, which produces more localized and controlled interventions. For the VCE task, we use $t = 30$ to allow a larger editing budget for achieving the target prediction.

COCO Subset Selection We construct our subset from the COCO `val2017` instance annotations. We retain images containing at least two annotated object categories whose masks each occupy at least 2% of the image area, and keep only the largest few objects by mask coverage when necessary. The minimum-area threshold removes objects whose spatial extent is too small to support visually meaningful scaling interventions, while requiring at least two objects ensures the presence of a non-queried concept that can be independently intervened on. This procedure yields $1{,}000$ images containing 2–6 retained objects each: 604 images contain two objects, 253 contain three, 109 contain four, and 34 contain five or six. The retained objects have a mean mask coverage of 0.133, ranging from 0.020 to 0.499. We rasterize the COCO polygon annotations into binary masks and store them using the same per-image representation as for the other datasets, allowing the downstream intervention pipeline to remain unchanged. Each selected image is additionally associated with its five human-written COCO captions.

Caption and Mask Generation We evaluate across three datasets with different levels of annotation, and the annotation each provides determines how much we must generate. COCO provides

human-annotated object categories together with instance segmentation masks, eliminating the need for external captioning or segmentation. HardImageNet and the 300-image ImageNet pair set are single-label classification benchmarks and we caption both with Qwen2.5-VL-7B-Instruct under greedy decoding, asking for one descriptive sentence and a comma-separated list in which each object is named in at most two words. For 300 image pairs, since these labels are fine-grained (*sorrel*, *Persian cat*), a captioner left unconstrained may return a coarse hypernym or a pluralized variant of the ground-truth class. We therefore provide the ground-truth label to the captioner and explicitly include it as the first concept, ensuring that the target concept is represented by construction rather than through synonym matching. For non-GT concepts, we exclude entries that do not correspond to a distinct, independently localizable region under our masking pipeline. These include parts of the target object (e.g., *eye*, *stripes*, and *fur*), which lie within the target object, and generic scene-level terms (e.g., *background* and *scene*) that do not define a distinct instance-level region. Each retained name is grounded to a binary mask with SAM, stored per image as $M_c \in 0, 1^{H \times W}$ and resized to 512×512 by nearest neighbor.

VLM Question Construction We follow the question design of POPE (Li et al., 2023) to construct binary yes/no questions for evaluating VLM responses. For each image, we generate up to six questions, consisting of at most three positive (yes) questions and three negative (no) questions. Positive questions are constructed from the usable concepts present in the image, with one question generated for each concept, up to three concepts. Negative questions query concepts that are absent from the image. For example, *Image 000000000802* contains two usable concepts, *refrigerator* and *oven*, occupying 19.3% and 9.7% of the image, respectively. We therefore construct two positive questions, such as "Is there a refrigerator in the image?", together with three negative questions, such as "Is there a bowl in the image?". Thus, images with two usable concepts contain five questions, whereas images with three or more usable concepts contain six questions.

Captioner Sensitivity We compare Gemini 2.5 Flash with Qwen and analyze the sensitivity of the concept inventory to the choice of captioner. Excluding the ground-truth class, Gemini yields 3.20 concepts per image on average, while Qwen yields 2.67, corresponding to a difference of 0.53 concepts per image. To assess whether the two captioners identify the same visual concepts, we match their concept inventories using Hungarian one-to-one assignment based on cosine similarity between Stable Diffusion text-encoder embeddings, with an acceptance threshold of 0.8. This yields 1.28 matched concept pairs per image on average. Using the Stable Diffusion text encoder, the matched pairs achieve an average cosine similarity of 0.9853. We further analyze the unmatched concepts. We find that 21.6% arise from naming preferences, such as *woman* versus *person*, while at least 25.2% reflect differences in semantic granularity, such as *hair* versus *person*. Another common source is background or scene elements retained by one captioner but omitted by the other. These results suggest that the disagreement between captioners is mainly attributable to naming and granularity differences rather than entirely distinct object selections.

Table 7: Mean semantic changes under different intervention coefficients on HardImageNet and COCO. We report responses from the in-loss ViT-H/14 backbone and two held-out models, ViT-B/32 and RN50.

		Absolute			Directional		
δ	N	ViT-H	ViT-B/32	RN50	ViT-H	ViT-B/32	RN50
HardImageNet							
-8	3143	-0.0090	-0.0010	-0.0012	-0.0706	-0.0197	-0.0155
-4	3143	-0.0068	-0.0006	-0.0009	-0.0735	-0.0186	-0.0143
-2	3143	-0.0050	-0.0004	-0.0006	-0.0734	-0.0186	-0.0138
0	3143	-0.0021	-0.0001	-0.0002	-0.0661	-0.0147	-0.0092
1	3143	-0.0000	$+0.0000$	-0.0000	$+0.0006$	-0.0001	$+0.0008$
2	3143	$+0.0026$	$+0.0002$	$+0.0002$	$+0.0740$	$+0.0147$	$+0.0112$
4	3143	$+0.0090$	$+0.0005$	$+0.0007$	$+0.0945$	$+0.0200$	$+0.0152$
8	3143	$+0.0250$	$+0.0011$	$+0.0013$	$+0.1111$	$+0.0226$	$+0.0195$

Continued on next page

Table 7 (continued)

		Absolute			Directional		
δ	N	ViT-H	ViT-B/32	RN50	ViT-H	ViT-B/32	RN50
COCO							
-8	2576	-0.0085	-0.0018	-0.0018	-0.0739	-0.0278	-0.0213
-4	2576	-0.0063	-0.0012	-0.0012	-0.0761	-0.0266	-0.0193
-2	2576	-0.0046	-0.0008	-0.0009	-0.0766	-0.0257	-0.0177
0	2576	-0.0019	-0.0003	-0.0003	-0.0704	-0.0238	-0.0127
1	2576	$+0.0000$	$+0.0000$	-0.0000	$+0.0006$	$+0.0015$	-0.0004
2	2576	$+0.0023$	$+0.0003$	$+0.0003$	$+0.0770$	$+0.0212$	$+0.0133$
4	2576	$+0.0075$	$+0.0008$	$+0.0009$	$+0.0962$	$+0.0265$	$+0.0184$
8	2576	$+0.0199$	$+0.0014$	$+0.0016$	$+0.1134$	$+0.0294$	$+0.0225$

A.3 Additional Experimental Results

A.3.1 Spatial and Semantic Change of Scaling-based intervention

We show the semantic response and spatial localization induced by scaling-based interventions in Tab. 7, negative scaling coefficients generally decrease the semantic similarity to the target concept, whereas positive coefficients increase it. This trend is consistent across the in-loss ViT-H/14 encoder and the held-out ViT-B/32 and RN50 encoders, indicating that the induced semantic changes are directionally consistent across different representation spaces.

Tab. 8 further evaluates the spatial distribution of the induced changes. Averaged across all intervention scales, 70.6% and 59.0% of the total pixel-value changes occur within the target masks on HardImageNet and COCO, respectively, despite the target regions covering only 26.2% and 13.3% of the image area on average. Moreover, the mean per-pixel change inside the target masks is 21.96 and 17.24 times that outside the masks, respectively, showing that the interventions are strongly concentrated within the corresponding concept regions.

A.3.2 Additional Results on HardImageNet

HardImageNet provides three object-level interventions on the original images. *Mask* replaces the pixels within the object segmentation mask with gray, *BBox* replaces the entire object bounding box with gray, and *Tile* replaces the bounding box region with an adjacent image patch. These interventions progressively alter or remove the foreground object while preserving the remaining image content. We show qualitative results in fig. 7. Compared to the three ablation-based interventions introduced by HardImageNet benchmark and the diffusion-based editing method Lance, our method induces visually subtle changes while effectively altering model's decisions.

A.3.3 Additional Results on Mitigation

We use scaling-based intervention at $\delta = -8$ to 750 images from the HardImageNet training split. For each image, we identify the non-GT concept whose suppression causes the largest decrease in the ground-truth prediction probability. For each class, the concept that is selected most frequently across its images is treated as the class-level spurious cue. The identified cues for all 15 classes are listed in tab. 9 below:

We use the spurious cues discovered by our framework and adapt three mitigation methods (Yang et al., 2023) to fine-tune RN-50 on the training split. We then evaluate the resulting models on the HardImageNet validation set. The results show that mitigation guided by the discovered cues can substantially improve overall accuracy, particularly for DRO and ERM, while also reducing performance degradation when the identified spurious cues are absent.

A.4 Additional Results on ImageNet

In fig. 9, We provide additional qualitative comparisons for visual counterfactual explanation on ImageNet. We compare MOCA with ACE ℓ_1, ACE ℓ_2, TIME, and MaskDiME across the evaluated

Algorithm 1: Optimization of MOCA

Input: concept directions $\{D_t\}$, masks $\{M_c\}$, intervention scale δ
Output: orthogonal concept frames $\{Q_t\}$
Initialize the learnable parameter matrix A_t as $A_t \leftarrow D_t, \quad \forall t \in \mathcal{T}_w$.
while not converged **do**
  $(\widetilde{Q}_t, R_t) \leftarrow \mathrm{QR}(A_t^\top), \qquad Q_t \leftarrow \widetilde{Q}_t^\top, \quad \forall t \in \mathcal{T}_w$ % QR-based orthogonalization
  $\mathcal{L} \leftarrow 0$
  for each concept $c \in \mathcal{C}$ **do**
    Extract the concept-specific directions $\{q_{t,c}\}_{c\in\mathcal{C}}$ from Q_t.
    Generate $\hat{x}_{c,\delta}$ by intervening along $\{q_{t,c}\}_{t\in\mathcal{T}_w}$ with scale δ
    $\Delta x_c \leftarrow \hat{x}_{c,\delta} - x_0$
    $\mathcal{L} \leftarrow \mathcal{L} + \lambda_{\mathrm{loc}} L_{\mathrm{loc}}^{(c)} + \lambda_{\mathrm{sem}} \mathcal{L}_{\mathrm{sem}}^{(c)} + \mathcal{L}_{\mathrm{align}}^{(c)}$
  end for
  $A \leftarrow \mathrm{Adam}(A, \nabla_A \mathcal{L})$
end while
return $\{Q_t\}_{t\in\mathcal{T}_w}$

Algorithm 2: Learnable Concept Direction $\Delta q_{t,c}$ for Steering Model Decisions

Input: Frozen concept frames $\{Q_t\}_{t\in\mathcal{T}_w}$, concept c, mask M_c, reconstruction x_{recon}, intervention scale δ, decision readout $P(\cdot)$, direction $\sigma \in \{-1, +1\}$
Output: Corrected concept axes $\{q_{t,c}^\star\}_{t\in\mathcal{T}_w}$
Initialize the learnable parameter matrix A with $A_t \leftarrow 0, \forall t \in \mathcal{T}_w$
$\mathcal{L}^\star \leftarrow +\infty$
$A^\star \leftarrow A$
while not converged **do**
  for each step $t \in \mathcal{T}_w$ **do**
    $\mathcal{O}_t \leftarrow Q_{t,-c}$
    $dq_t \leftarrow A_t - (A_t \mathcal{O}_t^\top)\mathcal{O}_t$
    $q'_{t,c} \leftarrow \dfrac{q_{t,c} + \Delta q_{t,c}}{\|q_{t,c} + \Delta q_{t,c}\|_2}$
    $Q'_t \leftarrow Q_t$ with row c replaced by $q'_{t,c}$
  end for
  Generate $\hat{x}_{c,\delta}$ by intervening along $\{q'_{t,c}\}_{t\in\mathcal{T}_w}$ with scale δ
  $\Delta x_c \leftarrow \hat{x}_{c,\delta} - x_0$
  $\hat{x}_c \leftarrow \hat{x}_{c,\delta} \odot M_c + x_0 \odot (1 - M_c)$
  $\mathcal{L}_{\mathrm{loc}} \leftarrow \dfrac{\|(\mathbf{1} - M_c) \odot \Delta x_c\|_1}{\|\Delta x_c\|_1 + \epsilon}$
  $\mathcal{L}_{\mathrm{dec}} \leftarrow -\sigma P(\hat{x}_c)$
  $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{loc}} + \lambda_{\mathrm{dec}} \mathcal{L}_{\mathrm{dec}}$
  if $\mathcal{L} < \mathcal{L}^\star$ **then**
    $\mathcal{L}^\star \leftarrow \mathcal{L}$
    $A^\star \leftarrow A$
  end if
  $A \leftarrow \mathrm{Adam}(A, \nabla_A \mathcal{L})$
end while
Rebuild $\{q_{t,c}^\star\}_{t\in\mathcal{T}_w}$ from $A^\star$ by the residualization and normalization above
return $\{q_{t,c}^\star\}_{t\in\mathcal{T}_w}$

class pairs. The examples further illustrate that MOCA can steer the prediction toward the target class through localized semantic changes while largely preserving the remaining image content.

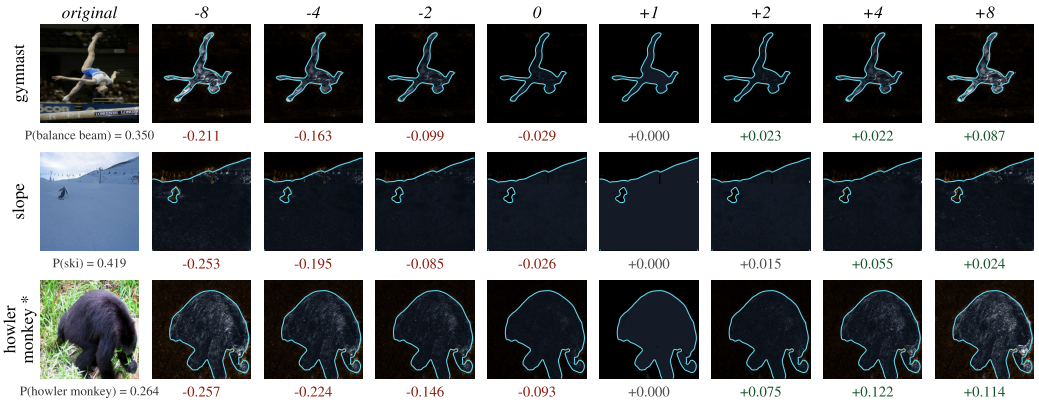

Figure 6: **Visualization of difference maps $x_c(\delta)$ between the intervened and reconstructed images.** We vary the scaling coefficient δ to obtain total intervention strengths of $\{-8, -4, -2, 0, +1, +2, +4, +8\}$ along target concept axis; Examples are selected from HardImageNet and evaluated using ResNet50; The number below each intervention indicates the change in the predicted probability of the corresponding ground-truth class.

Table 8: Spatial localization and magnitude of image changes under different intervention coefficients δ.

	HardImageNet					COCO				
δ	$S_{\text{in}}/S_{\text{tot}}$	$M_{\text{in}}/M_{\text{out}}$	M_{in}	M_{out}	$\lvert\text{diff}\rvert$	$S_{\text{in}}/S_{\text{tot}}$	$M_{\text{in}}/M_{\text{out}}$	M_{in}	M_{out}	$\lvert\text{diff}\rvert$
-8	0.678	18.28	6.926	0.783	1.689	0.585	16.64	**5.390**	0.392	0.889
-4	0.690	19.73	4.410	0.479	1.062	0.594	17.46	3.544	0.244	0.574
-2	0.694	20.15	2.884	0.308	0.689	0.597	17.68	2.397	0.159	0.384
0	0.681	18.01	1.066	**0.117**	0.256	0.573	15.96	0.941	0.064	0.152
2	0.698	19.87	1.119	**0.117**	0.266	0.576	16.28	0.952	**0.063**	0.152
4	0.738	26.19	3.264	0.291	0.748	**0.602**	18.40	2.489	0.158	0.391
8	**0.766**	31.52	**7.025**	0.555	1.567	**0.602**	18.28	4.829	0.311	0.756
Average	0.706	21.96	3.813	0.379	0.897	0.590	17.24	2.935	0.199	0.471

Table 9: Class-wise spurious cues identified on HardImageNet. For each class, we report the selected cue, the number of images in which the cue is present, and the number of cases in which suppressing the cue decreases the ground-truth prediction probability.

Class	Cue	Images	Drops
balance beam	walls	22/50	11
baseball player	bat	16/50	9
dog sled	snow	43/50	17
gymnastic horizontal bar	trees	24/50	13
hockey puck	labels	7/50	6
howler monkey	rocks	11/50	5
keyboard space bar	typewriters	12/50	9
miniskirt	walls	18/50	12
patio	chairs	21/50	15
seat belt	buckle	17/50	10
ski	walls	13/50	8
snorkel	goggles	13/50	7
sunglasses	people	8/50	6
swimming cap	water	36/50	16
volleyball	sky	27/50	15

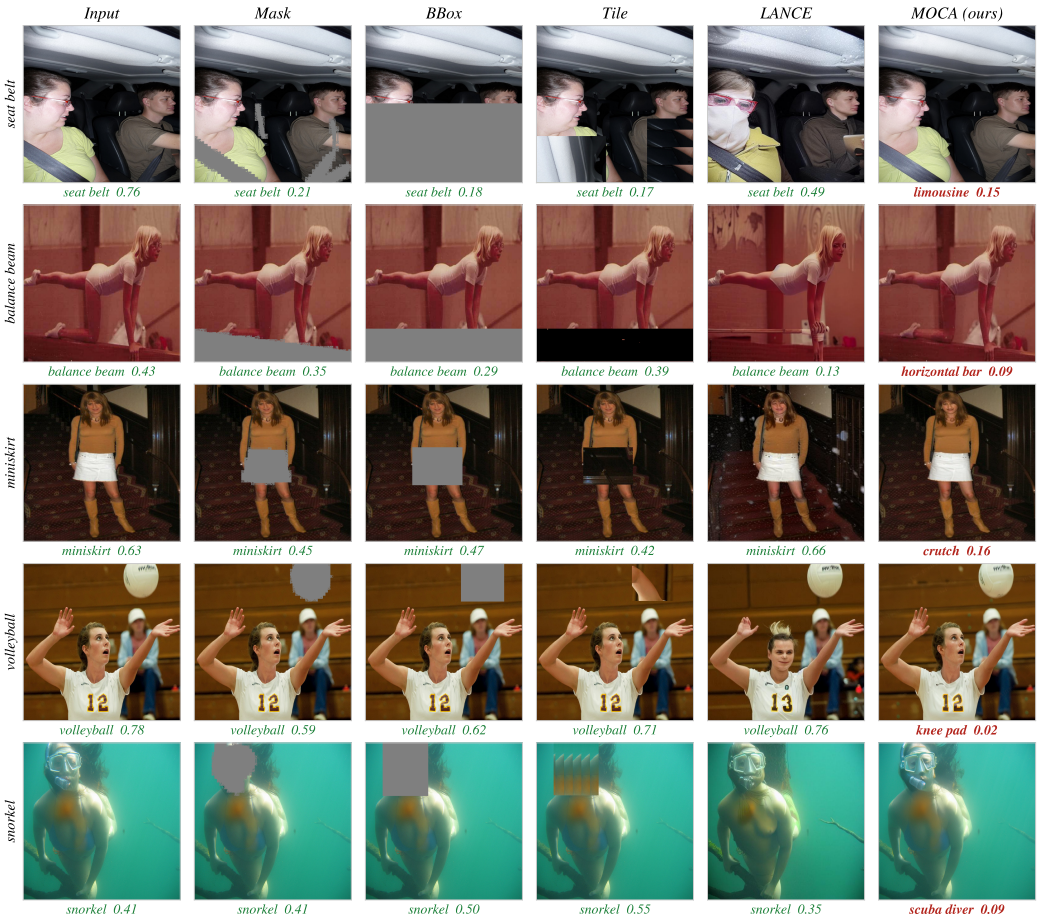

Figure 7: Additional qualitative results on HardImageNet. We compare MOCA with existing intput intervention and counterfactual editing methods, showing the resulting image changes and corresponding top-1 predictions.

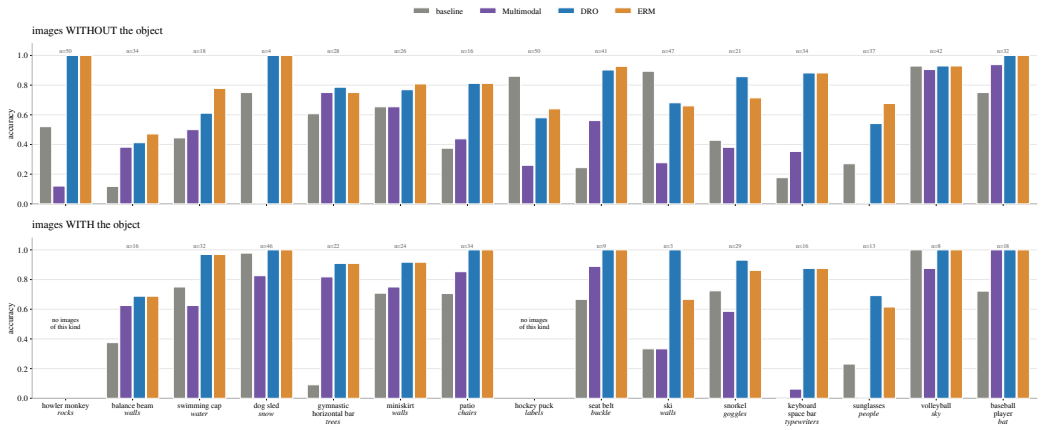

Figure 8: **Class-wise mitigation results on HardImageNet.** We report classification accuracy for each HardImageNet class, separately on images where the identified spurious cue is absent (top) or present (bottom). The identified cue for each class is shown in *italics* below the class name, and n denotes the number of images in each subgroup.

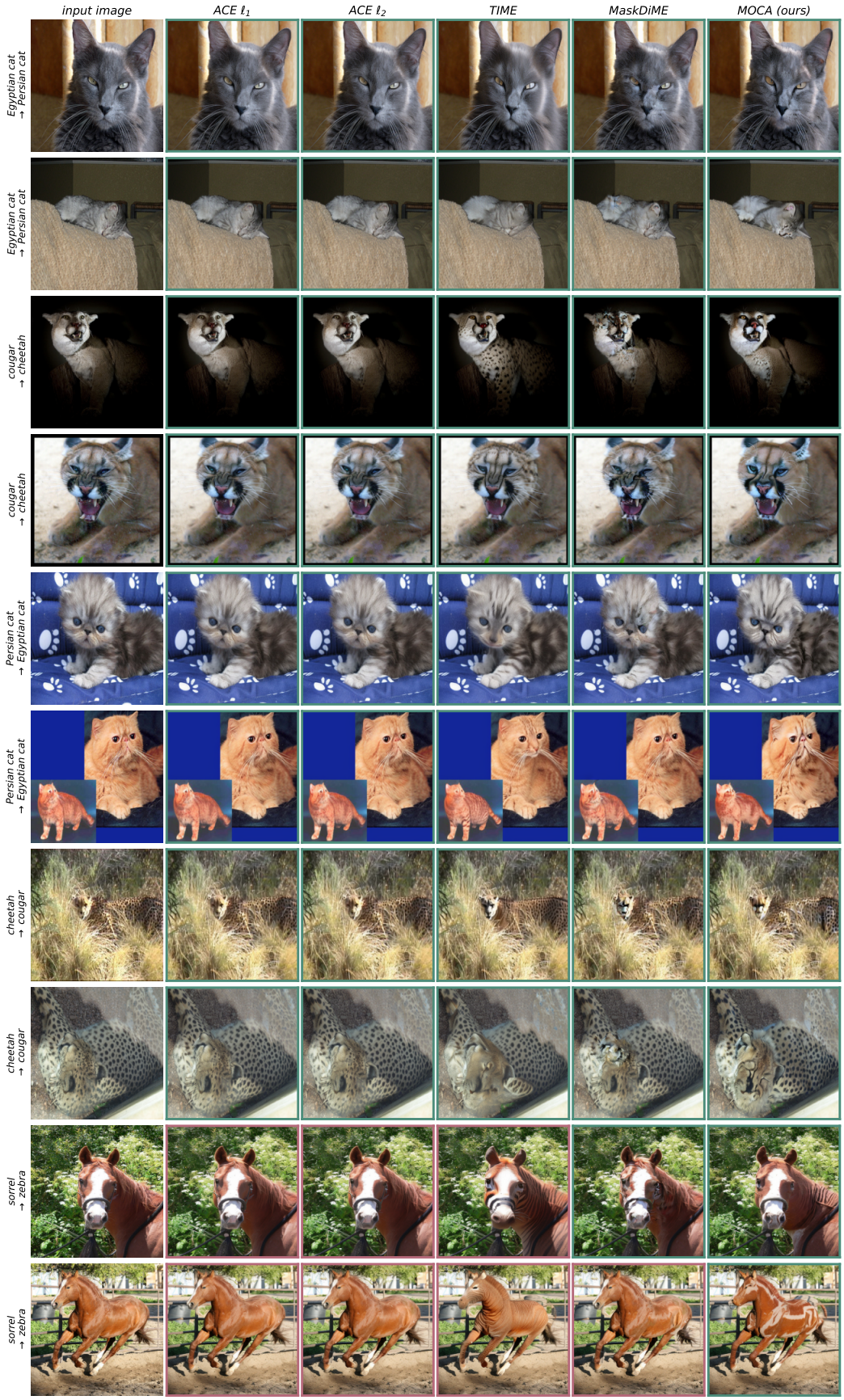

Figure 9: **Additional qualitative results of visual counterfactual explanation.**