INTERVENING IN CONCEPT BOTTLENECK MODEL VIA CAUSAL EFFECTS FROM PIXEL TO CONCEPT

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept bottleneck models (CBMs) employ human-interpretable concepts for predictions, enabling corrective interventions in neural networks. However, current intervention methods are usually limited to the concept bottlenecks encoded by features or their corresponding weight matrices, leading to inefficient interventions, limited interpretability, and dependence on human priors that restrict generalization. To address this, we propose CBM-PCCEI, a Pixel-to-Concept Causal Effect Intervention method for CBMs. It utilizes causal effects hierarchically to achieve multi-level intervention, improves effectiveness, and provides a more easily understood explanation of reasoning and intervention. To generalize improvements beyond the intervened cases and reduce dependence on human priors, we integrate the interventions with knowledge distillation, transferring the improved behavior to the base model. Evaluations on Flower-102, Aircraft, CUB-200, CIFAR-10, and CIFAR-100 with NFResNet-50 and DaViT show that our method improves validation accuracy by up to 11.01% with an average gain of 6.62% via single patch intervention, and achieves up to 3.82% with an average of 1.48% testing accuracy gain under generalization settings, outperforming both vanilla fine-tuning and state-of-the-art intervention strategies.

1 INTRODUCTION

Deep learning has achieved remarkable success in image classification. However, its limited interpretability raises concerns regarding security and reliability, driving substantial growth in explainable AI (XAI) research. Among various approaches, concept bottleneck models (CBMs) (Koh et al., 2020) have garnered significant attention due to their inherent interpretability and interveneability.

On the one hand, CBMs introduce an intervention mechanism based on an interpretable concept prediction structure, addressing the limitation of post-hoc explanation methods, such as feature attribution and concept-based explanations (Simonyan et al., 2013; Dammu & Shah, 2023; Selvaraju et al., 2016; Zhang et al., 2024; Zhou et al., 2018; Ghorbani et al., 2019; Zhang et al., 2020; Kamakshi et al., 2021; Fel et al., 2022; Kowal et al., 2024; Ge et al., 2021; Ahn et al., 2024), which often lack the capacity to correct model behavior.

On the other hand, the explanatory form of CBMs remains relatively simplistic, making them harder to comprehend compared to more elaborated conceptual explanations and attribution techniques. More critically, Chauhan et al. (2022) highlight that this mode of interaction heavily relies on expert knowledge, hindering its generalization in large-scale deployments. Furthermore, as noted by Shin et al. (2023), existing intervention methods typically operate only on the concept predictions (score) or the linear mappings (concept weight matrix), overlooking potential inherent biases in upstream feature extractors, leading to inefficient interventions.

Inspired by advances in handling illusion within multi-modal large models (Zhou et al., 2025), we explore whether a bridge can be constructed between the input image and the concept bottleneck, so as to enhance intervention effectiveness while preserving interpretability through hierarchical intervention. While gradient-based methods offer an intuitive pathway, they tend to introduce noise at both pixel and concept levels (Adebayo et al., 2018; Ghorbani et al., 2017; Rudin, 2018).

To address these issues, we propose CBM-PCCEI, a **P**ixel to **C**oncept **C**ausal **E**ffect **I**ntervention method based on two-stage Causal Effect Map (CEM) (Hu et al., 2024) for concept bottleneck models.

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.

Our approach hierarchically leverages causal effects (Average Treatment Effect, ATE) (Rubin, 1974) to uncover concepts/pixels detrimental to classification, from input pixels to latent concepts and onward to predictions, enabling interpretable interventions at pixel, concept, and joint levels. These interventions are further coupled with knowledge distillation to improve generalization ability of the original model. Our contributions are summarized as follows

i) *Hierarchical intervention.* It enhances interpretability and effectiveness of CBM interventions while overcoming the limitation that concept-level interventions cannot propagate corrections to upstream feature.

ii) *Explainable intervention.* By integrating causal attribution with concept-based explanations, it provides a clear, multi-level understanding of both the reasoning process and the effect of each intervention step.

iii) *Efficient intervention generalization.* Building upon an existing knowledge distillation-based intervention generalization architecture, it enables intervention generalization across all classes in CBMs while improving generalization performance.

2 Related Work

We first review literature related to concept bottleneck models, then examine intervention strategies used within them. Following this, we survey work on causal interventions and effect analyses. These research provide a basis for designing a pixel $\rightarrow$ concept $\rightarrow$ prediction intervention mechanism that supports interpretable, multi-level interventions.

2.1 Concept Bottleneck Models

Koh et al. (2020) first introduces CBM and proposes concept-level interventions driven by human priors. While straightforward, the method relies on expert knowledge and shows limited generalization at test time, as it requires concept labels or user interaction. Yuksekgonul et al. (2022) proposes PCBM, which augments the bottleneck with residual links to backbone features and enables global intervention on the concept weight matrix, although smoothing masks applied to the concept matrix yield only modest gains. Steinmann et al. (2023) develops CB2M, which stores user interventions in a memory module and replays them to enable autonomous test time corrections, thereby building a mistake memory that offers partial generalization. Chauhan et al. (2022) introduces an interactive policy that combines concept prediction uncertainty with a concept's influence on the target class to improve overall payoff. Xu et al. (2024) presents an ECBM that models relationships among concepts and supports transitive intervention effects.

2.2 Concept Intervention Strategies

Beyond specific CBMs, Shin et al. (2023) summarizes negative concept selection rules, including uncertain concept prediction (UCP), contribution to target prediction (CCTP), and expected contribution to prediction (ECTP). Concept style interventions have also been adapted to black box settings. Marcinkevics et al. (2024) reports black box interventions guided by concept signals. Notably, CBM-HNMU Xiong et al. (2025) selects negative concepts using global gradient contributions and demonstrates effectiveness across multiple backbone architectures and datasets.

2.3 Causal intervention and effect analysis

Causal intervention and effect analysis provide tools to quantify how changes to inputs or latent variables alter predictions (Karimi et al., 2022; Feder et al., 2021; Lu et al., 2022). With the rise of XAI, causal reasoning has been employed to construct counterfactual explanations (Laugel et al., 2019; Mothilal et al., 2019). Goyal et al. (2019) generates counterfactual visual explanations that localize image regions responsible for confusing class distinctions. Within the CBMs, Dominici et al. (2024a) proposes CF-CBM, which jointly trains a counterfactual generator with the bottleneck to encourage reliance on fewer salient concepts and to amplify the causal effect of concept interventions on predictions. Dominici et al. (2024b) presents CCGM, a concept causal graph model that matches the generalization performance of opaque networks while allowing human in the loop corrections

of hidden concepts and class predictions. For low-level vision tasks, Hu et al. (2024) introduces causal effect maps (CEMs) to quantify how individual input pixels influence specific output regions, reporting substantial performance gains by intervening on negatively associated pixels.

3 METHODOLOGY

We present CBM-PCCEI, a hierarchical causal effect intervention method that couples input pixels with latent concepts to steer class predictions in an interpretable manner. The framework is formalized and illustrated in **Figure 1**. Interventions follow a pixel → concept → prediction pathway. The procedure first identifies negative concepts, then modifies input pixel patches to suppress them while preserving the expression of non-negative concepts. This corrects model errors and improves classification performance while maintaining interpretability **(Sec. 3.2)**. Moreover, by linking inputs, latent concepts, and outputs through causal effects, our approach provides multi-level, clearer concept explanations that integrate visual and natural language cues **(Sec. 3.3)**. To minimize reliance on human prior knowledge, the method automatically selects negative pixels and concepts based on causal effects, and finally employs knowledge distillation to transfer the refined behavior to the original base model, thereby enabling generalization beyond directly intervened cases **(Sec. 3.4)**.

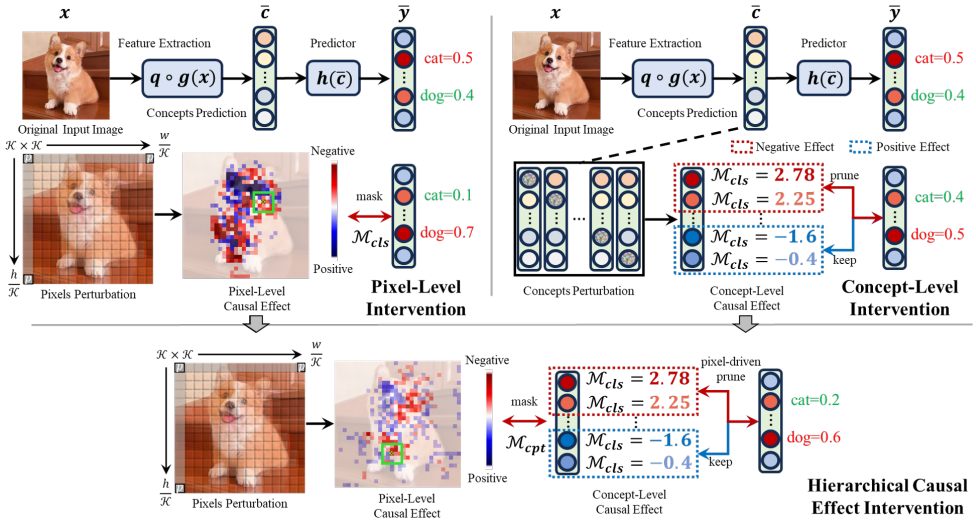


Figure 1: Overview of the Hierarchical Causal Effect Intervention. **Pixel-level intervention** directly utilizes causal effects to measure the relationship between input image patches and discriminant metrics ($\mathcal{M}_{\text{cls}}$), thereby identifying negative image regions for intervention to correct predictions. It can be applied to feature extraction and is more efficient, but lacks interpretability. **Concept-level intervention** uses the impact of each concept in the perturbation bottleneck on the distinction between the true and confused classes as $\mathcal{M}_{\text{cls}}$ to measure the causal effect. This helps identify concepts that are detrimental to distinguishing between the two (such as cats and dogs) and intervene to correct predictions, but it cannot address feature bias. After balancing the two factors, **Hierarchical Causal Effect Intervention** cascades causal effects from pixels to concepts ($\mathcal{M}_{\text{cpt}}$) and then to predictions, thereby achieving interpretable and efficient interventions (see architecture details in **Appendix. J**).

3.1 PROBLEM SETUP AND NOTATION

We adopt three data splits: train, validation, and test. For a given split, the dataset is defined as $\mathcal{D}^{\text{split}} = \{(x_i, y_i)\}_{i=1}^{n_{\text{split}}}$ with images $x_i \in \mathbb{R}^d$ (d refer to $w \times h \times ch$) and labels $y_i \in \{1, \ldots, n_{\text{cls}}\}$. In the CBM setting, each sample may also include a concept label $c_i \in \mathbb{R}^{n_{\text{cpt}}}$, yielding data points like $\{(x_i, y_i, c_i)\}_{i=1}^{n_{\text{split}}}$. A CBM applies three maps in sequence. The feature extractor g produces a hidden activation $v_i = g(x_i) \in \mathbb{R}^k$ and k is the dimensions of corresponding hidden layer feature. The concept head q compares v_i with the concept bottleneck $v_{\text{cpt}} \in \mathbb{R}^{n_{\text{cpt}} \times k}$ and returns concept scores $\bar{c}_i = q(v_i, v_{\text{cpt}}) \in \mathbb{R}^{n_{\text{cpt}}}$. The classifier h then predicts from these scores with $\bar{y}_i = h(\bar{c}_i) = \bar{c}_i W$ where $W \in \mathbb{R}^{n_{\text{cpt}} \times n_{\text{cls}}}$. In summary, a standard Deep Neural Network (DNN) maps x directly to y, while CBM instead factors the decision as $x \to c \to y$, thereby exposing an interpretable interface suitable for intervention. Classical CBMs supervise c using human-annotated labels. We replace these labels with concept targets derived from a shared image–text encoder (CLIP) (Radford et al.,

2021). Each input image is encoded by CLIP and then matched with concept embeddings in the concept bottleneck to form surrogate concept scores, which provide rich supervision without requiring additional annotation (Yang et al., 2022).

3.2 Hierarchical Causal Effect Intervention

Existing approaches improve CBM decisions by editing either (i) selected concept scores or (ii) selected entries of the concept weight matrix based on human prior knowledge and known data bias during the testing phase. Let $S \subseteq \{1, \ldots, P\}$ (Concept Indices) and $T \subseteq \{1, \ldots, P\} \times \{1, \ldots, K\}$ (Class Indices) be the edited indices, with replacement value $\mathcal{W}$ (typically 0), i.e.,

$$(\tilde{c})_j = \begin{cases} \mathcal{W}, & j \in S \\ \bar{c}_j, & j \notin S \end{cases} \quad (\tilde{W})_{u,j} = \begin{cases} \mathcal{W}, & (u,j) \in T \\ W_{u,j}, & (u,j) \notin T \end{cases}.$$

The intervention decision is

$$\hat{y}^\star = \arg\max_{\hat{y}} h(\tilde{c}) \quad \text{or} \quad \hat{y}^\star = \arg\max_{\hat{y}} h_{\tilde{W}}(\bar{c}).$$

Current intervention methods based on CBMs cannot alter the upstream representation $q \circ g$, and rely on human priors, which limits intervention when human knowledge or concept labels are unavailable and results in limited performance gains. To address these limitations, we introduce a hierarchical intervention method that follows a pixel $\rightarrow$ concept $\rightarrow$ prediction pathway. In order to balance interpretability with intervention effectiveness, our method begin with interventions of negative concepts driven by causal effects.

Concept-Level Intervention: For each input image with its corresponding class label (x, y), we identify concepts that impair the prediction of the true class. To obtain a reliable causal effect of each concept, we introduce a classification-based expected gain metric, $\mathcal{M}_{\text{cls}}$. The metric repeatedly perturbs the bottleneck layer and averages the resulting changes in class scores, thereby approximating the average treatment effect of concept interventions while mitigating noise from gradient estimation and forward propagation.

Specifically, for a concept index $i \in \{1, \ldots, n_{\text{cpt}}\}$ and a class index $j \in \{1, \ldots, n_{\text{cls}}\}$, we have a perturbed concept matrix by replacing the entry $W_{i,j}$ with a value $\mathcal{W} \in [0, 1]$ while keeping all other entries fixed. We focus on $j = y$ to assess how neutralizing concept i affects the true class. Let $s(x, W)$ denote the logits and define the margin

$$\text{mar}(x, W) = s_y(x, W) - \max_{k \neq y} s_k(x, W). \tag{1}$$

A single perturbation replaces $W_{i,y}$ by $\mathcal{W}$ (write $\tilde{W} = W^{(i \leftarrow \mathcal{W})}$) and induces the change

$$\mathcal{M}_{\text{cls}}(i, \mathcal{W}) = \text{mar}\big(x, W^{(i \leftarrow \mathcal{W})}\big) - \text{mar}(x, W). \tag{2}$$

Positive values indicate that disturbance concept i improves the decision of class y.

We adopt the coarse–fine procedure to balance computation and the reliability of causal effects. In the coarse stage, draw $\mathcal{W}_1, \ldots, \mathcal{W}_{n_{\text{coarse}}}$ independently from $[0, 1]$ and set

$$v_{\text{causal}}^{\text{coarse}}(i) = \frac{1}{n_{\text{coarse}}} \sum_{k=1}^{n_{\text{coarse}}} \mathcal{M}_{\text{cls}}(i, \mathcal{W}_k), \quad i = 1, \ldots, n_{\text{cpt}}. \tag{3}$$

To refine the coarse results of causal effect for sensitive concepts, choose a threshold $\theta_1 > 0$ and, for those with $|v_{\text{causal}}^{\text{coarse}}(i)| > \theta_1$, compute

$$v_{\text{causal}}^{\text{fine}}(i) = \frac{1}{n_{\text{fine}}} \sum_{k=1}^{n_{\text{fine}}} \mathcal{M}_{\text{cls}}(i, \mathcal{W}_k), \tag{4}$$

and take $v_{\text{causal}}^{\text{fine}}(i) = 0$ otherwise. The final score is

$$v_{\text{causal}}(i) = \begin{cases} v_{\text{causal}}^{\text{coarse}}(i), & v_{\text{causal}}^{\text{fine}}(i) = 0, \\ v_{\text{causal}}^{\text{fine}}(i), & \text{otherwise.} \end{cases} \tag{5}$$

We rank v_{causal} in descending order and select the top $\mathcal{N}$ indices $\mathcal{I}_{\text{neg}}^{\text{cpt}} \subseteq \{1, \ldots, n_{\text{cpt}}\}$, then set

$$W_{i,y} \leftarrow 0 \quad \text{for all } i \in \mathcal{I}_{\text{neg}}^{\text{cpt}},$$

which reduces negative conceptual influence and corrects the prediction.

As noted earlier, concept-level interventions cannot affect upstream features. To address this, we propose to directly link inputs with outputs, thereby shifting the causal scoring to the input level.

Pixel-Level Intervention: Based on the margin defined in equation 1, we estimate the effect of replacing local image content. Partition $x \in \mathbb{R}^{w \times h \times ch}$ into a uniform $\mathcal{K} \times \mathcal{K}$ grid of patches. For a patch index $(u = \frac{w}{\mathcal{K}}, v = \frac{h}{\mathcal{K}})$ we replace $p_{u,v}$ by a natural crop $\hat{p}$ drawn from DIV2K (Timofte et al., 2018), which stabilizes the estimate and avoids off-manifold artifacts. The single perturbation effect is

$$\Delta_{\text{patch}}(u, v, \hat{p}) = \text{mar}\big(x^{(u,v \leftarrow \hat{p})}, W\big) - \text{mar}(x, W). \tag{6}$$

In the coarse-stage, we sample $\hat{p}_1, \ldots, \hat{p}_{n_{\text{coarse}}}$ as random crops from DIV2K and compute CEM

$$m_{\text{causal}}^{\text{coarse}}(u, v) = \frac{1}{n_{\text{coarse}}} \sum_{k=1}^{n_{\text{coarse}}} \Delta_{\text{patch}}(u, v, \hat{p}_k). \tag{7}$$

To refine sensitive locations we choose a threshold $\theta_2 > 0$. For entries with $|m_{\text{causal}}^{\text{coarse}}(u, v)| > \theta_2$ we draw n_{fine} additional crops and set

$$m_{\text{causal}}^{\text{fine}}(u, v) = \frac{1}{n_{\text{fine}}} \sum_{k=1}^{n_{\text{fine}}} \Delta_{\text{patch}}(u, v, \hat{p}_k), \tag{8}$$

and take $m_{\text{causal}}^{\text{fine}}(u, v) = 0$ otherwise. The final map is

$$m_{\text{causal}}(u, v) = \begin{cases} m_{\text{causal}}^{\text{coarse}}(u, v), & m_{\text{causal}}^{\text{fine}}(u, v) = 0, \\ m_{\text{causal}}^{\text{fine}}(u, v), & \text{otherwise.} \end{cases} \tag{9}$$

We sort m_{causal} and select the top $\mathcal{P} \in \mathbb{N}$ locations to get $\mathcal{I}_{\text{neg}}^{\text{patch}}$. For the intervention we replace $p_{u,v} \leftarrow \hat{p}$ with $\hat{p}$ sampled from DIV2K for all $(u, v) \in \mathcal{I}_{\text{neg}}^{\text{patch}}$, which removes harmful pixels and increases the margin between the ground-truth class and its strongest competitor.

While end to end pixel-level interventions act directly on the feature extractor and influence the margin, they ignore the structure of the concept bottleneck. Consequently, such interventions are difficult to interpret, cannot target specific hidden concepts, and may even negatively impact generalization. We therefore couple pixels and concepts within a hierarchical intervention that links local image regions to concept predictions, and through them, to the classification decision.

Hierarchical Intervention from Pixel to Concept: We start from the negative concept set $\mathcal{I}_{\text{neg}}^{\text{cpt}}$ identified previously. For a patch index (u, v) and a random $\hat{p}$, let

$$\hat{c}(u, v, \hat{p}) = q\big(g(x^{(u,v \leftarrow \hat{p})}),\, v_{\text{cpt}}\big) \text{ and } \bar{c} = q\big(g(x),\, v_{\text{cpt}}\big).$$

Write $c^-(u, v, \hat{p})$ and c^+ for the restrictions to $\mathcal{I}_{\text{neg}}^{\text{cpt}}$ and its complement, respectively. We score a patch perturbation by penalizing negative concepts while preserving positive concepts:

$$\mathcal{M}_{\text{cpt}}(u, v, \hat{p}) = -\big\|\hat{c}^-(u, v, \hat{p})\big\|_2^2 \; - \; \big\|\hat{c}^+(u, v, \hat{p}) - \bar{c}^+\big\|_2^2.$$

$\mathcal{M}_{\text{cpt}}$ is then centered to distribute around zero. Large values indicate that the patch intervention suppresses harmful concepts without distorting supportive ones.

We next reuse the coarse–fine procedure in equation 7–equation 9 by replacing Δ_{patch} with $\mathcal{M}_{\text{cpt}}$. This yields a concept-aware map $m_{\text{causal}}^{\text{cpt}}(u, v)$ on the same grid. Finally, we select the top $\mathcal{P}$ locations and replace $p_{u,v}$ with $\hat{p}$ for all (u, v) in this set. In this way, pixel interventions are guided by concept objectives, negative concepts are smoothly suppressed, positive concepts are preserved, and the intervention remains interpretable while acting upstream on the features.

Algorithm 1 Concept Causal Attribution

Input Patch features p^{fea}, concept bottleneck v_{cpt}, crop embeddings $v^{\bar{y}}_{\text{crops}}$ on $\mathcal{D}_{crops}$, cluster $\mathcal{C}$, $\mathcal{Q}$
Output $\mathcal{I}^{\text{cpt}}_{\text{exp}} = \{\mathcal{I}^{\text{cpt}}_{\text{exp}}[c]\}^{\mathcal{C}}_{c=1}$, $\mathcal{I}^{\text{crops}}_{\text{exp}} = \{\mathcal{I}^{\text{crops}}_{\text{exp}}[c]\}^{\mathcal{C}}_{c=1}$

1: Run k-means on $\{p^{\text{fea}}_{u,v}\}$ to obtain $\mathcal{C}$ clusters and their centroids M_c
2: **for** $c = 1$ to $\mathcal{C}$ **do**
3: Compute the concept gap $G_c \leftarrow -|M_c - v_{\text{cpt}}|$
4: Select the Top–$\mathcal{Q}$ concept indices $\mathcal{I}^{\text{cpt}}_{\text{exp}}[c]$ from G_c
5: Form a representative concept embedding by averaging the selected rows of v_{cpt}

$$v^{\text{cpt}}_{\text{exp}}[c] \;=\; \frac{1}{\mathcal{Q}} \sum_{j \in \mathcal{I}^{\text{cpt}}_{\text{exp}}[c]} v_{\text{cpt}}[j]$$

6: Match $v^{\text{cpt}}_{\text{exp}}[c]$ to $v^{\bar{y}}_{\text{crops}}$ in the shared space and select the Top–$\mathcal{Q}$ crop indices $\mathcal{I}^{\text{crops}}_{\text{exp}}[c]$
7: **end for**

3.3 Explanation via Concept Causal Attribution

For abstract concepts, human often struggle to comprehend CBMs' reasoning and intervention basis. In this section, we employ the causal effect map $m^{\text{cpt}}_{\text{causal}}(u, v)$ to attribute concepts to sensitive image patches. Furthermore, our method integrates class specific crops extracted from the corresponding dataset ($\mathcal{D}_{\text{crops}}$) (Fel et al., 2022) with cross modally aligned encodings from CLIP, thereby generating visual concept explanations that correspond to natural language concepts.

Specifically, for each sample image x, we select patches where $\left|m^{\text{cpt}}_{\text{causal}}(u, v)\right| > \theta_3$ ($\theta_3 > 0$) and repeatedly perturb these patches using the replacement crop $\hat{p}$. We then calculate the average of $\hat{c}^-$ across the number of interventions as the clustering feature for the corresponding patch to get the patch feature p^{fea}. With all components prepared, concept causal attribution is performed as in **Alg. 1**. First, using K-means (Mcqueen, 1967) to group these features and patches, with a hyperparameter $\mathcal{C}$ pre-defined to set the number of groups. Next, within each cluster, we select the top $\mathcal{Q}$ natural language concepts with the largest G_c as the explanation of the corresponding causal effect map area.

Finally, the text encoder of CLIP embeds the natural language concepts matched to each cluster into a shared space, and their mean feature is computed as $v^{\text{cpt}}_{\text{exp}}$. Concurrently, the image encoder projects crop images from $\mathcal{D}_{\text{crops}}$ (spanning different classes) into the same space (v_{crops}), which are then matched against the mean feature of each concept cluster to obtain the corresponding visual concept explanations. In summary, through causal concept attribution, users can understand the model's reasoning and intervention processes from multiple modalities and complementary perspectives.

3.4 Intervention Generalization

Unlike intervention methods that require concept annotations or human priors at test time, our approach uses only ground-truth to construct hierarchical causal effects. To transfer the gains achieved through interventions without relying any labels during testing, we perform the intervention on $\mathcal{D}^{\text{val}}$ and distill the improved behavior into the CBM. For each $(x, y) \in \mathcal{D}^{\text{val}}$, we compute the baseline concepts $\bar{c} = q(g(x), v_{\text{cpt}})$ and intervened concepts $\hat{c}$ from the intervened input (e.g., $x^{(u,v \leftarrow \hat{p})}$ aggregated as in **Sec. 3.2**). We retain the sample only if the margin $\mathcal{M}_{\text{cls}}$ increases after intervention (i.e., the intervention is beneficial). The model is then updated to match intervened logits and concepts while respecting the class label, i.e,

$$\mathcal{L}_{\text{int}} = \alpha \left\|h(\hat{c}) - h(\bar{c})\right\|^2_2 + \beta \left\|\hat{c} - \bar{c}\right\|^2_2 + \gamma\, \mathcal{L}_{\text{cls}}, \tag{10}$$

$$\mathcal{L}_{\text{cls}} = -\sum_{k=1}^{n_{\text{cls}}} \mathbf{1}[k{=}y]\; \log \pi_k,$$

where $\pi = \text{softmax}\big(h(\bar{c})\big)$. For a fair comparison, we also fine-tune the original CBM on $\mathcal{D}^{\text{val}}$ using only $\mathcal{L}_{\text{cls}}$ and the same optimization settings. Importantly, the samples for fine-tuning the model are consistent with the positive samples selected during the distillation. This ensures that the distinction between intervention generalization and label-based fine-tuning remains clear and unambiguous.

4 EXPERIMENTS

4.1 INTERVENTION BENEFITS ON VALIDATION

We evaluate interventions on two backbones, a CNN (NFResNet-50) and a ViT (DaViT), across five datasets: Flower-102 (Nilsback & Zisserman, 2008), Aircraft (Maji et al., 2013), CUB-200 (Wah et al., 2011), for fine-grained recognition, and CIFAR-10 and CIFAR-100 (Krizhevsky, 2009) for low-resolution images. We evaluate the intervention effects at different levels, including concept-level intervention (CI), pixel-level intervention (PI), and hierarchical intervention (HI), with CBM-HNMU included as a competitor. Results are summarized in **Table 1**.

Table 1: Evaluation on intervention benefits ($\mathcal{K} = 8, \mathcal{N} = n_{\text{cpt}} * 7.8\%, \mathcal{P} = 1$). *Accuracy* (ACC) list the classification effects of the model baseline (Base), after concept-level intervention (CI), pixel-level intervention (PI), and hierarchical intervention (HI) on the validation set ($\mathcal{D}^{\text{val}}$). $\mathcal{M}_{\text{cls}}$ is used to measure the model's ability to distinguish true class. The larger it is, the greater the distance between the logits of the true class and the most confusion class. $\triangle \uparrow$ and $\% \uparrow$ are represent the increase value/percent after intervention compare with the baseline. The pink row represents the improvement in classification accuracy of CBM-HNMU and our method compared to baseline after intervention, as well as the gain difference between the two ($\triangle$). The underlined data are those following HI. Due to the randomness introduced by $\hat{p}$, all pixel-oriented interventions are tested three times, denoted as AVG±STD. Further comparisons with CBM-HNMU are provided in **Appendix. O.**

		Flower-102			Aircraft			CUB-200			CIFAR-10			CIFAR-100			
		Base	CI	HI / PI	Base	CI	HI / PI	Base	CI	HI / PI	Base	CI	HI / PI	Base	CI	HI / PI	Avg ↑ (HI)
NFResNet-50 (Mittal et al., 2022)	ACC (%)	96.08	96.33	$\frac{97.69 \pm 0.09}{97.67 \pm 0.06}$	63.72	65.56	$\frac{74.73 \pm 0.36}{76.24 \pm 0.15}$	63.60	65.50	$\frac{73.13 \pm 0.11}{73.71 \pm 0.32}$	96.20	96.56	$\frac{98.65 \pm 0.07}{98.77 \pm 0.02}$	81.14	81.98	$\frac{89.64 \pm 0.16}{90.33 \pm 0.19}$	**6.62** ↑
	$\mathcal{M}_{\text{cls}}(\triangle \uparrow)$	-	0.04	$\frac{0.23 \pm 0.009}{0.34 \pm 0.003}$	-	0.34	$\frac{0.82 \pm 0.009}{1.29 \pm 0.008}$	-	0.48	$\frac{1.05 \pm 0.034}{1.45 \pm 0.027}$	-	0.94	$\frac{0.63 \pm 0.004}{0.69 \pm 0.006}$	-	0.32	$\frac{1.35 \pm 0.009}{1.87 \pm 0.019}$	**0.82** ↑
	.vs CBM-HNMU ACC (% ↑)	Target	Ours	△	Target	Ours	△	Target	Ours	△	Target	Ours	△	Target	Ours	△	**6.13** ↑
		0.12	**1.61**	**1.49** ↑	0.87	**11.01**	**10.14** ↑	0.80	**9.53**	**8.73** ↑	0.32	**2.45**	**2.13** ↑	0.32	**8.50**	**8.18** ↑	
		Base	CI	HI	Base	CI	HI	Base	CI	HI	Base	CI	HI	Base	CI	HI	Avg ↑ (HI)
DaViT (Ding et al., 2022)	ACC (%)	96.88	97.06	$\underline{97.59 \pm 0.08}$	72.93	73.92	$\underline{82.62 \pm 0.11}$	69.85	71.35	$\underline{78.45 \pm 0.33}$	97.68	97.96	$\underline{98.62 \pm 0.07}$	79.18	80.50	$\underline{85.50 \pm 0.11}$	**5.25** ↑
	$\mathcal{M}_{\text{cls}}(\triangle \uparrow)$	-	0.03	$\underline{0.12 \pm 0.002}$	-	0.221	$\underline{0.623 \pm 0.004}$	-	0.38	$\underline{0.98 \pm 0.022}$	-	3.69	$\underline{0.22 \pm 0.007}$	-	0.35	$\underline{0.94 \pm 0.005}$	**0.80** ↑

Quantitatively, HI exceeds CBM-HNMU by as much as 10.14% with an average gain of 6.13%. Relative to the baseline, HI achieves up to 11.01% higher accuracy, with an average gain of 6.62%. Overall, CI yields modest gains because interventions at the concept-level do not influence upstream feature extraction, which leaves backbone biases in place. PI acts directly on the input, improves accuracy and $\mathcal{M}_{\text{cls}}$ substantially. However, PI provides weaker interpretability since it bypasses the concept pathway. HI combines these strengths by coupling pixel intervention with concept control, so it preserves interpretability and yields improvements comparable to PI. See **Appendix. A** for comparisons with LCP, CCTP, and ECTP, and **Appendix. B** for comparisons with PCBM. HI consistently demonstrates superior performance across all evaluated intervention strategies.

Table 2: Evaluation on generalization benefits of hierarchical intervention (HI-Gen) and compared with the corresponding black box of the same backbone using CBM-HNMU intervention.

		Flower-102		Aircraft		CUB-200		CIFAR-10		CIFAR-100		
		Base	Fine-Tune / HI-Gen	Base	Fine-Tune / HI-Gen	Base	Fine-Tune / HI-Gen	Base	Fine-Tune / HI-Gen	Base	Fine-Tune / HI-Gen	Avg ↑ (HI-Gen)
NFRes50	ACC (%)	96.26	96.63 ± 0.04 $\underline{\mathbf{97.18 \pm 0.02}}$	63.63	65.71 ± 0.22 $\underline{\mathbf{67.45 \pm 0.15}}$	63.46	64.92 ± 0.04 $\underline{\mathbf{65.27 \pm 0.05}}$	95.42	95.49 ± 0.03 $\underline{\mathbf{95.66 \pm 0.07}}$	81.33	81.55 ± 0.02 $\underline{\mathbf{81.93 \pm 0.04}}$	**1.48** ↑
	$\mathcal{M}_{\text{cls}}(\triangle \uparrow)$	-	0.0348 ± 0.0941 $\underline{\mathbf{0.8048 \pm 0.0123}}$	-	0.2563 ± 0.0774 $\underline{\mathbf{0.7441 \pm 0.0096}}$	-	0.1609 ± 0.0440 $\underline{\mathbf{0.3068 \pm 0.0001}}$	-	-0.0506 ± 0.6999 $\underline{\mathbf{3.3282 \pm 0.4384}}$	-	-0.0039 ± 0.1016 $\underline{\mathbf{0.1196 \pm 0.0139}}$	**1.06** ↑
	.vs CBM-HNMU ACC (% ↑)	Target	Ours (△)	Target	Ours (△)	Target	Ours (△)	Target	Ours (△)	Target	Ours (△)	**0.55** ↑
		1.08	0.92 (0.16 ↓)	0.31	**3.82 (3.51** ↑)	0.26	**1.81 (1.55** ↑)	2.64	0.24 (2.40 ↓)	0.34	**0.60 (0.26** ↑)	
		Base	Fine-Tune / HI-Gen	Base	Fine-Tune / HI-Gen	Base	Fine-Tune / HI-Gen	Base	Fine-Tune / HI-Gen	Base	Fine-Tune / HI-Gen	Avg ↑ (HI-Gen)
DaViT	ACC (%)	97.72	97.56 ± 0.21 $\underline{\mathbf{98.01 \pm 0.01}}$	74.74	77.14 ± 0.06 $\underline{\mathbf{78.52 \pm 0.13}}$	70.31	72.09 ± 0.05 $\underline{\mathbf{72.97 \pm 0.25}}$	97.39	97.61 ± 0.06 $\underline{\mathbf{97.68 \pm 0.01}}$	78.13	79.94 ± 0.07 $\underline{\mathbf{80.28 \pm 0.39}}$	**1.83** ↑
	$\mathcal{M}_{\text{cls}}(\triangle \uparrow)$	-	-0.1053 ± 0.2880 $\underline{\mathbf{0.2789 \pm 0.0094}}$	-	0.1762 ± 0.0024 $\underline{\mathbf{0.4033 \pm 0.0479}}$	-	0.2118 ± 0.0105 $\underline{\mathbf{0.3818 \pm 0.0379}}$	-	-1.1350 ± 0.9384 $\underline{\mathbf{0.2616 \pm 0.0176}}$	-	0.2815 ± 0.0137 $\underline{\mathbf{0.4187 \pm 0.0760}}$	**0.35** ↑

4.2 INTERVENTION GENERALIZATION PERFORMANCE

Using **Eq. 10**, we distill the corrections produced by HI on $\mathcal{D}^{\text{val}}$ into the original CBM, and we also fine-tune a comparison branch with $\mathcal{L}_{\text{cls}}$ only. **Table 2** reports the evaluation results on $\mathcal{D}^{\text{test}}$. On NFResNet-50, the distilled model (HI-Gen) improves test accuracy by up to 3.82% with an average gain of 1.48% over the baseline, and by up to 3.51% with an average gain of 0.55% relative to CBM-HNMU. On DaViT, the gains reach 3.78% at the best, with an average improvement of 1.83% over the baseline. The margin metric $\mathcal{M}_{\text{cls}}$ increases consistently after intervention. We also

compare the generalization of CI-Gen, PI-Gen, and HI-Gen (see **Appendix. C**). HI-Gen always achieves the best performance, showing that concept-aware pixel intervention balances effectiveness and generalizability better than pure pixel or pure concept intervention.

4.3 Ablation Study

We evaluate the effectiveness of hierarchical intervention (HI) by contrasting HI with concept-level intervention (CI), random concept and random patch interventions on NFResNet-50 across five datasets ($\mathcal{D}^{\text{val}}$), while varying the number of intervention concepts $\mathcal{N}$ and patches $\mathcal{P}$ (**Figure 2**).

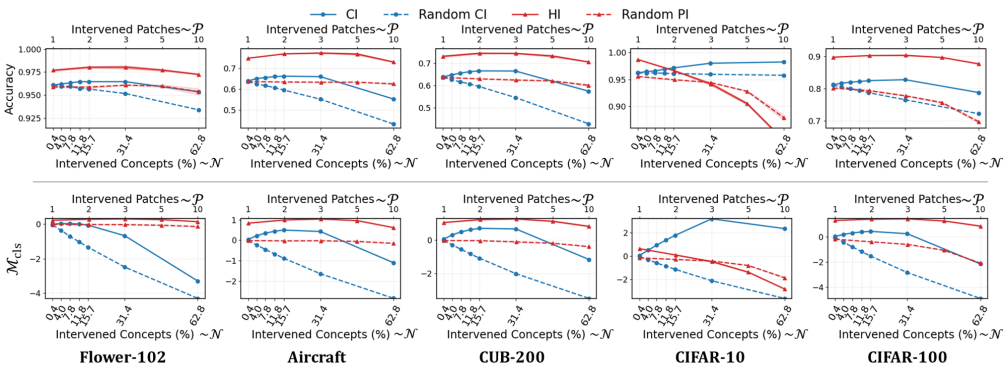


Figure 2: Ablation study on $\mathcal{N}$ and $\mathcal{P}$, with comparison against corresponding random interventions.

Table 3: Ablation on $\mathcal{P}$ for intervention generalization with hierarchical intervention (HI-Gen).

$\mathcal{P}=$	1	2	3	5	10
Flower-102					
Accuracy (%)	**97.18**	97.04	96.83	96.83	96.99
$\mathcal{M}_{\text{cls}}(\triangle\uparrow)$	**0.80**	0.13	0.12	0.15	0.17
Aircraft					
Accuracy (%)	**67.45**	66.34	66.46	66.52	66.37
$\mathcal{M}_{\text{cls}}(\triangle\uparrow)$	**0.74**	0.37	0.39	0.40	0.42
CUB-200					
Accuracy (%)	**65.27**	64.84	64.69	64.50	64.81
$\mathcal{M}_{\text{cls}}(\triangle\uparrow)$	**0.31**	0.19	0.21	0.21	0.20

Consequently, random interventions consistently reduce both accuracy and $\mathcal{M}_{\text{cls}}$, with the degradation worsening as $\mathcal{N}$ and $\mathcal{P}$ increase. In contrast, CI and HI remain more stable and yield positive gains. Moreover, HI (red solid line) consistently outperforms CI (blue solid line), confirming effectiveness of our hierarchical approach. The optimal proportion of negative intervention concepts typically falls in $(7.8\%\text{ to }31.4\%)\times n_{\text{cpt}}$, and the recommended number of pixel patch is $\mathcal{P}=3$ for most datasets on validation. Nevertheless, as for ablation in testing generalization (**Table 3**), we observe $\mathcal{P}=1$ yields better performance when using NFResNet-50 with $\mathcal{N}=7.8\%\times n_{\text{cpt}}$, which again points to the inherent limitations of the distillation process.

4.4 Reasoning and Intervention Explanations

Our method integrates causal effect with feature clustering to attribute image regions ($|m^{\text{cpt}}_{\text{causal}}(u,v)|>\theta_3$) that are sensitive to negative concepts to specific multi-modal explanations. Each concept cluster is visualized with distinct colors on the corresponding pixel patches, and semantically similar visual concepts are assigned using CLIP. **Figure 3** illustrates an example where the causal effect map for the true class prediction is divided into four color-coded regions. The multi-modal explanations of the green region shows strong association with the reproductive organ at the flower's center characterized by conical structure, seed head concept and orange-yellow attributes. In the causal effect map, patches in this region appear red, indicating that these patches act as negative cues for distinguishing the true class.

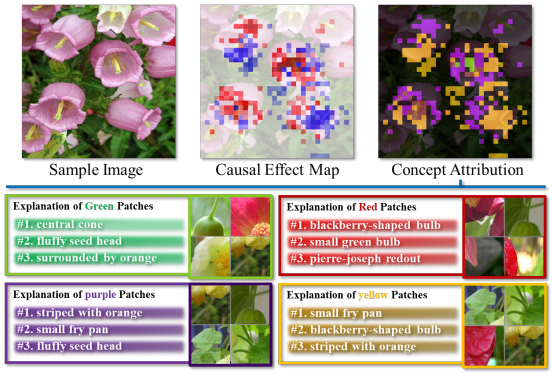


Figure 3: An example for reasoning explanation based on concept causal attribution ($\mathcal{C}=4$, $\mathcal{Q}=4$).

By intervening on pixels associated with the negative concepts, the corresponding concept scores are reduced, correcting model errors and providing a clear explanation of the intervention process. We observe that the intervention concepts are often correlated with both the ground-truth and the confused class, frequently sharing spurious visual features (see **Sec. 4.5** for quantification). For example, in the left case of **Figure 4**, the CBM misclassifies "Canterbury Bells" as "Mallow". Using the $\mathcal{M}_{\text{cls}}$, the top four negative concepts: "orange-yellow", "central cone", "vibrant yellow" and "striped with orange" are identified. Intervening on patches that are positively correlated with these concepts lowers their prediction scores, thereby steering the CBM toward the correct classification.

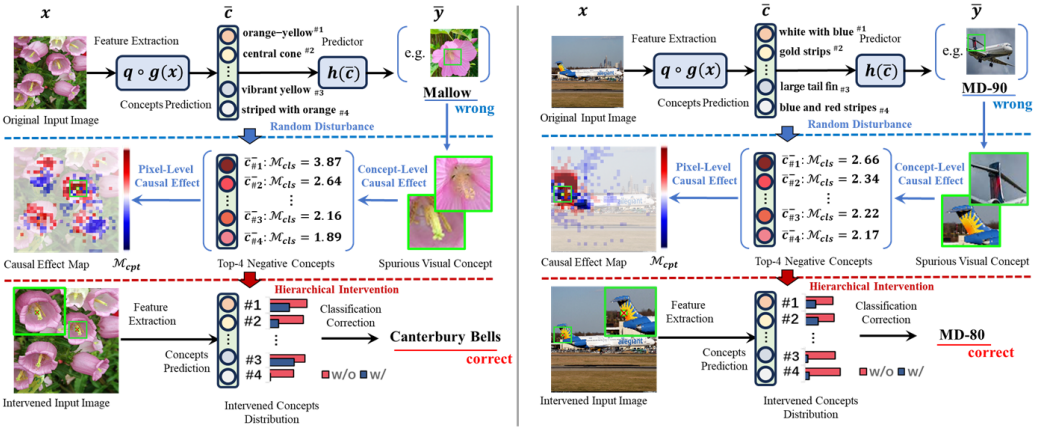


Figure 4: Illustration of the intervention explanation. For erroneous predictions, CBM-PCCEI leverages $\mathcal{M}_{\text{cls}}$ to identify negative concepts and associates them with image patches via $\mathcal{M}_{\text{cpt}}$. The method then intervenes on patches that are positively correlated with the negative concepts, thereby correcting the decision.

4.5 User Study

To quantify alignment with human perception, we conducted a sample-level user study following the CT–CF–CR protocol. For each corrected sample, scored every intervened concept as a true-class cue (CT), a false-class cue for the confused class (CF), or a shared cue relevant to both classes (CR). We recruited 30 non-expert participants and evaluated three fine-grained datasets. The interface and aggregate results appear in **Figure 5**.

For the three datasets we assessed "26", "365" and "181" corrected samples, respectively. Each case presented top 20 concepts of $\mathcal{I}_{\text{neg}}^{\text{cpt}}$, yielding "520", "7,300", and "3,620" judgments. The results show that the concepts selected by our method align with human visual perception of the confused classes and deliver a clear improvement over CBM-HNMU across the evaluated datasets.

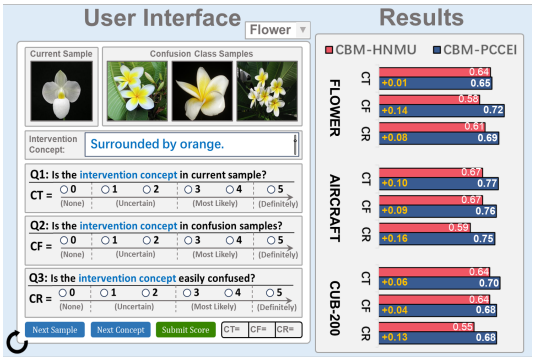


Figure 5: Overview of the user interface (left) and evaluation results (right) on NFResNet-50. The higher the CT and CF values, the more the visual cues of the intervention concept align with human perception within the corresponding class samples. CR further indicates whether the intervention concept is a key cues in cognitive confusion (higher value indicate greater consistency).

5 Conclusion

We propose an interpretable hierarchical intervention strategy for concept bottleneck models. This approach overcomes a key limitation of traditional CBM-based interventions by explicitly accounting for upstream feature bias, thereby significantly boosting intervention effectiveness while enhancing interpretability in a manner aligned with human visual perception. Furthermore, CBM PCCEI integrates causal effect map to deliver a stable and high performance extension of CBM interventions that requires no human priors and achieves efficient generalization.

AI USE STATEMENT

In this work, we only used generative AI tools to polish writing. We have not used generative AI tools for other tasks with required disclosure. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

For the intervention explanation, this work includes a user study, with the full set of instructions provided to participants available in **Appendix. E**. We confirm that the user study conducted in this work does not involve sensitive issues concerning human ethics, the human body, or psychological well-being.

REPRODUCIBILITY STATEMENT

We commit to releasing all code and data upon acceptance of the paper to facilitate reproducibility. The manuscript and appendix provide sufficient methodological and experimental details to enable the full replication of our methods and reproduction of all reported results. Please refer to **Sec. 3** for details of the proposed methods and to **Appendix. F.2** for hyperparameters and other experimental settings.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Neural Information Processing Systems*, 2018. URL https://api.semanticscholar.org/CorpusID:52938797.

Yong Hyun Ahn, Hyeon Bae Kim, and Seong Tae Kim. Www: A unified framework for explaining what, where and why of neural networks by interpretation of neuron concepts. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10968–10977, 2024. URL https://api.semanticscholar.org/CorpusID:268063605.

Kushal Chauhan, Rishabh Tiwari, Jana von Freyberg, Pradeep Shenoy, and Krishnamurthy Dj Dvijotham. Interactive concept bottleneck models. In *AAAI Conference on Artificial Intelligence*, 2022. URL https://api.semanticscholar.org/CorpusID:254685619.

Preetam Prabhu Srikar Dammu and Chirag Shah. Detecting spurious correlations via robust visual concepts in real and ai-generated image classification. *ArXiv*, abs/2311.01655, 2023. URL https://api.semanticscholar.org/CorpusID:265019367.

Mingyu Ding, Bin Xiao, Noel C. F. Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *European Conference on Computer Vision*, 2022. URL https://api.semanticscholar.org/CorpusID:248006188.

Gabriele Dominici, Pietro Barbiero, Francesco Giannini, Martin Gjoreski, Giuseppe Marra, and Marc Langheinrich. Counterfactual concept bottleneck models. In *International Conference on Learning Representations*, 2024a. URL https://api.semanticscholar.org/CorpusID:267406323.

Gabriele Dominici, Pietro Barbiero, Mateo Espinosa Zarlenga, Alberto Termine, Martin Gjoreski, Giuseppe Marra, and Marc Langheinrich. Causal concept graph models: Beyond causal opacity in deep learning. In *International Conference on Learning Representations*, 2024b. URL https://api.semanticscholar.org/CorpusID:270062750.

Amir Feder, Katherine A. Keith, Emaad A. Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M Stewart, Victor Veitch, and Diyi Yang. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 2021. URL https://api.semanticscholar.org/CorpusID:237386009.

Thomas Fel, Agustin Picard, Louis Béthune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2711–2721, 2022. URL https://api.semanticscholar.org/CorpusID:253708233.

Yun Ping Ge, Yao Xiao, Zhi Xu, Meng Zheng, Srikrishna Karanam, Terrence Chen, Laurent Itti, and Ziyan Wu. A peek into the reasoning of neural networks: Interpreting with structural visual concepts. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2195–2204, 2021. URL https://api.semanticscholar.org/CorpusID:233481581.

Amirata Ghorbani, Abubakar Abid, and James Y. Zou. Interpretation of neural networks is fragile. In *AAAI Conference on Artificial Intelligence*, 2017. URL https://api.semanticscholar.org/CorpusID:22172746.

Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. Towards automatic concept-based explanations. In *Neural Information Processing Systems*, 2019. URL https://api.semanticscholar.org/CorpusID:184487319.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. *ArXiv*, abs/1904.07451, 2019. URL https://api.semanticscholar.org/CorpusID:119309561.

Jinfan Hu, Jinjin Gu, Shiyao Yu, Fanghua Yu, Zheyuan Li, Zhiyuan You, Chaochao Lu, and Chao Dong. Interpreting low-level vision models with causal effect maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47:6396–6409, 2024. URL https://api.semanticscholar.org/CorpusID:271533796.

Vidhya Kamakshi, Uday Gupta, and N. C. Krishnan. Pace: Posthoc architecture-agnostic concept extractor for explaining cnns. *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2021. URL https://api.semanticscholar.org/CorpusID:237363956.

Amir-Hossein Karimi, Krikamol Muandet, Simon Kornblith, Bernhard Scholkopf, and Been Kim. On the relationship between explanation and prediction: A causal view. *ArXiv*, abs/2212.06925, 2022. URL https://api.semanticscholar.org/CorpusID:254636466.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. *ArXiv*, abs/2007.04612, 2020. URL https://api.semanticscholar.org/CorpusID:220424448.

Matthew Kowal, Richard P. Wildes, and Konstantinos G. Derpanis. Visual concept connectome (vcc): Open world concept discovery and their interlayer connections in deep models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10895–10905, 2024. URL https://api.semanticscholar.org/CorpusID:268876088.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, X. Renard, and Marcin Detyniecki. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. In *International Joint Conference on Artificial Intelligence*, 2019. URL https://api.semanticscholar.org/CorpusID:198147903.

Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Scholkopf. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2022. URL https://api.semanticscholar.org/CorpusID:251647196.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *ArXiv*, abs/1306.5151, 2013. URL https://api.semanticscholar.org/CorpusID:2118703.

Ricards Marcinkevics, Sonia Laguna, Moritz Vandenhirtz, and Julia E. Vogt. Beyond concept bottleneck models: How to make black boxes intervenable? *ArXiv*, abs/2401.13544, 2024. URL https://api.semanticscholar.org/CorpusID:267199978.

J. Mcqueen. Some methods for classi cation and analysis of multivariate observations. 1967. URL https://api.semanticscholar.org/CorpusID:235144755.

Tanish Mittal, Preyansh Agrawal, Esha Pahwa, and Aarya Makwana. Nfresnet: Multi-scale and u-shaped networks for deblurring. *ArXiv*, abs/2212.05909, 2022. URL https://api.semanticscholar.org/CorpusID:254563859.

Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2019. URL https://api.semanticscholar.org/CorpusID:159041688.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729, 2008. URL https://api.semanticscholar.org/CorpusID:15193013.

Chengyao Qian, Trung Le, and Mehrtash Harandi. A good teacher adapts their knowledge for distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1239–1248, October 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206 – 215, 2018. URL https://api.semanticscholar.org/CorpusID:182656421.

Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336 – 359, 2016. URL https://api.semanticscholar.org/CorpusID:15019293.

Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. A closer look at the intervention procedure of concept bottleneck models. *ArXiv*, abs/2302.14260, 2023. URL https://api.semanticscholar.org/CorpusID:257233107.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013. URL https://api.semanticscholar.org/CorpusID:1450294.

David Steinmann, Wolfgang Stammer, Felix Friedrich, and Kristian Kersting. Learning to intervene on concept bottlenecks. *ArXiv*, abs/2308.13453, 2023. URL https://api.semanticscholar.org/CorpusID:261214749.

Radu Timofte, Shuhang Gu, Jiqing Wu, and Luc Van Gool. Ntire 2018 challenge on single image super-resolution: Methods and results. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 965–96511, 2018. URL https://api.semanticscholar.org/CorpusID:52034468.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. URL https://api.semanticscholar.org/CorpusID:16119123.

Nuoye Xiong, Anqi Dong, Ning Wang, Cong Hua, Guangming Zhu, Lin Mei, Peiyi Shen, and Liang Zhang. Intervening in black box: Concept bottleneck model for enhancing human neural network mutual understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2836–2845, October 2025.

Xin-Chao Xu, Yi Qin, Lu Mi, Hao Wang, and Xiaomeng Li. Energy-based concept bottleneck models: Unifying prediction, concept intervention, and probabilistic interpretations. In *International Conference on Learning Representations*, 2024. URL https://api.semanticscholar.org/CorpusID:268031821.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19187–19197, 2022. URL https://api.semanticscholar.org/CorpusID:253735286.

Mert Yuksekgonul, Maggie Wang, and James Y. Zou. Post-hoc concept bottleneck models. *ArXiv*, abs/2205.15480, 2022. URL https://api.semanticscholar.org/CorpusID:249209990.

Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A. Ehinger, and Benjamin I. P. Rubinstein. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In *AAAI Conference on Artificial Intelligence*, 2020. URL https://api.semanticscholar.org/CorpusID:235349029.

Xianren Zhang, Dongwon Lee, and Suhang Wang. Comprehensive attribution: Inherently explainable vision model with feature detector. In *European Conference on Computer Vision*, 2024. URL https://api.semanticscholar.org/CorpusID:271533378.

Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *European Conference on Computer Vision*, 2018. URL https://api.semanticscholar.org/CorpusID:51903952.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. In *The Thirteenth International Conference on Learning Representations*, 2025.

APPENDIX

Our supplementary material for the manuscript can be summarizes as follows.

A. Comparison with LCP, CCTP and ECTP
B. Comparison with Post-hoc CBM
C. Comparison between CI-Gen, PI-Gen, and HI-Gen
D. Concept Constraint
E. User Study Details
F. Detailed Experimental Setup
    F.1 Concept Bottleneck Settings
    F.2 Hyper-parameters Settings
    F.3 Hardware Requirement
    F.4 Runtime and Scalability
G. Limitations
H. Discussions
I. More Visualizations of Reasoning and Intervention
J. Hierarchical Causal Effect Intervention Architecture Details
K. DIV2K-related Ablation
L. Statistics of Error Correction and Misjudgment
M. Discussion and Reflection on Performance Limitations in Intervention Generalization
N. Ablation Study on CLIP
O. Ablation with CBM-HNMU

Section **A** compares the intervention benefits of HI and LCP, and CCTP and ECTP on the validation set.

Section **B** further compares CBM-PCCEI, CBM and Post-hoc CBM (PCBM) based on generalization accuracy and intervention gain, further demonstrating the effectiveness and superiority of our method.

Section **C** compares the classification performance of CI, PI, and HI after intervention generalization.

Section **D** demonstrates the changes of negative and non-negative concepts before and after the hierarchical intervention (HI) to show that our approach can explainably weaken harmful concepts while keeping the others unchanged.

Section **E** lists some use cases of user study, as well as detailed settings and scoring criteria.

Section **F**, **G** and **H** provide more experimental details and more discussion about our work.

Section **I** integrates reasoning and intervention explanations and provides more visual examples of CBM-PCCEI.

Section **J** demonstrates the architectural details of the hierarchical causal intervention approach.

Section **K** uses grayscale masks to dissolve the natural image patches provided by DIV2K, ensuring that DIV2K does not help the classifier cheat.

Section **L** presents statistics on the number of samples corrected by CBM-PCCEI intervention on different datasets and architectures, as well as examples of cases where intervention actually led to prediction errors.

Section **M** introduces the bottleneck of performance gain that arise during intervention generalization, compares distillation with other generalization methods (CB2M-like), and presents experiments on optimizing performance bottlenecks using Teacher Adapter.

Sections **N** and **O** provide further analysis of the contributions of CBM-PCCEI and its components, as well as the effectiveness of the Hierarchical Intervention (HI) framework.

A COMPARISON WITH LCP, CCTP AND ECTP

To further compare the performance of our method with other concept intervention strategies, we evaluate HI and the following intervention strategies using an NFResNet-50 backbone-based CBM as a baseline

- LCP: It selects concepts with the largest loss on concept prediction compared to the ground-truth (the concept prediction scores obtained from CLIP are used as pseudo-labels).
- CCTP: It selects concepts with the highest contribution on target prediction;
- ECTP: It selects concepts with the highest expected change in the target predictive distribution with respect to intervention.

The results are shown in **Table I**. The metrics of the three concept intervention strategies compared here are the maximum values in ablation with a concept intervention number of $\mathcal{N} = \{\}\% * n_{\text{cpt}}$. The results show that, across all five datasets, HI consistently achieves the greatest intervention benefit and classification performance compared to the other three intervention strategies (Shin et al., 2023).

Table I: Comparative evaluation of intervention strategies on the validation set. Accuracy reports classification performance, while $\mathcal{M}_{\text{cls}}$ measures the separation between the true-class and most-confusing-class logits. ↑ denotes improvement over the baseline.

Flower-102					
Metric	**Base**	**LCP**	**CCTP**	**ECTP**	**CBM-PCCEI (HI)**
Accuracy	96.08	96.08 (-)	96.33 (↑ 0.25)	96.08 (-)	**97.89 (↑ 1.61)**
$\mathcal{M}_{\text{cls}}$	–	−0.0949	0.0459	0.0044	**0.2314**
Aircraft					
Accuracy	63.72	63.85 (↑ 0.13)	65.59 (↑ 1.87)	63.72 (-)	**74.73 (↑ 11.01)**
$\mathcal{M}_{\text{cls}}$	–	−0.0232	0.3609	0.0589	**0.8164**
CUB-200					
Accuracy	63.60	63.75 (↑ 0.15)	65.55 (↑ 1.95)	63.60 (-)	**73.13 (↑ 9.53)**
$\mathcal{M}_{\text{cls}}$	–	−0.0141	0.5077	0.0621	**1.0481**
CIFAR-10					
Accuracy	96.20	96.22 (↑ 0.02)	96.64 (↑ 0.44)	97.62 (↑ 1.42)	**98.65 (↑ 2.45)**
$\mathcal{M}_{\text{cls}}$	–	−0.1895	1.0510	**2.4273**	0.6267
CIFAR-100					
Accuracy	81.14	81.14 (-)	82.04 (↑ 0.90)	81.14 (-)	**89.64 (↑ 8.50)**
$\mathcal{M}_{\text{cls}}$	–	−0.0755	0.3403	0.0480	**1.3465**

B COMPARISON WITH POST-HOC CBM

Traditional CBMs relies on concept annotations and human priors (such as inherent data bias) for intervention generalization. However, image classification dataset generally does not necessarily contain concept labels, and data bias varies for models with different architectures and even different parameter sizes. In this section, we train Post-hoc CBM (PCBM) using the pseudo-label dynamic generation and model alignment strategy based on CLIP text-image linking capabilities used in our method and compare the CBM, PCBM, and CBM-PCCEI in terms of baseline performance and intervention benefits. All intervention comparisons used NFResNet-50 as the backbone, and the results are shown in **Table II**, **III** and **IV**.

Table II shows the global intervention gain and corresponding baseline metrics for each method. It can be seen that the hierarchical intervention (HI) of CBM-PCCEI consistently achieves the largest intervention gain, while the gains of CBM and PCBM based on concept label intervention are very weak and unstable (even resulting in ACC or $\mathcal{M}_{\text{cls}}$ loss). Looking back at the concept-level intervention (CI) based on causal effects in the manuscript, its intervention gain is also generally higher than that of CBM (INT) and PCBM (INT). These results further demonstrate the effectiveness and superiority of our method in selecting pixels/concepts based on causal effects.

Table III and **IV** show the ablation results of CBM (INT) and PCBM (INT) based on concept label intervention. PCBM, by incorporating residual connections of upstream features, is more tolerant of large-scale concept interventions, while traditional CBM suffers a precipitous performance drop when too many concepts are intervened due to excessive feature drift.

Table II: The gains of CBM and PCBM after intervention (INT) using pseudo-labels generated by CLIP based on image concept matching compared to the corresponding baselines are listed (ACC ↑ & $\mathcal{M}_{\text{cls}}$). Note that the accuracy gain and $\mathcal{M}_{\text{cls}}$ for both CBM and PCBM are the maximum values during ablations (see Tab. III and IV), and do not necessarily mean that they can be obtained simultaneously.

Split	**Flower-102** (Accuracy % / $\mathcal{M}_{\text{cls}}$ ↑)				
	CBM (Base)	**CBM (INT)**	**PCBM (Base)**	**PCBM (INT)**	**CBM-PCCEI (HI)**
Val	96.08/−	96.08(− ↑)/ − 0.095	95.96/−	96.20(0.24 ↑)/ − 0.042	**97.69(1.61 ↑)/0.231**
Test	96.26/−	96.35(0.09 ↑)/ − 0.096	96.43/−	96.51(0.08 ↑)/ − 0.041	**97.18(0.92 ↑)/0.8048**
	Aircraft (Accuracy % / $\mathcal{M}_{\text{cls}}$ ↑)				
Val	63.72/−	63.85(0.13 ↑)/ − 0.023	70.33/−	70.36(0.03 ↑)/ − 0.032	**74.73(11.01 ↑)/0.816**
Test	63.63/−	63.78(0.15 ↑)/ − 0.021	70.93/−	71.02(0.09 ↑)/ − 0.034	67.45 **(3.82 ↑)/0.7441**
	CUB-200 (Accuracy % / $\mathcal{M}_{\text{cls}}$ ↑)				
Val	63.60/−	63.75(0.15 ↑)/ − 0.014	70.60/−	70.60(− ↑)/ − 0.016	**73.13(9.53 ↑)/1.048**
Test	63.46/−	63.43(− ↑)/ − 0.015	70.37/−	70.42(0.05 ↑)/ − 0.017	65.27 **(1.81 ↑)/0.3068**
	CIFAR-10 (Accuracy % / $\mathcal{M}_{\text{cls}}$ ↑)				
Val	96.20/−	96.22(0.02 ↑)/ − 0.190	95.56/−	95.56(− ↑)/ − 0.001	**98.65(2.45 ↑)/0.627**
Test	95.42/−	95.45(0.03 ↑)/ − 0.187	94.89/−	94.89(− ↑)/ − 0.001	**95.66(0.24 ↑)/3.3282**
	CIFAR-100 (Accuracy % / $\mathcal{M}_{\text{cls}}$ ↑)				
Val	81.14/−	81.14(− ↑)/ − 0.076	82.30/−	82.46(0.16 ↑)/ − 0.040	**89.64(8.50 ↑)/1.347**
Test	81.33/−	81.38(0.05 ↑)/ − 0.076	82.32/−	82.36(0.04 ↑)/ − 0.039	81.93 **(0.60 ↑)/0.1196**

Table III: Ablation of CBM (INT). Bold data represents the optimal metric for classification accuracy and $\mathcal{M}_{\text{cls}}$ in corresponding branch (Val or Test). A negative value of $\mathcal{M}_{\text{cls}}$ indicates that the discrimination of the true class decreased after the intervention.

CBM (INT)	Val							Test						
Metrics	**Flower-102**													
$\mathcal{N} = \{\}\% * n_{\text{cpt}}$	0.4	4.0	7.8	11.8	15.7	31.4	62.8	0.4	4.0	7.8	11.8	15.7	31.4	62.8
Accuracy	**96.08**	96.02	96.02	95.96	89.65	2.39	2.39	96.26	**96.35**	96.26	96.14	89.57	2.35	2.35
$\mathcal{M}_{\text{cls}}$	**−0.095**	−1.170	−2.583	−4.157	−5.730	−11.02	−14.40	**−0.096**	−1.182	−2.610	−4.202	−5.787	−11.09	−14.47
Aircraft														
Accuracy	63.69	**63.85**	63.79	63.33	55.78	2.46	2.85	63.67	**63.78**	63.58	63.07	55.03	2.28	2.85
$\mathcal{M}_{\text{cls}}$	**−0.023**	−0.271	−0.619	−1.101	−1.687	−4.489	−6.553	**−0.021**	−0.260	−0.600	−1.067	−1.637	−4.432	−6.508
CUB-200														
Accuracy	63.65	**63.75**	63.25	62.55	59.80	1.20	1.60	**63.43**	63.38	63.39	63.03	59.73	1.10	1.45
$\mathcal{M}_{\text{cls}}$	**−0.014**	−0.266	−6.32	−1.110	−1.764	−5.137	−7.384	**−0.015**	−0.255	−0.610	−1.092	−1.756	−5.163	−7.429
CIFAR-10														
Accuracy	96.20	96.20	96.20	96.20	**96.22**	96.20	96.16	95.43	95.44	**95.45**	95.45	95.45	95.45	95.41
$\mathcal{M}_{\text{cls}}$	**−0.190**	−1.572	−2.703	−3.658	−4.432	−6.381	−7.434	**−0.187**	−1.550	−2.664	−3.609	−4.374	−6.294	−7.331
CIFAR-100														
Accuracy	81.10	**81.14**	81.14	80.84	82.24	3.50	3.28	81.34	**81.38**	81.25	81.10	80.50	3.43	3.33
$\mathcal{M}_{\text{cls}}$	**−0.076**	−0.985	−2.099	−3.340	−4.592	−8.700	−11.53	**−0.076**	−0.999	−2.127	−3.378	−4.644	−9.065	−11.60

Table IV: Ablation of PCBM (INT). Bold data represents the optimal metric for classification accuracy and $\mathcal{M}_{\text{cls}}$ in corresponding branch (Val or Test). A negative value of $\mathcal{M}_{\text{cls}}$ indicates that the discrimination of the true class decreased after the intervention.

PCBM (INT)	Val							Test						
Metrics	**Flower-102**													
$\mathcal{N} = \{\}\% * n_{\text{cpt}}$	0.4	4.0	7.8	11.8	15.7	31.4	62.8	0.4	4.0	7.8	11.8	15.7	31.4	62.8
Accuracy	95.96	95.96	95.96	96.02	96.02	96.08	**96.20**	96.43	96.43	96.43	96.43	**96.51**	96.39	96.35
$\mathcal{M}_{\text{cls}}$	**−0.042**	−0.495	−0.997	−1.531	−2.053	−3.689	−4.531	**−0.041**	−0.490	−0.990	−1.528	−2.054	−3.703	−4.550
Aircraft														
Accuracy	70.33	70.27	70.27	**70.36**	70.36	70.33	70.27	70.93	70.93	**71.02**	70.96	70.96	70.99	71.02
$\mathcal{M}_{\text{cls}}$	**−0.032**	−0.267	−0.461	−0.632	−0.775	−1.135	−1.277	**−0.034**	−0.274	−0.473	−0.647	−0.793	−1.161	−1.306
CUB-200														
Accuracy	**70.60**	70.55	70.35	70.45	70.60	70.30	70.25	70.37	**70.42**	70.37	70.31	70.30	70.23	69.87
$\mathcal{M}_{\text{cls}}$	**−0.016**	−0.200	−0.409	−0.635	−0.864	−1.689	−2.340	**−0.017**	−0.212	−0.432	−0.665	−0.900	−1.742	−2.384
CIFAR-10														
Accuracy	**95.56**	95.56	95.56	95.56	95.56	95.56	95.56	**94.89**	94.88	94.87	94.87	94.87	94.87	94.87
$\mathcal{M}_{\text{cls}}$	**−0.001**	−0.002	−0.004	−0.004	−0.005	−0.005	−0.007	**−0.001**	−0.002	−0.004	−0.005	−0.006	−0.007	−0.007
CIFAR-100														
Accuracy	82.28	82.30	82.32	82.36	**82.46**	82.30	81.98	82.33	82.34	**82.36**	82.33	82.36	82.24	82.08
$\mathcal{M}_{\text{cls}}$	**−0.040**	−0.460	−0.915	−1.372	−1.792	−2.983	−3.603	**−0.039**	−0.459	−0.914	−1.372	−1.795	−2.987	−3.608

C COMPARISON BETWEEN CI-GEN, PI-GEN, AND HI-GEN

During the validation phase, a comparison of CI, PI, and HI revealed that PI had the highest intervention benefit. However, we believe that skipping targeted interventions at the concept-level not only results in a loss of interpretability but also lead to significant biases in conceptual predictions, thereby limiting the generalization ability of the refined concepts. Conversely, HI, which focuses on negative concepts while preserving non-negative concepts, might achieve better results. Therefore, using NFResNet-50 as the backbone, we compared the generalization performance of CI-Gen, PI-Gen, and HI-Gen on the test set. The results, shown in **Table V**, confirm our hypothesis that HI-Gen generally achieved the best performance metrics across all five datasets.

Table V: Comparative evaluation of intervention generation models on test sets.

Flower-102				
Metric	**Base**	**CI-Gen**	**PI-Gen**	**HI-Gen**
Accuracy	96.26	96.35	96.75	**97.18**
$\mathcal{M}_{\text{cls}}$	–	0.0290	0.1672	**0.8048**
Aircraft				
Accuracy	63.63	64.54	66.01	**67.45**
$\mathcal{M}_{\text{cls}}$	–	0.1770	0.4074	**0.7441**
CUB-200				
Accuracy	63.46	64.25	64.72	**65.27**
$\mathcal{M}_{\text{cls}}$	–	0.1038	0.2355	**0.3068**
CIFAR-10				
Accuracy	95.42	**95.67**	95.59	95.66
$\mathcal{M}_{\text{cls}}$	–	1.1294	0.2617	**3.3282**
CIFAR-100				
Accuracy	81.33	81.82	81.85	**81.93**
$\mathcal{M}_{\text{cls}}$	–	0.0822	**0.1801**	0.1196

D CONCEPT CONSTRAINTS

In manuscript, we set the benefit measurement function for concept-level causal effects to weaken negative concepts while keeping other concepts unchanged, thereby obtaining a stable and interpretable intervention gain. Herein, we evaluate the changes in negative and non-negative concepts, further demonstrating that concept pruning in CBM-PCCEI is controllable.

Due to uncertainty from pixel-level interventions, we repeat the intervention multiple times and report the resulting concept-level shift using the average value. The shift is quantified by computing the cosine similarity between the vector of each kind of concept prediction score before and after the intervention, and then subtracting the value from 1 to reflect the degree of deviation, i.e.,

$$\mathrm{Shift}(c^-) = 1 - \cos\left(\bar{c}^-, \hat{c}^-\right), \quad \mathrm{Shift}(c^+) = 1 - \cos\left(\bar{c}^+, \hat{c}^+\right),$$

where c^- and c^+ represent the negative concepts and its complement, respectively. We further define the $\mathrm{Shift\ Gap}$ between the two and the $\mathrm{Shift\ Ratio}$ of negative concept shift in the total shift as more intuitive metrics.

We use NFResNet-50 as the backbone to perform concept shift evaluation based on a hierarchical intervention (HI) on five datasets: Flower-102, Aircraft, CUB-200, CIFAR-10 and CIFAR-100. We listed the changes in negative concepts, the changes in other concepts that are not negative concepts, and the difference between the changes in the two in **Table VI**. In the results, $\mathrm{Shift}(c^-)$ is usually much larger than $\mathrm{Shift}(c^+)$ leading positive $\mathrm{Shift\ Gap}$ and high $\mathrm{Shift\ Ratio}$, which indicates that pixel-driven concept pruning in HI focuses on negative concepts and ensures that the predictions of other concepts are relatively stable. It is worth noting that the exception to CIFAR-10 is predictable. Because the concepts bottleneck of CIFAR-10 is too small (see **Sec. F.1**), the reasoning relationships between concepts and classes may be denser, making it more difficult to control concept changes caused by intervening pixel patches.

Table VI: Comparative assessment of concept shift before and after hierarchical intervention (HI).

Flower-102	$\mathbf{Shift}(c^-)$	$\mathbf{Shift}(c^+)$	**Shift Gap**	**Shift Ratio** %
	0.01923	0.00044	0.01879	97.76
Aircraft	$\mathbf{Shift}(c^-)$	$\mathbf{Shift}(c^+)$	**Shift Gap**	**Shift Ratio** %
	0.35953	0.07404	0.2855	82.92
CUB-200	$\mathbf{Shift}(c^-)$	$\mathbf{Shift}(c^+)$	**Shift Gap**	**Shift Ratio** %
	0.55289	0.16399	0.38890	77.12
CIFAR-10	$\mathbf{Shift}(c^-)$	$\mathbf{Shift}(c^+)$	**Shift Gap**	**Shift Ratio** %
	0.04381	0.03764	0.00616	53.79
CIFAR-100	$\mathbf{Shift}(c^-)$	$\mathbf{Shift}(c^+)$	**Shift Gap**	**Shift Ratio** %
	0.00067	0.00003	0.00064	95.71

E User Study Details

For the user study, we provide each recruited non-expert with a 30-minute experiment explanation and metric (CT-CF-CR) introduction, and set up a pre-test. Formal evaluation is only conducted after each person fully understood and passed the pre-test. Detailed evaluation information for three fine-grained visual classification datasets is shown in **Table VIIa**, and detailed records of the user study process are shown in **Table VIIb**. The data details from the user study is shown in **Table VIII**.

Here, we provide some user test cases and explain in detail some scoring rules and their meanings in **Figure I**. *In use case* $\#414$ *of the Flower-102 dataset*, the score for the concept "surrounded by orange" largely conforms to the metric definition. It is worth noting that CLIP has limitations in its ability to model concepts, so the semantic meaning of "surrounded" can be split and ignored or its weight reduced in the scoring. We emphasize this to users and encourage them to focus more on word-level semantics rather than the entire sentence to ensure that all attributes of the concept can be identified and scored. *In use case* $\#284$ *of the Aircraft dataset*, the concept of being unable to associate visual features is proposed, and users are instructed to skip this concept; our overall evaluation will also ignore these concepts. *In use case* $\#446$ *of the CUB-200 dataset*, the visual concepts in the background should also be taken into consideration, because we cannot ignore the shortcuts obtained in model data-driven approaches (such as background dependence).

Table VII: Statistical data from user study.

(a) Datasets' details. The use cases are selected from all validation samples that are corrected after intervention in each dataset, and all of these use cases will appear in the evaluation. CIFAR-10 and CIFAR-100 are excluded from the user study because their image resolution is too low, making it difficult for users to identify objects, let alone concepts.

Entry / Dataset	Flower-102	Aircraft	CUB-200
Samples Number	26	365	181
Concepts Number	520	$7,300$	$3,620$

(b) Users' details. Repeat Times is the average number of repetitions (until successful) of all 30 recruits for each dataset's pre-test; Time Consuming is the average completion time of all participants for the formal evaluation on each dataset.

Entry / Dataset	Flower-102	Aircraft	CUB-200
Time Consuming (min)	47.65	45.00	46.32
Repeat Times	2.00	1.47	1.82

Table VIII: Detailed results of user study.

Flower-102			
Metric	**CT**	**CF**	**CR**
CBM-HNMU	0.6365 ± 0.1241	0.5831 ± 0.0896	0.6111 ± 0.1432
CBM-PCCEI	$\mathbf{0.6544 \pm 0.1404}$	$\mathbf{0.7191 \pm 0.1163}$	$\mathbf{0.6885 \pm 0.1071}$
Aircraft			
CBM-HNMU	0.6712 ± 0.0568	0.6687 ± 0.0577	0.5902 ± 0.1618
CBM-PCCEI	$\mathbf{0.7723 \pm 0.1488}$	$\mathbf{0.7636 \pm 0.1457}$	$\mathbf{0.7496 \pm 0.1361}$
CUB-200			
CBM-HNMU	0.6421 ± 0.0806	0.6406 ± 0.0647	0.5518 ± 0.1755
CBM-PCCEI	$\mathbf{0.7048 \pm 0.1893}$	$\mathbf{0.6775 \pm 0.1509}$	$\mathbf{0.6793 \pm 0.1532}$

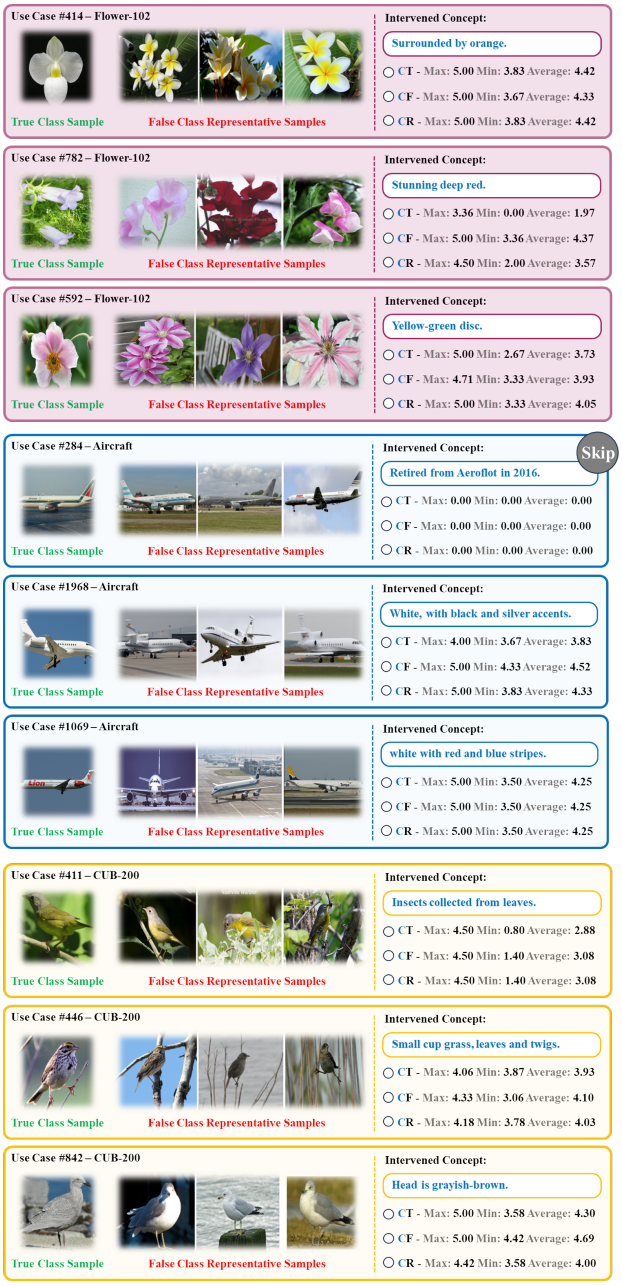


Figure I: Test cases and scores from some user study are shown here. The use case number only represents the index of the true class sample and is not related to the number of samples being evaluated.

User Study Guidelines and Participant Instructions

Overview and Recruitment

This user study evaluated the quality of "intervention concepts" (harmful or confusing concepts) identified by our method in three fine-grained visual classification datasets (Flower-102, Aircraft, and CUB-200), i.e., whether they are consistent with human perception.

Prerequisites: Each participant will receive 30 to 45 minutes of instruction covering the experimental objectives, the definition of CT-CF-CR indicators, and scoring criteria.

Qualification: Participants must pass a pre-test to demonstrate their full understanding before formal evaluation can begin.

Evaluation Procedure

The evaluation is conducted via a local web service deployed over the network. During the quiz, the website will only collect users' ratings and personal IDs for the purpose of calculating and distributing compensation.

Step-by-Step Instructions

1. **Login:** Enter a unique username (initials or nickname) to track individual progress.
2. **Dataset Selection:** Select one of the three datasets from the dropdown menu.
3. **Sample Assessment:** For each sample, the interface displays:
   - **Current Sample:** The target image being evaluated.
   - **Confusion Class Samples:** Some representative images from the class likely to be confused with the current sample.
   - **Intervention Concept:** The specific intervention concept discovered by the model.
4. **Scoring:** Users answer three questions (Q1-Q3) for 20 concepts per sample.
5. **Completion:** Data is saved only after all samples in a dataset are evaluated. A pop-up notification confirms the completion of a dataset.

Scoring Metrics and Rules

Participants rated the intervention concept on a scale of 0 to 5 (integers) according to the following three criteria:

Core Questions

Q1 (CT): Is the intervention concept in current sample?

Q2 (CF): Is the intervention concept in confusion samples?
(Appears in at least one of the listed confused samples)

Q3 (CR): Is the concept easily confused?
(Between the current sample's corresponding class and the confused sample's corresponding class)

Scoring Scale

- **0:** Completely inconsistent.
- ≤ 2**:** Uncertain / Weakly consistent.

- ≤ 4**:** Mostly consistent.
- **5:** Completely consistent.

SPECIAL HANDLING RULES

1. **Abstract Concepts:** If a concept cannot be associated with visual features (e.g., *"delivery of 156 aircraft by 2017"*), users are instructed to skip it using the "Next Concept" button. These are ignored in the final evaluation.
2. **Simple Association:** We encourage users to rate abstract concepts that have practical significance, such as *"a fiery red sunset"*. This can be represented by the concrete concept of the color red through simple association.
3. **Word-Level Semantics:** Users are encouraged to focus on the meaning at the word level, rather than the strict grammar at the sentence level. For example, if a concept is *"surrounded by orange"*, even if the spatial logic of *"surrounded"* is not clear, users should rate it according to the strength of the presence of the *"orange"* attribute.
4. **Background Context:** Visual concepts in the background (e.g., specific environments) must be scored, as the model may rely on these as shortcuts for classification. For example, concepts such as *"bushes"* and *"branches"* often appear together with two different bird species, which may be a clue that makes it easy for the model to confuse the two when generalizing.

F EXPERIMENTAL SETUP DETAILS

F.1 CONCEPT BOTTLENECK SETTINGS

We adopt the same concept bottlenecks based on prompt templates automatically generated by ChatGPT-3.5 for all five datasets as for LaBo and CBM-HNMU, and performed statistical analysis on the natural language concept set for each dataset, as detailed in **Table IX**.

Table IX: Concept count statistics: The Abstract column represents the number of abstract concepts, the Concrete column represents the number of visually relevant concepts, and the Total column represents the sum of all concepts in the concept bottleneck.

Dataset / Entry	Total	Abstract	Concrete (Color; Shape; Texture; Others)
Flower-102	$2,550$	$1,358$	$1,192\,(648; 292; 64; 188)$
Aircraft	$5,100$	$2,919$	$2,181\,(1,653; 281; 70; 177)$
CUB-200	$10,000$	$2,115$	$7,885\,(5,480; 384; 386; 1635)$
CIFAR-10	500	214	$286\,(58; 131; 63; 34)$
CIFAR-100	$5,000$	$1,609$	$3,391\,(956; 1,478; 522; 435)$

F.2 HYPER-PARAMETERS SETTINGS

- Training phase: The image resolution in the dataset is uniformly scaled to 256×256. The learning rate is uniformly set at $1e^{-4}$ with a decay weights 0.99. The training epochs are uniformly set to 100, and the process is stopped after 10 epochs without any optimizations. The training batch size is uniformly set to 32.
- Causal Effect Calculation: The number of intervention repetitions in the coarse-grained stage, n_{coarse}, is uniformly set to 3 and while the number (n_{fine}) of intervention repetitions in the fine-grained stage was set to 50. The threshold θ_1 is set to retain only the concept of 0.02 for refined calculation of concept-level causal effects, while θ_2 is set to retain only patches of 0.05 for refined calculation of pixel-level causal effects. When performing causal attribution, θ_3 is set to retain 75% of the patch to filter out causal insensitive regions.
- Distillation phase: The learning rate is uniformly set at $1e^{-6}$ with a decay weights 0.99. The fine-tuning epochs are uniformly set to 100, and the process is stopped after 10 epochs without any optimizations. The loss weighting coefficients α, β, and γ for the phase are all set to 1. We conducted a grid search over various loss-weight configurations and found that tuning the loss weights yielded negligible performance gains, with no consistent improvement across datasets.
- Unified Miscellaneous: All training processes used Adam as the optimizer, and the partitioning of training, validation, and test data remained faithful to the original dataset; we did not make any additional adjustments.

F.3 HARDWARE REQUIREMENT

Training, inference, and other accelerated computing require only a single 3090 graphics card with 24GB of VRAM. The platform is equipped with an Intel(R) Xeon(R) Silver 4210R CPU 2.40GHz core and 256GB of physical memory with a frequency of 1600MHz.

F.4 RUNTIME AND SCALABILITY

We demonstrate that the pixel/concept perturbation-based causal intervention method is more stable and performs better than other negative concept intervention strategies for CBM. However, time consumption is unavoidable in the two-stage causal computation process, and the computation time results for the two levels of causal effect map are shown in **Table X**. The time taken for HI can be approximated as the sum of the times cost for CI and PI.

Table X: Runtime of causal interventions for NFResNet-50 ($n_{\text{coarse}} = 3$, $n_{\text{fine}} = 50$, inference $\text{bs} = 64$).

Datasets & Settings	Samples Number	Cost-Level	CI	HI
Flower-102	1633	Sample Cost	12.87 s	6.36 s
		Dataset Cost	5.83 h	2.88 h
Aircraft	3333	Sample Cost	25.05 s	6.36 s
		Dataset Cost	23.19 h	5.88 h
CUB-200	2000	Sample Cost	49.13 s	6.36 s
		Dataset Cost	27.29 h	3.53 h

Faster, but requires a trade-off of some performance, using gradient-based methods such as CCTP instead of coarse-stage causal calculation to reduce the computation. At the concept-level, $v_{\text{causal}}^{\text{coarse}}$ is calculated using gradient contributions. At the pixel-level, back-propagation is performed on the prediction score for each concept, generating n_{cpt} gradient maps. These are then merged to obtain $m_{\text{causal}}^{\text{coarse}}$. The time costs and performance evaluation results of replacing the coarse-stage with the gradient algorithm for PI and CI are shown below (**Table XI, XII and XIII**).

Table XI: A comparison of the time cost of CI and CI computation using gradient algorithms (CCTP) to replace the coarse-stage causal effect calculation (NFResNet-50).

Datasets & Settings	Cost-Level	CI (Ours)	Replace CI w/ Grad (Coarse-Stage)
Flower-102	Sample Cost	12.87 s	0.043 (coarse) + 3.21 (fine) s
($n_{\text{coarse}} = 3, n_{\text{fine}} = 50, n_{\text{cpt}} = 2550$)	Dataset Cost	5.83 h	1.47 h
Aircraft	Sample Cost	25.05 s	0.046 (coarse) + 6.26 (fine) s
($n_{\text{coarse}} = 3, n_{\text{fine}} = 50, n_{\text{cpt}} = 5100$)	Dataset Cost	23.19 h	5.84 h
CUB-200	Sample Cost	49.13 s	0.051 (coarse) + 12.28 (fine) s
($n_{\text{coarse}} = 3, n_{\text{fine}} = 50, n_{\text{cpt}} = 10000$)	Dataset Cost	27.29 h	6.87 h

Table XII: A comparison of the time cost of PI and PI computation using gradient algorithms (CCTP) to replace the coarse-stage causal effect calculation (NFResNet-50).

Datasets & Settings	Cost-Level	PI (Ours)	Replace PI w/ Grad (Coarse-Stage)
Flower-102	Sample Level	6.36 s	0.042 (coarse) + 2.89 (fine) s
($n_{\text{coarse}} = 3, n_{\text{fine}} = 50$)	Dataset Cost	2.88 h	1.32 h
Aircraft	Sample Level	6.36 s	0.044 (coarse) + 2.89 (fine) s
($n_{\text{coarse}} = 3, n_{\text{fine}} = 50$)	Dataset Cost	5.88 h	2.71 h
CUB-200	Sample Level	6.36 s	0.049 (coarse) + 2.89 (fine) s
($n_{\text{coarse}} = 3, n_{\text{fine}} = 50$)	Dataset Cost	3.53 h	1.63 h

By using gradient-based algorithm, we significantly reduce the time cost while retaining the majority of the intervention gain. It is possible to sacrifice more gain to fully utilize gradients to discover

Table XIII: Compare the performance of HI with that of HI which uses CCTP to replace coarse-stage causal computation at the concept-level (CI) and pixel-level (PI), respectively (NFResNet-50).

Datasets & Metrics	HI (Ours)	HI - Replace CI w/ Grad	HI - Replace PI w/ Grad	Original
Flower-102 (ACC)	**0.9769 ± 0.0019**	0.9751 ± 0.0008	0.9724 ± 0.0010	0.9608
$\mathcal{M}_{\text{cls}} \uparrow$	**0.2314 ± 0.0094**	0.2248 ± 0.0047	0.2022 ± 0.0064	—
Aircraft (ACC)	**0.7473 ± 0.0036**	0.7437 ± 0.0020	0.7327 ± 0.0039	0.6372
$\mathcal{M}_{\text{cls}} \uparrow$	**0.8164 ± 0.0091**	0.7964 ± 0.0049	0.7253 ± 0.0204	—
CUB-200 (ACC)	**0.7313 ± 0.0011**	0.7267 ± 0.0015	0.7165 ± 0.0061	0.6360
$\mathcal{M}_{\text{cls}} \uparrow$	**1.0481 ± 0.0338**	0.9538 ± 0.0116	0.8840 ± 0.0154	—

negative concepts, which only takes a few minutes for a dataset. This confirms the framework can be significantly accelerated for larger datasets.

To further investigate the impact of replacing our intervention mechanisms with gradient-based methods on generalization, we compare the original HI-Gen with variants that replace Concept-Level Intervention (CI) and Pixel-Level Intervention (PI) with gradient-based alternatives, as well as with the baseline and fine-tuning approaches. Results are reported in **Table XIV** using NFResNet-50 as the backbone.

Table XIV: Compare the performance of HI-Gen with that of HI-Gen which uses CCTP to replace coarse-stage causal computation at the concept-level (CI) and pixel-level (PI), respectively (NFResNet-50).

Dataset	Metric	Baseline	Fine-Tune	HI-Gen	HI-Gen (Replace CI w/ Grad)	HI-Gen (Replace PI w/ Grad)
Flower-102	ACC	96.26	96.63 ± 0.04	$\mathbf{97.18 \pm 0.02}$	96.70 ± 0.08	96.52 ± 0.06
	$\mathcal{M}_{\text{cls}}$	–	0.0348 ± 0.0941	$\mathbf{0.8048 \pm 0.0123}$	0.0923 ± 0.0079	0.0626 ± 0.0147
Aircraft	ACC	63.63	65.71 ± 0.22	$\mathbf{67.45 \pm 0.15}$	66.07 ± 0.06	64.90 ± 0.03
	$\mathcal{M}_{\text{cls}}$	–	0.2563 ± 0.0774	$\mathbf{0.7441 \pm 0.0096}$	0.2966 ± 0.0132	0.1716 ± 0.0029
CUB-200	ACC	63.46	64.92 ± 0.04	$\mathbf{65.27 \pm 0.05}$	64.78 ± 0.05	64.26 ± 0.15
	$\mathcal{M}_{\text{cls}}$	–	0.1609 ± 0.0440	$\mathbf{0.3068 \pm 0.0001}$	0.1774 ± 0.0017	0.1186 ± 0.0110

At the Application Level, compared to gradient-based derivative methods, HI indeed entails higher computational overhead, traded for gains in terms of intervention and generalization. However, we need to clearly point out that CBM-based models are primarily designed for high-risk decision-making scenarios. In these scenarios, interpretability, reliability, and controllability are often more critical than deployment efficiency alone. Moreover, many such applications do not impose strict requirements on real-time inference latency or the computational cost of offline model intervention and optimization. In practical CBM applications, including medical diagnosis, pathology analysis, and scientific decision support, intervention is typically conducted offline during model validation, auditing, or model refinement. Therefore, the additional computational cost is a reasonable trade-off for obtaining a more reliable, interpretable, and high-performing model. For example, in medical image analysis based on chest X-rays, interpretable concepts such as "opacity", "pleural effusion", and "lung consolidation" are commonly used to support clinically meaningful predictions. However, due to dataset biases or spurious correlations, models may incorrectly associate irrelevant image regions with disease-related concepts. By introducing hierarchical intervention (HI), which establishes a correction pathway from pixel-level evidence to concept-level reasoning, our method can identify and mitigate such feature-level biases before transferring the corrected knowledge back to the original model. This process ultimately improves the robustness and reliability of the deployed model while preserving the interpretability advantages of CBMs.

G LIMITATIONS

- Concept bottleneck size sensitivity: In experiments, we observed that CIFAR-10, unlike other datasets, exhibits strong concept sensitivity in ablation experiments ($\mathcal{N}$ and $\mathcal{P}$) and concept shift evaluation. This means that even slight modifications to pixels and concepts can cause significant feature shifts and performance degradation. Analogous to black-box using MLP as the classification head, this might be due to the insufficient number of concept nodes, causing each class prediction node to be densely connected to the concept prediction during model fitting. In other words, for each class prediction, there are few or no redundant concepts. Therefore, when we slightly increase the intervention degree, even a small change in concept prediction can cause a significant shift in class prediction. However, it is now quite simple to expand the concept bottleneck using LLMs, so small-scale and sparse concepts will not have a significant impact on the application of CBM-PCCEI.
- Performance bottleneck caused by distillation: We found that CBM-PCCEI yields huge gains when applied to the validation set using hierarchical intervention, but a mismatched effect when generalized to the test set using distillation. Editing specific patches of the input image on the validation set can directly place sample points in favorable classification positions within the feature space, while distillation fine-tuning adjusts the data distribution across the entire sample space to find a better decision boundary. However, due to limitations in model parameters and architecture, the optimization range is limited, and overly demanding adjustments (intervening in multiple patches) may even lead to performance degradation. For reference, this phenomenon is also mentioned in (Qian et al., 2025). When the teacher model performs exceptionally well, the gap between its distribution and the student's original capacity becomes a "distillation ceiling". We discuss and supplement the optimization related to this limitation in **Appendix. M**.
- Finer-grained pixel intervention is expensive: Computing causal effects in pixels instead of patches might provide more detailed explanations and intervention granularity, but the computational cost is too expensive, i.e. for a input $x \in \mathbb{R}^{256\times256\times3}$, patch-oriented analyze have to calculate $\mathcal{M}_{\text{cls}}$ for $3,072$ times while pixel-oriented require $196,608$ times when $\mathcal{K} = 8$ and $n_{\text{coarse}} = 3$. For a detailed discussion of computation time and optimization, please refer to **Appendix. F.4**.

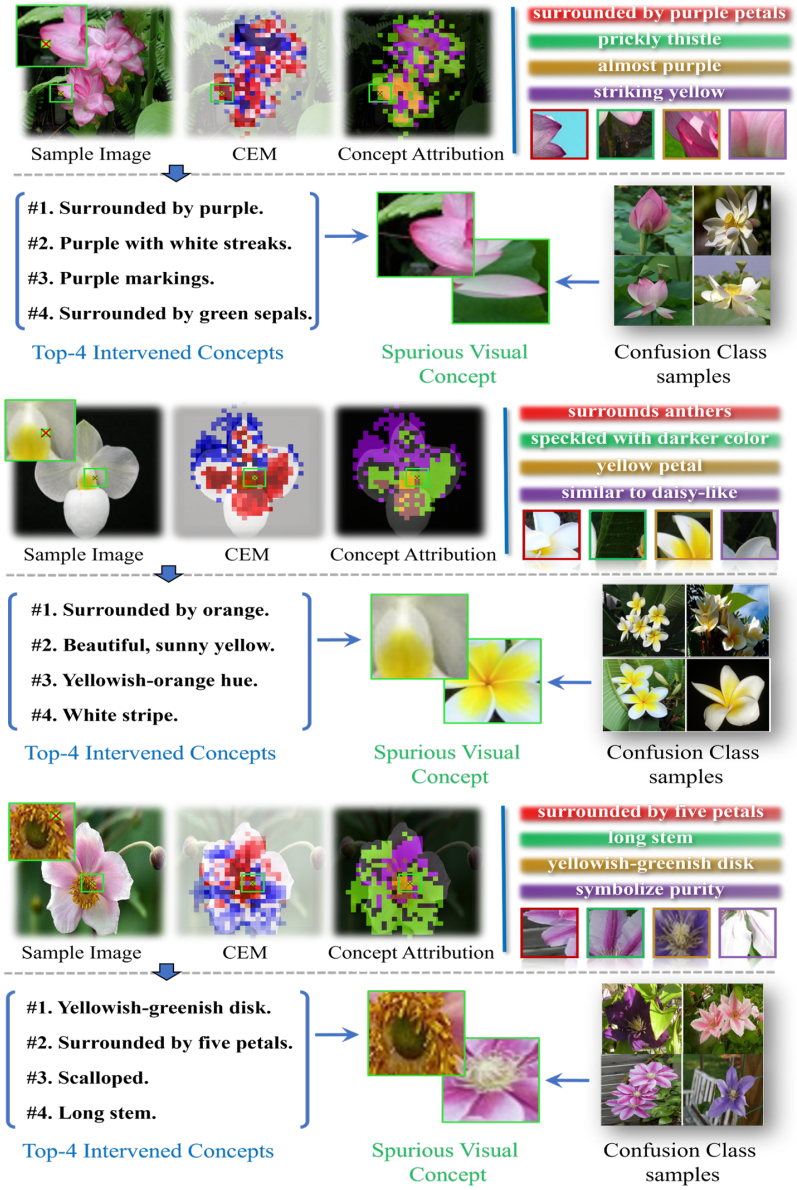
H DISCUSSIONS

Method extension: CBM-PCCEI can be applied to virtually any concept-bottleneck-based method, such as PCBM and CBM-HNMU, without requiring additional annotations. Hierarchical intervention capabilities are enabled simply by setting a metric function for the optimization objective and calculating concept-level and pixel-level causal effects.

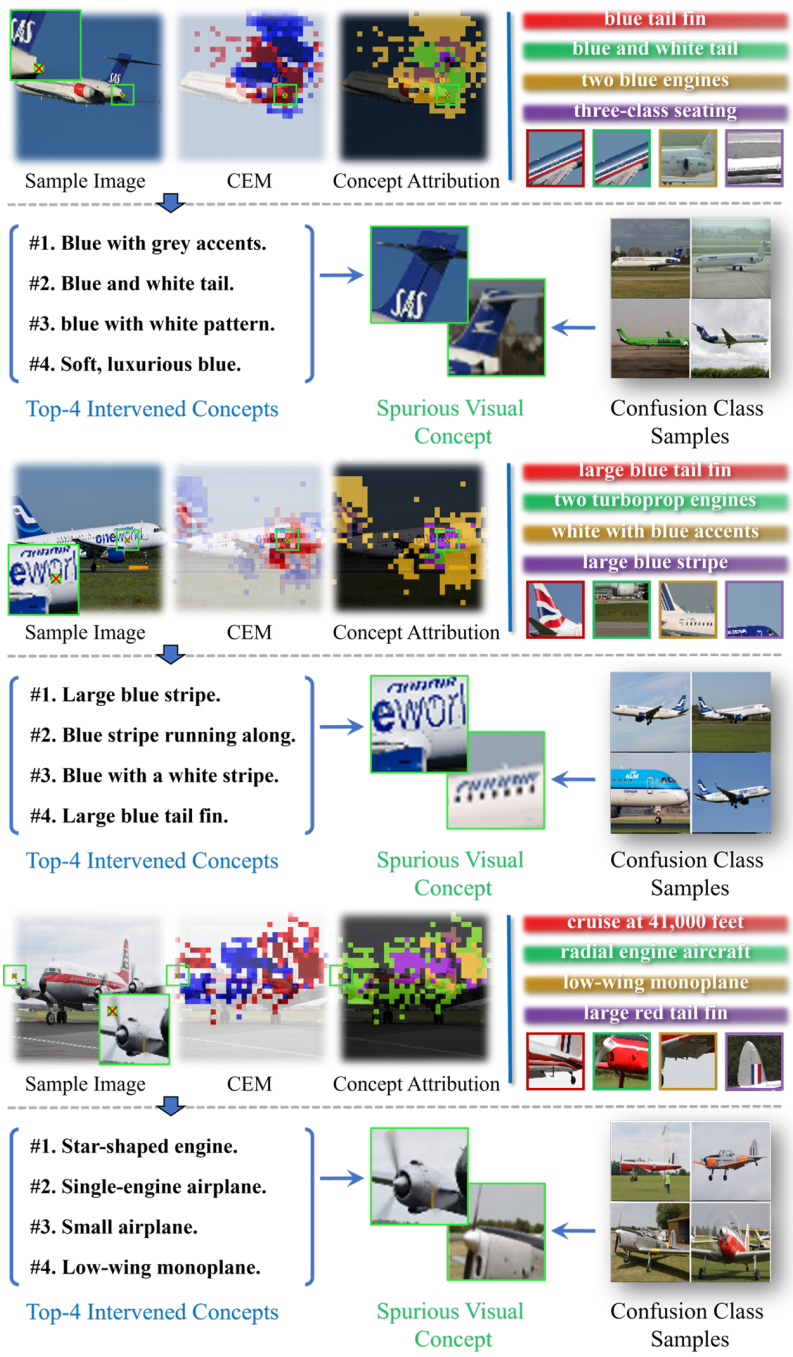
Future work: Due to the generalization bottleneck caused by the distillation, future work can focus on optimizing the distillation method and other sample-based plug-and-play intervention methods. For the former, this could involve increasing the explicit constraints of the optimization objective by adapting the original CBM feature space to alleviate the feature shift of the distilled CBM while simultaneously optimizing the feature distribution of misclassified sample points. For the latter, research could explore optimizing methods such as memory bank or introducing intervention pattern learning to record sample features and intervention patterns during the intervention process on the validation set and generalize them to the test set.

I MORE VISUALIZATIONS OF REASONING AND INTERVENTION

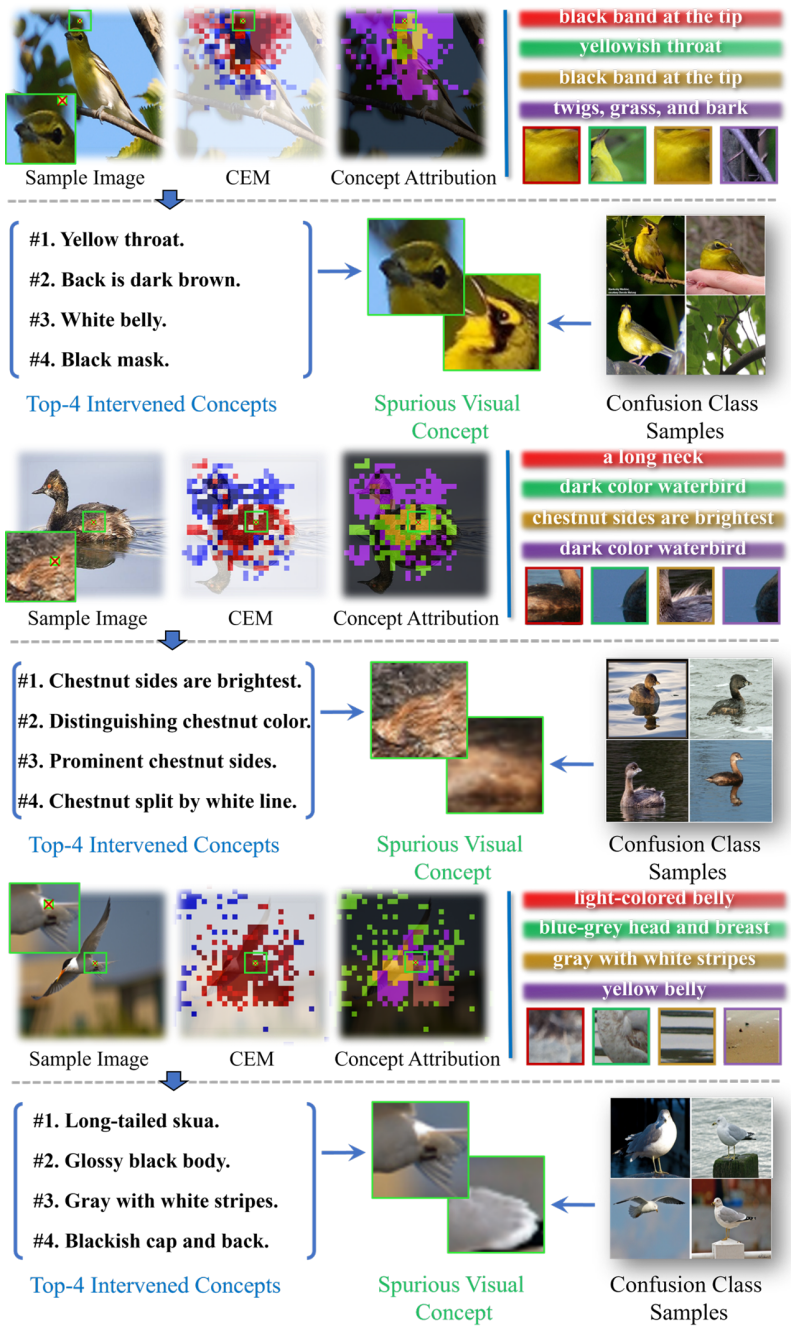
We integrate reasoning and intervention explanations and provide more visual examples here.



Samples of Flower-102.

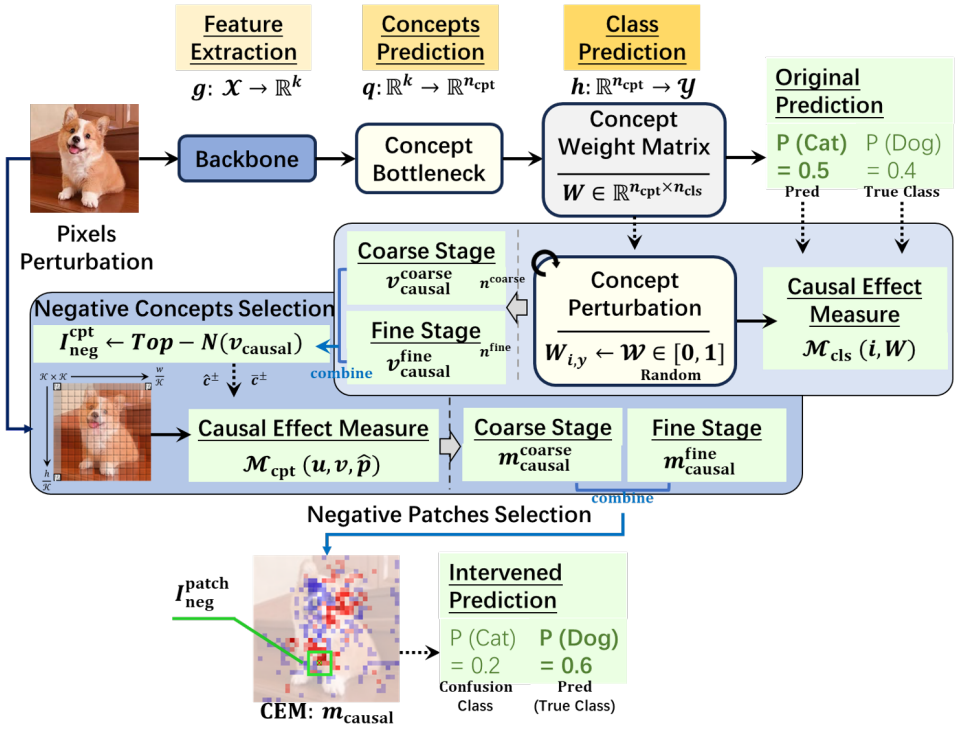


Samples of Aircraft.



Samples of CUB-200.

J HIERARCHICAL CAUSAL EFFECT INTERVENTION ARCHITECTURE DETAILS



K DIV2K-RELATED ABLATION

To ensure DIV2K doesn't provide "cheating" cues, we conducted an ablation study using **Grayscale Masks** vs. **DIV2K Crops**. We perform ablation on four scenarios using DIV2K and Grayscale Mask as patch for pixel causal effect calculation and intervention: ❶ Causal effect calculation and intervention using DIV2K; ❷ Causal effect calculation using DIV2K, but intervention using a Grayscale Mask; ❸ Causal effect calculation and intervention using a Grayscale Mask; ❹ Causal effect calculation using a Grayscale Mask, but intervention using DIV2K.

Table XV: A comparison of performance of HI using different crops for causal calculation and intervention (NFResNet-50).

Datasets	Causal Calcu by **Grayscale Mask**	Causal Calcu by **DIV2K**	**Original** ACC / $\mathcal{M}_{\text{cls}}$ ↑
Flower-102	ACC / $\mathcal{M}_{\text{cls}}$ ↑	ACC / $\mathcal{M}_{\text{cls}}$ ↑	—
Intervention by **Grayscale Mask**	0.9769±0.0019 / 0.2250±0.0096	0.9769±0.0016 / 0.2179±0.0156	0.9608 / —
Intervention by **DIV2K**	0.9769±0.0019 / 0.2171±0.0115	0.9769±0.0009 / 0.2314±0.0094	0.9608 / —
Aircraft	ACC / $\mathcal{M}_{\text{cls}}$ ↑	ACC / $\mathcal{M}_{\text{cls}}$ ↑	—
Intervention by **Grayscale Mask**	0.7423±0.0030 / 0.7987±0.0173	0.7350±0.0069 / 0.7445±0.0510	0.6732 / —
Intervention by **DIV2K**	0.7394±0.0015 / 0.7864±0.0073	0.7473±0.0036 / 0.8164±0.0091	0.6732 / —
CUB-200	ACC / $\mathcal{M}_{\text{cls}}$ ↑	ACC / $\mathcal{M}_{\text{cls}}$ ↑	—
Intervention by **Grayscale Mask**	0.7343±0.0057 / 1.1135±0.0372	0.7303±0.0055 / 1.0264±0.0376	0.6360 / —
Intervention by **DIV2K**	0.7135±0.0026 / 0.8773±0.0152	0.7313±0.0011 / 1.0481±0.0338	0.6360 / —

As shown in **Table XV**, the results are generally similar. Using only Grayscale Mask on CUB-200 even achieved the best accuracy and class discrimination. This indicates that DIV2K does not provide any additional assistance or cheating to the classifier. We chose DIV2K mainly to maintain the natural distribution of the image during causal estimation and avoid the sharp edges of artificial masks.

L STATISTICS OF ERROR CORRECTION AND MISJUDGMENT

Our method filters negative concepts by calculating the measure $\mathcal{M}_{\text{cls}}$ corresponding to multiple random intervention values $\mathcal{W} \in [0, 1]$, rather than 0. This introduces a kind of "random uncertainty" that, to some extent, fills the gap in the impact of the intrinsic correlation between intervention concepts on the prediction, thus ensuring that we can identify concepts with robust harmfulness. We provide the Correction/Error-Correction (CORR/ERR-CORR) statistics below. ERR-CORR represents rare cases where intervening on a correct sample leads to misclassification.

Table XVI: Statistical analysis of Correction (CORR) / Error-Correction (ERR-CORR) samples after intervention based on HI across all five datasets and two model architectures.

Models	**Flower-102**	**Aircraft**	**CUB-200**	**CIFAR-10**	**CIFAR-100**
NFRes-50 (CORR/ERR-CORR)	CI: 4/0 HI: 26/0	CI: 61/0 HI: 371/1	CI: 38/0 HI: 191/7	CI: 18/0 HI: 130/4	CI: 42/0 HI: 441/1
DaVIT (CORR/ERR-CORR)	CI: 3/0 HI: 9/0	CI: 33/0 HI: 316/2	CI: 30/0 HI: 169/3	CI: 14/0 HI: 50/5	CI: 66/0 HI: 320/2

As shown in **Table XVI**, across 5 datasets and two model architectures, our Hierarchical (HI) and Concept-Level (CI) Intervention achieves significant corrections with negligible "side-effect" misclassifications. It is worth discussing that joint interventions using complex combinations and distributions of concepts may produce better results and be more in line with the real world and human perception, although this would incur greater computational overhead.

M Discussion and Reflection on Performance Limitations in Intervention Generalization

M.1 Optimization of distillation by Teacher Adapter

As mentioned in **Appendix. G**, there are limitations to using distillation to achieve intervention generalization (knowledge transfer), which results in the intervention gain on the validation set being higher than the gain during intervention generalization. We attempted to use a Teacher Adapter to narrow the distributional differences between the intervention model's features and logits (on validation) and the original model. As shown in **Table XVII**, the results show that using the adapter does help alleviate the distillation bottleneck, but the gain is still limited. We believe that the capacity/learning ability of the model itself also limits the distillation effect. *Overall, distillation is currently more stable and effective compared to other generalization methods. Furthermore, given the experimental results and analysis above, there is still room for improvement and exploration in the generalization of interventions based on distillation.*

Table XVII: The results of using Teacher Adapter to optimize intervention generalization (NFResNet-50).

Mods & Datasets	**Flower-102**	**Aircraft**	**CUB-200**
Baseline-Val	0.9608	0.6372	0.6550
HI-Val (Ours)	0.9769	0.7473	0.7313
Baseline-Gen	0.9626	0.6363	0.6346
HI-Gen (Ours)	0.9718	0.6745	0.6527
Adapter-Gen	**0.9740**	**0.6878**	**0.6632**

M.2 LEARNING TO MATCH INTERVENTION FEATURES

To avoid distillation bottlenecks, we also consider intervening in generalization autonomy "at test time" so that the model can correct classification errors without any training (distillation) phase. We explored two proactive intervention strategies using memory caching (inspired by CB2M) to test whether the model could autonomously intervene based on feature matching at test time without distillation:

ImgF-Based: Caching Dict - {image features: intervened concept scores}.

CptS-Based: Caching Dict - {original concept scores: intervened concept scores}.

We compared the intervention generalization results implemented using the memory caching method and the distillation method (HI-Gen), and the results are as **Table XVIII**:

Table XVIII: HI-Gen .vs ImgF-Based and CptS-Based Method (NFResNet-50).

Datasets	**Metric**	**Baseline**	**ImgF-Based**	**CptS-Based**	**HI-Gen (Ours)**
Flower-102	ACC	0.9626	0.9639	0.9642	**0.9718**
	$\mathcal{M}_{\text{cls}} \uparrow$	—	0.0501	0.0897	**0.8048**
Aircraft	ACC	0.6363	0.6265	0.6157	**0.6745**
	$\mathcal{M}_{\text{cls}} \uparrow$	—	-0.0708	-0.0789	**0.7441**
CUB-200	ACC	0.6346	0.6205	0.6134	**0.6527**
	$\mathcal{M}_{\text{cls}} \uparrow$	—	-0.1439	-0.1700	**0.3068**
CIFAR-10	ACC	0.9542	0.9449	0.9470	**0.9566**
	$\mathcal{M}_{\text{cls}} \uparrow$	—	-0.0480	-0.0082	**3.3282**
CIFAR-100	ACC	0.8133	0.8046	0.8003	**0.8193**
	$\mathcal{M}_{\text{cls}} \uparrow$	—	-0.0188	0.0576	**0.1196**

Most of these "policy-based" caching methods even failed to outperform the baseline model. This may be because the representation of intervention strategies is too complex, and specific, limited feature representations cannot effectively match or learn specific patterns. Therefore, referring to CBM-HNMU, we found that knowledge distillation is more effective for intervention generalization. As a future work (see **Appendix. H**), improving memory-bank-based feature matching methods, optimizing distillation patterns, and exploring new intervention mode learning strategies all have great exploratory value.

N Ablation of CLIP Supervision

To separately evaluate the gains contributed by CLIP supervision to our proposed method, we removed the CLIP-based supervision while keeping the intervention mechanism unchanged, and measured intervention effectiveness (validation) and generalization performance under the same settings. The results are shown below (using NFResNet-50 as the backbone network, **Table XIX**):

Table XIX: Ablation study of CLIP supervision on HI and HI-Gen.

Dataset	Method	w/ CLIP (Ours)	w/o CLIP
Flower-102	HI	$\mathbf{97.69 \pm 0.09}$	97.55 ± 0.06
	HI-Gen	$\mathbf{97.18 \pm 0.02}$	97.01 ± 0.16
Aircraft	HI	$\mathbf{74.73 \pm 0.36}$	74.03 ± 0.18
	HI-Gen	$\mathbf{67.45 \pm 0.15}$	67.17 ± 0.58
CUB-200	HI	73.13 ± 0.11	$\mathbf{74.27 \pm 0.38}$
	HI-Gen	65.27 ± 0.05	$\mathbf{65.27 \pm 0.04}$

The results show that removing CLIP supervision has negligible impact on both validation-time intervention and generalization performance, indicating that the improvements mainly come from the proposed intervention mechanism rather than CLIP supervision.

O ABLATION WITH CBM-HNMU

To further isolate the specific contribution of CBM-PCCEI and evaluate whether the performance gains originate from the proposed hierarchical intervention mechanism, we replaced the hierarchical intervention strategy with the concept selection strategy based on Global Gradient Contribution from CBM-HNMU. We then compared this gradient-based variant with our original method under identical experimental settings, including both validation-time intervention performance and test-time generalization performance. The results are presented below (**Table XX**)):

Table XX: Comparison of intervention strategies between CBM-PCCEI and CBM-HNMU based on global gradient contribution.

Dataset	Setting	Metric	HI (Ours)	CBM-HNMU
Flower-102	**INT**	ACC	$\mathbf{97.69 \pm 0.09}$	96.57
		$\mathcal{M}_{\text{cls}}$	0.2314 ± 0.0094	$\mathbf{0.2853}$
	Gen	ACC	$\mathbf{97.18 \pm 0.02}$	96.68 ± 0.05
		$\mathcal{M}_{\text{cls}}$	$\mathbf{0.8048 \pm 0.0123}$	0.2463 ± 0.0072
Aircraft	**INT**	ACC	$\mathbf{74.73 \pm 0.36}$	66.94
		$\mathcal{M}_{\text{cls}}$	$\mathbf{0.8164 \pm 0.0091}$	0.6851
	Gen	ACC	$\mathbf{67.45 \pm 0.15}$	65.34 ± 0.26
		$\mathcal{M}_{\text{cls}}$	$\mathbf{0.7441 \pm 0.0096}$	0.2636 ± 0.0191
CUB-200	**INT**	ACC	$\mathbf{73.13 \pm 0.11}$	66.85
		$\mathcal{M}_{\text{cls}}$	$\mathbf{1.0481 \pm 0.0338}$	0.8972
	Gen	ACC	$\mathbf{65.27 \pm 0.05}$	64.36 ± 0.11
		$\mathcal{M}_{\text{cls}}$	$\mathbf{0.3068 \pm 0.0001}$	0.1601 ± 0.0219

The results demonstrate that our method consistently achieves higher intervention efficiency and stronger generalization performance compared with the Global Gradient Contribution. Replacing our hierarchical intervention mechanism with the global gradient contribution algorithm produces results comparable to those obtained by interventions relying solely on concept-level gradients or causal effects. This further highlights the importance of the proposed pixel-level cascaded intervention mechanism, which enables more precise and effective corrections by establishing hierarchical causal effects ranging from the pixel level to the conceptual level.