VFL: VISUAL FUNCTION LOCALIZATION FOR AUTOREGRESSIVE IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In autoregressive image generation, visual functions refer to hypothesized internal processes that support behaviors such as introducing attributes and refining local details. These processes remain poorly understood. Circuit analysis explains how internal computations contribute to behavior in language models, but extending it to image generation presents several challenges. Visual concepts typically span multiple image tokens, attention heads may interact, and intervention effects must be evaluated on the final image. We introduce VFL (Visual Function Localization), a framework for studying these functions through activation interventions, i.e., controlled changes to internal activations. VFL first identifies image-token positions where interventions most clearly produce the behavior associated with each visual function. It then learns how strongly to intervene at each attention head at those positions. We evaluate VFL on four visual functions using single-object images with simple backgrounds. Across all four functions, the interventions produce visual changes consistent with the corresponding functional hypotheses. The learned masks assign weight to multiple attention heads rather than concentrating on a single head.

1 INTRODUCTION

Understanding how image generators transform prompts into visual content remains a central challenge in mechanistic interpretability. In particular, the internal processes that introduce attributes, compose structures, and propagate information across image tokens remain poorly understood. We refer to these hypothesized processes as *visual functions*. Existing interpretability research has primarily focused on diffusion models, using cross-attention attribution and intermediate representations to trace how visual concepts are represented and updated during generation (Tang et al., 2023; Toker et al., 2024). In contrast, autoregressive image generators (Yu et al., 2022; Sun et al., 2024; Wang et al., 2024) predict tokens sequentially rather than through iterative denoising. This difference limits the direct application of diffusion-based analyses to their internal computations.

Like autoregressive language models, autoregressive image generators predict tokens sequentially, with each generated token influencing subsequent predictions. This similarity motivates extending circuit analysis to autoregressive image generation. Circuit analysis is a mechanistic interpretability framework widely used to study the internal computations that support specific behaviors in language models. At a high level, it employs interventions that modify selected internal computations and measure the resulting behavioral changes. This analysis involves three stages. Intervention-position selection determines where in the token sequence to intervene. Function localization identifies where inside the model to intervene, and intervention-effect evaluation measures the resulting change in the target behavior. Together, these stages provide a systematic framework for studying visual functions in autoregressive image generation.

Extending circuit analysis to autoregressive image generation requires addressing distinct challenges in all three stages. First, text tokens often carry explicit semantic meaning, whereas recognizable visual concepts typically emerge only across multiple image tokens. This difference makes intervention positions harder to identify than in language-model circuits. Second, the computations supporting a visual function may be distributed across the model rather than concentrated in a few independently important attention heads. Transformer computations can interact, encode redundant information, and compensate for one another (Elhage et al., 2021; McGrath et al., 2023). Con-

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.

sequently, interventions on isolated heads may not capture their collective contribution to a visual function. Third, visual intervention effects require image-level evaluation. In language-model circuits, target-token logits often provide a direct behavioral measure because the target token itself carries task-level semantic meaning. Individual image tokens, by contrast, do not fully represent visual concepts, and a change at one position can propagate to subsequent predictions. We therefore assess intervention effects on the final decoded image rather than on a single token.

To address these challenges, we introduce **VFL** (**V**isual **F**unction **L**ocalization). VFL uses activations from one generation to modify the corresponding activations in another. We call the source generation the *donor* and the generation being modified the *recipient*. For **intervention-position selection**, VFL compares candidate *token sets*, each specifying a subset of image-token positions at which to intervene. It selects the set whose intervention best matches the target visual function through **Function-Guided Causal Probing (FCP)**. For **function localization**, VFL holds the selected token set fixed and employs **Differentiable Mask Learning (DML)** to learn how strongly to intervene at each attention head. For **intervention-effect evaluation**, we apply the learned intervention at the selected positions during full autoregressive generation and measure its effect on the final decoded image. This procedure links function-specific token positions and attention-head weights to their observable effects on generated images.

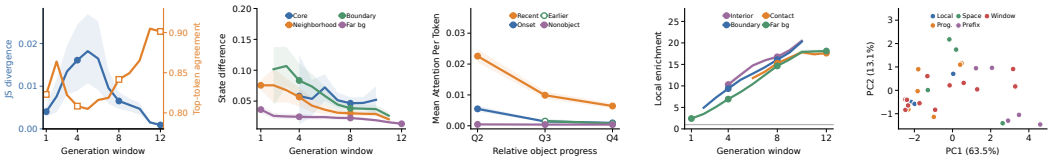
Following prior mechanistic localization work (Basu et al., 2024), we evaluate VFL on single-object images with simple backgrounds and vary the prompt-specified object color. Across the evaluated autoregressive image generators, interventions at the selected positions produce visual changes consistent with the corresponding functional hypotheses. The learned masks distribute weight across multiple attention heads rather than concentrating on a single head. We also compose masks for different visual functions to test whether their visual effects can appear together in a single generation. These results support VFL as an intervention-based approach to localizing visual functions across token positions and attention heads in autoregressive image generation.

Our main contributions are as follows. (1) To our knowledge, VFL is the first intervention-based framework for localizing visual functions in autoregressive image generation. (2) We develop two complementary components. **Function-Guided Causal Probing (FCP)** selects a token set for each visual function, and **Differentiable Mask Learning (DML)** learns the intervention strength at each attention head. We evaluate the resulting effects on the final decoded image. (3) In a controlled single-object setting, we show that interventions combining selected token sets and learned attention-head weights produce function-specific visual changes across autoregressive image generators.

2 Related Work

Circuit analysis in autoregressive models. Circuit analysis localizes behavior-relevant tokens, heads, and paths through activation interventions, causal probing, and model editing (Elhage et al., 2021; Wang et al., 2022; Conmy et al., 2023; Vig et al., 2020; Meng et al., 2022; Zhang & Nanda, 2024). Recent work extends this with path and attribution patching, automated circuit discovery, and analyses of distributed and function-level structure (Goldowsky-Dill et al., 2023; Syed et al., 2024; Lepori et al., 2023; Hsu et al., 2025; Geiger et al., 2024; Makelov et al., 2024; Todd et al., 2024; Marks et al., 2025). Autoregressive image generators preserve token ordering (Parmar et al., 2018; Ramesh et al., 2021), while related systems study structured steering and staged optimization (Li et al., 2026b; Song et al., 2026); VFL brings intervention-based localization to visual generation through token-set probing, layer–head weighting, and image-level evaluation.

Interpretability and perturbation learning for visual models. Visual-model interpretability studies visual units, diffusion representations, layer localization, and concept-related heads (Bau et al., 2018; Kong et al., 2024; Basu et al., 2024; Park et al., 2025). Recent work analyzes causal blocks, head-level mechanisms, diffusion concepts, cross-attention interventions, and multimodal causal representations (Zarei et al., 2025; Wang et al., 2026; Shi et al., 2025; Tinaz et al., 2025; Maliha & Hougen, 2026; Bae et al., 2026; Li et al., 2026a). Other methods control visual models through cross-attention editing, differentiable perturbations, and learnable gates (Hertz et al., 2022; Song et al., 2025; Xiao et al., 2026); VFL instead targets internal activation interventions in frozen autoregressive image generators.



(a) (b) (c) (d) (e)

Figure 1: **Generation-wide diagnostics and token-set profiles.** Panels (a) to (e) summarize the observations that motivate the four visual functions. In panels (a), (b), and (d), each generation window groups consecutive image-token generation steps.

3 Empirical Motivation for Visual Functions

We first examine how image-token predictions evolve during autoregressive generation and how they depend on the prompt and previously generated image tokens. These observations motivate four visual functions. Figure 1 summarizes the results, and Appendix D provides the detailed measurement procedures.

Figure 1(a) reports Jensen-Shannon divergence, which measures the difference between two next-token distributions, and top-token agreement, which indicates whether they assign the highest probability to the same token. Both measurements show that prompt-dependent differences are stronger early in generation and weaken later. Because generated tokens condition subsequent predictions, these early differences can influence the remaining image sequence. We therefore define **Fun. 1 (Attribute transmission)** as the transfer of prompt-specified attribute information into the image sequence during early generation.

Figure 1(b) shows that prompt changes affect hidden states differently in object-related regions and the far background. Here, state difference is the L2 distance between donor and recipient hidden states at the same image-token position. This spatial pattern suggests that prompt-specified attributes can remain associated with the intended object rather than affect the image uniformly. We define **Fun. 2 (Object-attribute ownership)** as the process that maintains this association.

Figure 1(c) examines attention from the current object token to previously generated image tokens. Relative object progress indicates how early or late the current token appears within the object, while mean attention per token is the average attention weight assigned to a previous token. Later object tokens assign substantially more attention to recent tokens from the same object than to non-object tokens. Figure 1(d) reports local enrichment, which measures whether nearby previous tokens receive more attention than under a uniform distribution over the image-token history. The generally increasing values indicate a stronger preference for nearby tokens as generation proceeds. Together, these patterns suggest that established object information may propagate through recent and nearby tokens as the object is generated. This motivates **Fun. 3 (Object-internal propagation)**, the stepwise transfer of established information to later positions within the same object.

The increasing local enrichment in Figure 1(d) also motivates examining whether nearby generated content supports local refinement. Figure 1(e) further shows that intervention effects vary across token sets. Based on these observations, we define **Fun. 4 (Local rendering refinement)** as the use of nearby generated visual information to refine local image content.

These observations motivate the four visual functions. In Section 4, we introduce VFL to localize them through interventions.

4 Methodology

4.1 Method Overview

We seek to identify where in the image-token sequence, and through which attention heads, the generator supports each visual function in Section 3.

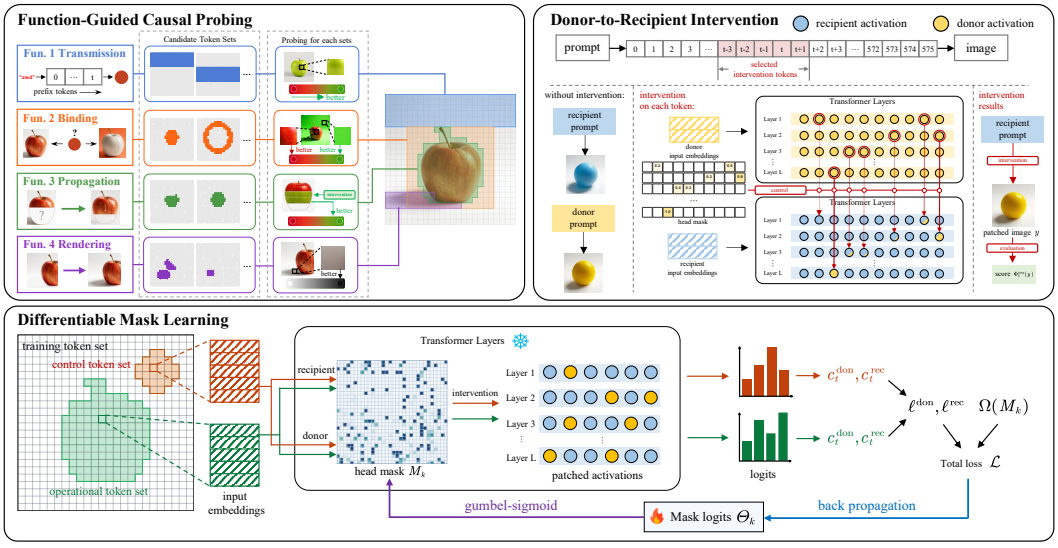


Figure 2: Overview of VFL. VFL localizes visual functions by identifying function-relevant image-token positions and their supporting attention heads through donor-to-recipient interventions, with effects evaluated on the generated image.

We use *activation interventions* to examine how selected attention heads contribute to a visual function. At chosen image-token positions, we adjust their activation outputs in the recipient toward the corresponding donor outputs, while keeping the generator weights and recipient prompt fixed. For example, we use a red-apple donor to test whether these internal changes shift the recipient apple toward red, even though its prompt still specifies green (Section 4.2).

Function-Guided Causal Probing. FCP aims to select the token set whose intervention best matches the intended visual effect. We therefore test predefined candidate sets through *all-head interventions*, which modify all heads in all layers at the positions in each position of the selected token set. Comparing the resulting images, together with the function definition, guides the selection of a token set (Section 4.3).

Differentiable Mask Learning. With the selected token set fixed, DML learns a head mask, i.e., one intervention weight per attention head. We jointly optimize these weights to make donor tokens more likely at function-relevant positions and recipient tokens more likely at positions whose content should be preserved, e.g., background positions during object-color transfer. The objective also favors using fewer heads; only mask parameters are trained (Section 4.4).

Validate the predicted visual effect. We retain a specified number of heads with the highest mask weights, keeping their learned weights and setting all other weights to zero. We apply this mask at the selected token set during full autoregressive generation and check whether the intended change occurs while the rest of the image remains largely unchanged. We also test whether intervening on other heads or token sets produces a weaker effect. Combining masks tests whether their visual effects can appear together (Section 4.5). Figure 2 outlines the framework.

4.2 Donor-to-Recipient Activation Intervention

We use activation interventions to examine how selected attention heads contribute to a visual function. At chosen image-token positions, we adjust their outputs in the recipient toward the corresponding donor outputs, while keeping the recipient prompt and generator weights fixed. We then continue generation and completed generating the final image.

To examine how specific token positions and attention heads affect generation, we mix donor head outputs into the recipient at selected positions, with a head mask controlling the contribution of each attention head. For pair i, let $y_i^{\text{rec}}, y_i^{\text{don}}$ and $v_i^{\text{rec}}, v_i^{\text{don}}$ denote the independently generated recipient and donor images and token sequences. Let $A_{i,t,l,h}^{\text{rec}}$ and $A_{i,t,l,h}^{\text{don}}$ be the head outputs under

their respective prompts at image-token position t, layer l, and head h, before the attention output projection. Given token set $\mathcal{T}$ and head mask M, we modify the recipient output as

$$\widetilde{a}_{i,t,l,h} = (1 - w_{l,h})A^{\text{rec}}_{i,t,l,h} + w_{l,h}A^{\text{don}}_{i,t,l,h}, \qquad w_{l,h} = \alpha M_{l,h}, \quad t \in \mathcal{T}, \tag{1}$$

where $\widetilde{a}$ is the modified output, $M_{l,h} \in [0, 1]$ weights each head, and $\alpha \geq 0$ controls overall strength. The output interpolates between recipient ($w_{l,h} = 0$) and donor ($w_{l,h} = 1$); $w_{l,h} > 1$ amplifies their difference. Thus, $\mathcal{T}$ selects positions and M selects and weights heads. Setting $\alpha = 1$ and all mask weights to one gives an all-head intervention. Setting $\alpha = 1$ and all mask weights to one gives an all-head intervention at the selected positions. The recipient then continues autoregressive generation using the modified activations and its own sampled tokens. We decode the completed sequence into the image $y_i(\mathcal{T}, M)$.

4.3 Intervention-Position Selection with Function-Guided Causal Probing

FCP seeks a token set whose intervention best matches the intended visual effect. For all four functions, we define candidate token sets, intervene on each set, and compare the resulting images with those from interventions elsewhere before selecting a set. We explain this shared process through **Fun. 2 (Object-attribute ownership)** as an example. This function keeps an attribute on its object, e.g., transferring a donor's red color to a recipient apple without similarly changing the background. The candidate positions selection and color scores below are specific to Fun. 2.

Which candidate token sets should we consider? Fun. 2 concerns whether a color change stays on the object, so its candidates distinguish positions within and around it. We map segmentation masks to token positions to form six sets comprising the inner and outer parts of the object interior, boundary, nearby background, boundary with nearby background, and *object neighborhood*. The neighborhood combines the outer interior, boundary, and nearby background (Figure 3(b)).

Does the color change stay with the object? We apply the all-head intervention in Eq. 1 separately at each candidate token set $\mathcal{T}$, obtaining the image $y_i(\mathcal{T}) = y_i(\mathcal{T}, \mathbf{1})$. For each resulting image, we compare progress toward the donor color in fixed object and background regions $\mathcal{O}$ and $\mathcal{B}$. Let $\bar{\mathbf{c}}_{\mathcal{A}}(y)$ be the mean RGB vector in region $\mathcal{A}$, and let $\mathbf{d}_{\mathcal{A},i} = \bar{\mathbf{c}}_{\mathcal{A}}(y_i^{\text{don}}) - \bar{\mathbf{c}}_{\mathcal{A}}(y_i^{\text{rec}})$ be the color difference between donor and recipient. We define regional color progress $P_{\mathcal{A},i}$ and the *ownership score* $\Phi^{\text{img}}_{2,i}$ as

$$P_{\mathcal{A},i}(y) = \frac{\langle \bar{\mathbf{c}}_{\mathcal{A}}(y) - \bar{\mathbf{c}}_{\mathcal{A}}(y_i^{\text{rec}}), \mathbf{d}_{\mathcal{A},i} \rangle}{\|\mathbf{d}_{\mathcal{A},i}\|_2^2 + \varepsilon}, \qquad \mathcal{A} \in \{\mathcal{O}, \mathcal{B}\}, \quad \Phi^{\text{img}}_{2,i}(y) = P_{\mathcal{O},i}(y) - P_{\mathcal{B},i}(y). \tag{2}$$

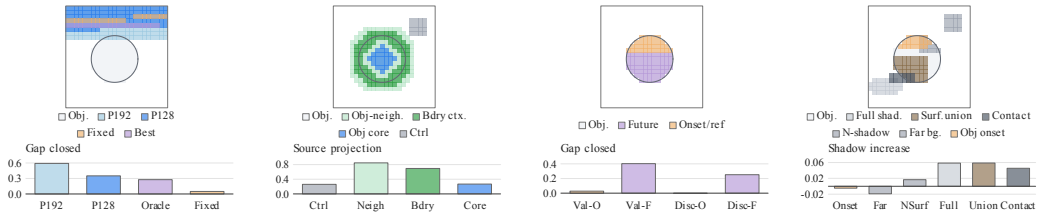
Here, $\varepsilon > 0$ prevents division by zero. An unchanged regional mean gives zero progress; positive progress indicates a change in the direction of the donor color. Subtracting background progress distinguishes object-color transfer from a similar change in both regions.

Could other token sets produce the same effect? A color change alone cannot distinguish a candidate if intervening elsewhere produces the same effect. We therefore repeat the intervention on two *control token sets* of the same size as $\mathcal{T}$, one in the far background and one randomly sampled from the background outside the object neighborhood. Equal sizes keep the number of intervened positions fixed. Let $\mathcal{C}_i(\mathcal{T})$ contain these two sets. We compare the candidate with the stronger control,

$$J_{2,i}(\mathcal{T}) = P_{\mathcal{O},i}(y_i(\mathcal{T})) - \max_{\mathcal{T}' \in \mathcal{C}_i(\mathcal{T})} P_{\mathcal{O},i}(y_i(\mathcal{T}')) \,. \tag{3}$$

Positive $J_{2,i}$ indicates greater object-color progress than either control. This difference compares intervention positions, whereas the ownership score compares regions within each image.

Which token set best matches each visual function? For Fun. 2, the object neighborhood ranks highest in mean object progress and ownership score across pairs, with a positive mean control difference, so we select it as the . For the other functions, we consider their definitions alongside the candidate comparisons, adopting an early prefix for Fun. 1, later object positions for Fun. 3, and shadow positions for Fun. 4. The selected set $\mathcal{T}^+_{k,i}$ for function k and pair i defines the *operational positions*, i.e., the positions used in DML and evaluation. Appendix A defines the image-level scores $\Phi^{\text{img}}_{k,i}$.



(a) Fun. 1. (b) Fun. 2. (c) Fun. 3. (d) Fun. 4.

Figure 3: **Function-specific token-set comparisons.** Scales differ across functions.

The selected rule ρ_k^{op} yields the operational token positions $\mathcal{T}_{k,i}^{+}$ for each pair i: an early prefix, object neighborhood, future-object region, and shadow region for Fun. 1–4, respectively.

4.4 Function Localization with Differentiable Mask Learning

With positions fixed by FCP, DML seeks heads whose joint intervention produces the intended change. We define token targets and a loss, then learn head weights. For function k, each *mask logit* $\Theta_{k,l,h}$ is a trainable score controlling a weight in M_k; only these scores are optimized.

What should change or stay the same? For Fun. 2, object-color transfer should change the apple's color but leave its background unchanged. At function-relevant positions, the prediction target is the donor token; at positions to preserve, including control token sets, it is the recipient token. We denote this *target token* by $v_{i,t}^{\mathrm{tar}}$ and use the other sequence's token at that position, $v_{i,t}^{\mathrm{cmp}}$, for comparison. Both come from the independently generated sequences, and both position types use the same mask.

The Training Process. To compare the prompts, both runs must receive the same earlier image tokens. We form one input sequence using donor tokens at function-relevant positions and recipient tokens elsewhere. Each run receives the tokens before position t from this sequence but keeps its own prompt. After applying Eq. 1, continued computation gives the next-token prediction scores $\widetilde{\mathbf{g}}_{i,t}(M_k)$.

Training should make the target token more likely than the other token. For pairs and positions sampled from the training distribution $\mathcal{D}_k$, we minimize

$$\Theta_k^{\star} = \arg\min_{\Theta_k} \mathbb{E}_{(i,t)\sim\mathcal{D}_k}\left[\ell_k\big(\widetilde{\mathbf{g}}_{i,t}(M_k), v_{i,t}^{\mathrm{tar}}, v_{i,t}^{\mathrm{cmp}}\big)\right] + \mathcal{R}_k(M_k). \tag{4}$$

The loss ℓ_k uses cross-entropy, i.e., the target token's negative log probability, and an optional margin penalty when its log probability is not sufficiently above the other token's. The regularizer $\mathcal{R}_k$ controls total mask weight and favors weights near zero or one.

How can head selection be learned? We seek head weights that make the intervention favor the target tokens. To optimize these weights, we employ Gumbel-Sigmoid (Maddison et al., 2016), which allows gradients to pass through sampled soft masks to the mask logits Θ_k. At each training update, we sample

$$M_{k,l,h} = \sigma\left(\frac{\Theta_{k,l,h} + \epsilon_{l,h}}{\tau}\right), \qquad \epsilon_{l,h} = \log U_{l,h} - \log(1 - U_{l,h}), \qquad U_{l,h} \sim \mathcal{U}(0,1). \tag{5}$$

Here, σ is the sigmoid function, $U_{l,h}$ are independent uniform samples, and $\epsilon_{l,h}$ is the resulting noise. The temperature $\tau > 0$ controls how close the soft weights are to zero or one. Initially, these soft weights are used directly in the intervention. Late in training, we instead use a binary mask in the forward pass, setting weights above $1/2$ to one and the others to zero, while computing gradients through the soft weights. This is a *straight-through estimator* (Bengio et al., 2013). After training, we omit the sampling noise and compute the deterministic soft mask from the optimized scores,

$$M_{k,l,h}^{\star} = \sigma\left(\frac{\Theta_{k,l,h}^{\star}}{\tau_{\mathrm{det}}}\right), \tag{6}$$

where $\tau_{\mathrm{det}} > 0$ is a fixed temperature. The resulting mask retains continuous weights rather than the binary values used in late-stage training.

4.5 INTERVENTION-EFFECT EVALUATION

We test whether learned masks produce the intended image changes, whether other heads work just as well, and whether different effects can appear together.

Does the image change as intended? Token predictions alone do not show the final visual effect. We therefore evaluate full autoregressive generations with an *evaluation mask* $\widehat{M}_k$. It retains the K_k highest weights of $M_k^\star$ unchanged, where K_k is the chosen head count, and sets the rest to zero. Applying it at $\mathcal{T}_{k,i}^+$ gives $y_{k,i}^{\mathrm{int}} = y_i(\mathcal{T}_{k,i}^+, \widehat{M}_k)$, scored by $s_{k,i}^{\mathrm{eval}} = \Phi_{k,i}^{\mathrm{img}}(y_{k,i}^{\mathrm{int}})$ (Section 4.3).

Would other heads work just as well? We move the learned weights to other heads within each layer, leaving the token set and per-layer head counts unchanged. These *head controls* test whether changing the heads weakens the image-level effect.

Can different visual effects appear together? We test whether separately learned effects can appear together, e.g., color and shadow changes. We combine the masks in one generation, each at its selected positions. With $r_k(t) = \mathbf{1}[t \in \mathcal{T}_{k,i}^+]$, the weight is $m_{k,t,l,h} = r_k(t)\widehat{M}_{k,l,h}$. Functions sharing donor q form the set S_q, with combined weight $m_{t,l,h}^{(q)} = 1 - \prod_{k \in S_q}(1 - m_{k,t,l,h})$. For each donor, we add the weighted activation difference $\Delta a^{(q)} = A^{\mathrm{don},q} - A^{\mathrm{rec},q}$ to a^{cur}, the recipient output immediately before intervention,

$$\widetilde{a}_{i,t,l,h} = a_{i,t,l,h}^{\mathrm{cur}} + \sum_q \frac{m_{t,l,h}^{(q)}}{\max\left(1, \sum_{q'} m_{t,l,h}^{(q')}\right)} \Delta a_{i,t,l,h}^{(q)}. \tag{7}$$

The denominator keeps the total donor weight at most one when masks overlap.

5 EXPERIMENTS

We examine whether VFL produces function-specific visual changes and how these depend on the selected heads, generator, and evaluation or training choices.

5.1 EXPERIMENTAL SETUP

Our setup makes the intended visual effects easier to assess. We use single-object images with simple backgrounds. Janus-Pro (Chen et al., 2025) is the main generator; Emu3 (Wang et al., 2024) provides supplementary cross-model evaluation. Both generators remain frozen. Interventions and image-level scores follow Section 4 and Appendix A.

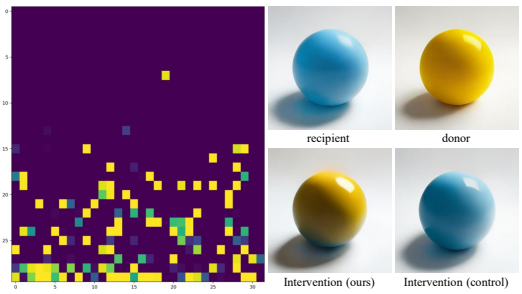
5.2 INTERVENTION VALIDATION

We first test whether interventions produce the intended visual effects while preserving the rest of the image. In Figure 4, the ideal result changes the object color toward the donor's for Fun. 1 to 3, or the ground shadow toward the donor's for Fun. 4, while retaining the rest of the recipient image. The observed changes are consistent with the four functional hypotheses, with smaller changes in background or early object regions (Figure 5(a)).

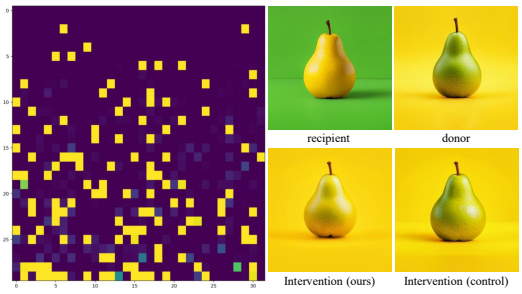
We next test whether other heads produce the same visual effects. Learned masks outperform random and lowest-weight head controls where tested, but some matched head controls remain close (Figure 5(b)).

5.3 CROSS-MODEL EVALUATION ON EMU3

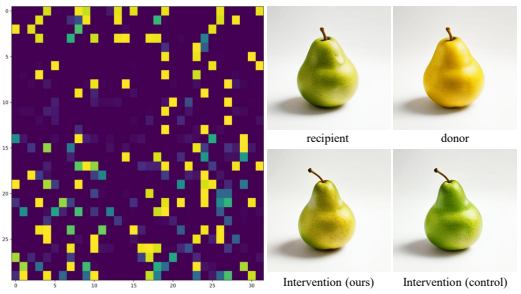
This supplementary experiment tests whether VFL is specific to Janus-Pro. On frozen Emu3, the same FCP and DML framework produces changes consistent with all four functional hypotheses, supporting application beyond Janus-Pro. Figure 6 shows representative results.



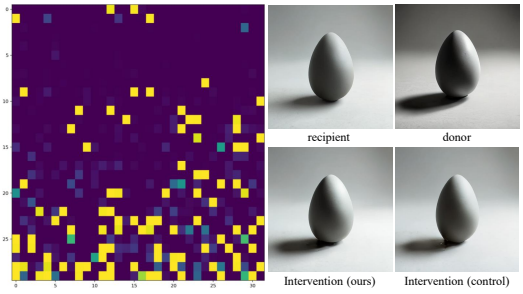
(a) Fun. 1: Attribute transmission.



(b) Fun. 2: Object–attribute ownership.



(c) Fun. 3: Object-internal propagation.



(d) Fun. 4: Local rendering refinement.

Figure 4: **Qualitative function-specific interventions.** Head mask, recipient, donor, our intervention, and head-control result for each function. The ideal result changes only the object color (Fun. 1 to 3) or ground shadow (Fun. 4) toward the donor, leaving the rest as in the recipient.

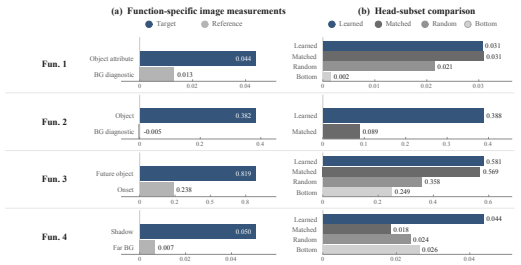
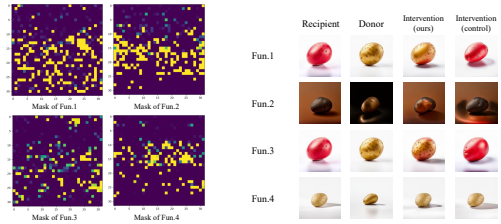


Figure 5: **Visual effects and head controls.** (a) Function-specific changes. BG denotes background; onset/future denote early/later object positions. (b) Head controls retain learned weights at the selected token sets. Random and Bottom select random and lowest-weight heads.



(a) Layer–head masks (b) Interventions

Figure 6: **Cross-model evaluation on Emu3.** (a) Learned head masks. (b) Representative interventions and head controls. FCP and DML are applied to frozen Emu3. The visual effects are consistent with the functional hypotheses, supporting application beyond Janus-Pro.

5.4 Further Analysis and Ablation Studies

We next examine whether visual effects can be combined, how they relate to mask weights, and how they vary with evaluation and training choices.

Can separately learned visual effects appear together? The combined intervention changes both color and shadow (Figure 7). Its joint score, which measures both effects, decreases with alternative token sets or donors, while some head controls remain comparable (Figure 9(d)). **What do mask weights reveal about head contributions?** Ranked mask weights are distributed across multiple heads (Figure 8(a)). Retraining on different subsets changes masks more for Fun. 2 and 3 (b). Adding head groups can increase or decrease scores (e). **Do visual effects persist across evaluation settings?** For Fun. 2, learned masks outperform matched head controls on most pairs not used for mask training (Figure 8(c)). For Fun. 3, learned masks outperform random and lowest-weight head controls on average at each tested strength (d).

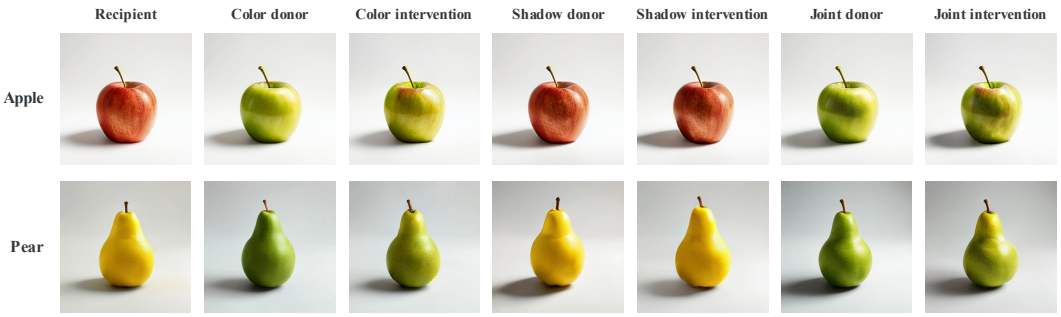
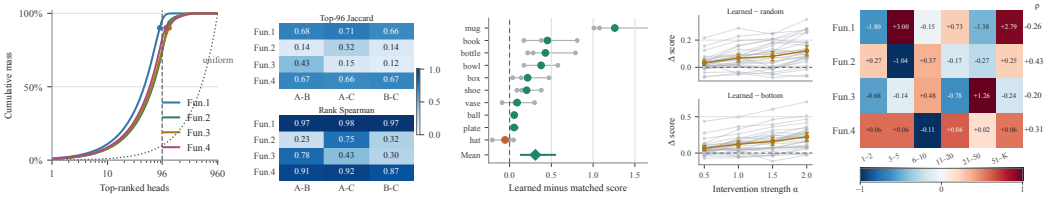
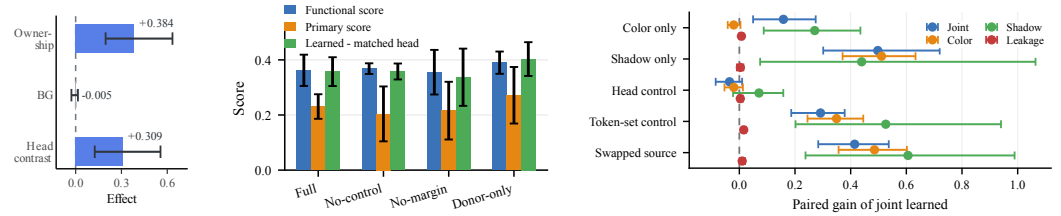


Figure 7: **Joint color and shadow interventions.** Single-function results provide comparisons for the combined effect.



(a) (b) (c) (d) (e)

Figure 8: **Head-mask structure and intervention effects.** (a) Cumulative mask weights. (b) Mask similarity across training subsets A/B/C. (c,d) Fun. 2 and 3 score differences from head controls. Gray marks show individual results; colored marks show means. Colored bars give 95% confidence intervals for overall means. (e) Score changes from adding head groups in decreasing weight order. Colors are normalized per row; ρ is the Spearman correlation with mean group weight; K is the retained head count.



(a) (b) (c)

Figure 9: **Evaluation and training ablations.** (a) Fun. 2 held-out scores separate the token-set effect, the background effect, and the learned-minus-matched-head contrast. (b) No-control removes control-position losses; No-margin removes the margin penalty; Source-only uses donor-token targets alone. (c) Paired gains compare the joint learned intervention with head, token-set, and donor controls. Leakage denotes changes to image content that should remain unchanged.

Figure 9 collects three additional ablations. For Fun. 2, the token-set effect is positive, the background effect stays near zero, and the learned heads improve over the matched head control in (a). Across objective variants, all scores remain positive, and Source-only is highest on this metric in (b).

6 Conclusion

We present a functional framework for understanding attention heads in autoregressive image generators through donor-to-recipient activation intervention. With function-specific token sets and sparse head masks, our experiments reveal interpretable visual function in attribute transfer, binding, propagation, and local rendering, showing how these models compose images token by token.

Ethics Statement

This work does not involve human or animal subjects or the collection of personally identifiable information. The experimental prompt pairs were manually constructed for controlled comparisons rather than derived from personal records or sensitive data. Because the analyzed pretrained generator may inherit biases from its training data, the findings should be interpreted within the controlled prompts and model setting evaluated here.

Reproducibility Statement

To support reproducibility, the main paper describes the model setting, intervention framework, evaluation protocol, and causal controls. The appendix provides additional details on the formal score definitions,token-set and token-set construction, mask training and deployment, prompt-pair construction, experimental configurations, and ablation protocols. The experimental prompt pairs were manually constructed for controlled comparisons.

AI Use Statement

Generative AI tools were used to assist with manuscript preparation, including language editing, structural refinement and scientific visualization. All AI-assisted outputs were reviewed and verified by the authors against the underlying experiments and source materials. The authors take full responsibility for the final content of the paper.

References

Jun-Hyun Bae, Wonyong Jo, Jaehyup Lee, and Heechul Jung. Mechanistic dissection of cross-attention subspaces in text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 19657–19665, 2026.

Samyadeep Basu, Keivan Rezaei, Priyatham Kattakinda, Vlad I Morariu, Nanxuan Zhao, Ryan A Rossi, Varun Manjunatha, and Soheil Feizi. On mechanistic knowledge localization in text-to-image generative models. In *Forty-first International Conference on Machine Learning*, 2024.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1811.10597*, 2018.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pp. 160–187. PMLR, 2024.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control.(2022). *URL https://arxiv.org/abs/2208.01626*, 3:3, 2022.

Aliyah Hsu, Georgia Zhou, Yeshwanth Cherapanamjeri, Yaxuan Huang, Anobel Odisho, Peter Carroll, and Bin Yu. Efficient automated circuit discovery in transformers using contextual decomposition. In *International Conference on Learning Representations*, volume 2025, pp. 57880–57898, 2025.

Xianghao Kong, Ollie Liu, Han Li, Dani Yogatama, and Greg Ver Steeg. Interpretable diffusion via information decomposition. In *International Conference on Learning Representations*, 2024.

Michael A Lepori, Thomas Serre, and Ellie Pavlick. Uncovering intermediate variables in transformers using circuit probing. *arXiv preprint arXiv:2311.04354*, 2023.

Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Weitao Ma, and Xiachong Feng. Causal tracing of object representations in large vision language models: Mechanistic interpretability and hallucination mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31645–31653, 2026a.

Xinze Li, Yuqing Lan, Zhenghao Liu, Haidong Xin, Yukun Yan, Shuo Wang, Zheni Zeng, Sen Mei, Ge Yu, and Maosong Sun. Seek: Steering llm reasoning for rag via internal reasoning sketches. *arXiv preprint arXiv:2601.09402*, 2026b.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, volume 2024, pp. 26486–26515, 2024.

Maisha Maliha and Dean F Hougen. Mechanistic interpretability of text-to-image diffusion models via cross-attention interventions. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25287–25299, 2026.

Samuel Marks, Can Rager, Eric Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations*, volume 2025, pp. 23888–23923, 2025.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Jungwon Park, Jungmin Ko, Dongnam Byun, Jangwon Suh, and Wonjong Rhee. Cross-attention head position patterns can align with human visual concepts in text-to-image generative models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International conference on machine learning*, pp. 4055–4064. PMLR, 2018.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Yingdong Shi, Changming Li, Yifan Wang, Yongxiang Zhao, Anqi Pang, Sibei Yang, Jingyi Yu, and Kan Ren. Dissecting and mitigating diffusion bias via mechanistic interpretability. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8192–8202. IEEE, 2025.

Zifan Song, Xiao Gong, Guosheng Hu, Shuguang Dou, Qingsong Zhao, and Cairong Zhao. Dpl++: Advancing the network performance via image and label perturbations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Zifan Song, Kaitao Song, Guosheng Hu, Ding Qi, Junyao Gao, Xiaohua Wang, Dongsheng Li, and Cai Zhao. Trade in minutes! rationality-driven agentic system for quantitative financial trading. In *International Conference on Learning Representations*, volume 2026, pp. 27140–27156, 2026.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416, 2024.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Türe. What the daam: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5644–5659, 2023.

Berk Tinaz, Zalan Fabian, and Mahdi Soltanolkotabi. Emergence and evolution of interpretable concepts in diffusion models. *Advances in neural information processing systems*, 38:166943, 2025.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. Function vectors in large language models. In *International conference on learning representations*, volume 2024, pp. 17282–17333, 2024.

Michael Toker, Hadas Orgad, Mor Ventura, Dana Arad, and Yonatan Belinkov. Diffusion lens: Interpreting text encoders in text-to-image pipelines. *arXiv preprint arXiv:2403.05846*, 2024.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

Binxu Wang, Jingxuan Fan, and Xu Pan. Circuit mechanisms for spatial relation generation in diffusion transformers. *arXiv preprint arXiv:2601.06338*, 2026.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Jing Xiao, Minghai Jiao, Peng Xiao, Ende Zhang, Shuang Kan, Wenyan Jiang, Jinyao Li, Yixian Liu, and Haidong Xin. Cagmamba: Context-aware gated cross-modal mamba network for multimodal sentiment analysis. In *International Conference on Intelligent Computing*, pp. 299–311. Springer, 2026.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Arman Zarei, Samyadeep Basu, Keivan Rezaei, Zihao Lin, Sayan Nag, and Soheil Feizi. Localizing knowledge in diffusion transformers. *Advances in Neural Information Processing Systems*, 38: 14584–14612, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.