DON'T IGNORE THEM: SCAFFOLD FEATURES MATTER FOR STABLE SAE STEERING IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoders (SAEs) enable concept editing in diffusion models by manipulating internal features, and existing methods typically select editing targets based on the target semantics. In contrast, we identify a class of features based on reuse across prompts, without relying on target semantics. Their effects vary with the generated content and are difficult to characterize with a single concept, yet they respond strongly to concept editing. Blocking their responses impairs editing results more than blocking the responses of matched ordinary features. These results show that these features are not high-frequency background unrelated to the target edit, but important internal variables that carry shared generative context. We therefore call them scaffold features. We further incorporate scaffold features into editing and propose Scaffold-Coordinated Feature Editing (SCFE), a method built on existing SAEs that requires no additional training. Using small concept interventions, SCFE estimates spatial responses for the current prompt and adds constrained scaffold-feature updates along the corresponding directions while manipulating the target concept. Experiments show that, under stronger editing, SCFE both increases target-semantic change and improves overall image preservation, and these gains generalize across diffusion-model architectures and their associated SAEs. These findings show that a feature's value for concept editing is not limited to the semantics it represents, as scaffold features selected without relying on target semantics can also improve targeted editing.

1 INTRODUCTION

Sparse autoencoders (SAEs) decompose dense activations inside diffusion models into sparse features, enabling researchers to identify interpretable visual concepts within complex representations (Surkov et al., 2025; Kim et al., 2025; Shabalin et al., 2025). These features also provide a direct control interface for concept editing. Enhancing or suppressing target-associated features can change image attributes, style, and composition during generation (Tinaz et al., 2025; Kim & Ghadiyaram, 2025; Kamenetsky et al., 2026). This approach requires no retraining of the diffusion model and connects editing operations to specific internal representations. It therefore allows researchers not only to control generated outputs but also to analyze how the model realizes semantic changes through feature interventions.

However, features with clear semantics are not necessarily ideal control variables for editing. Prior work on object erasure finds that directly intervening on SAE latents can introduce out-of-distribution activations and produce visible artifacts (Cassano et al., 2026a). Related studies also show that target interventions can affect associated non-target representations, and that activation steering applied without distinguishing the generated content can degrade images that do not require modification (Shi et al., 2026; Chrabąszcz et al., 2026). These findings distinguish identifying a target concept from effectively editing it. What a feature represents provides a basis for selecting an editing direction, but how to manipulate it also requires considering the intervention's effects on the rest of the generation process.

Yet we observe a counterintuitive phenomenon: features selected without using target semantics also play an important role in realizing those semantics in generated images. Unlike approaches that locate editing features around a given concept (Cywiński & Deja, 2025; Kim & Ghadiyaram, 2025), we use the statistical property of cross-prompt reuse to identify shared features that recur

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

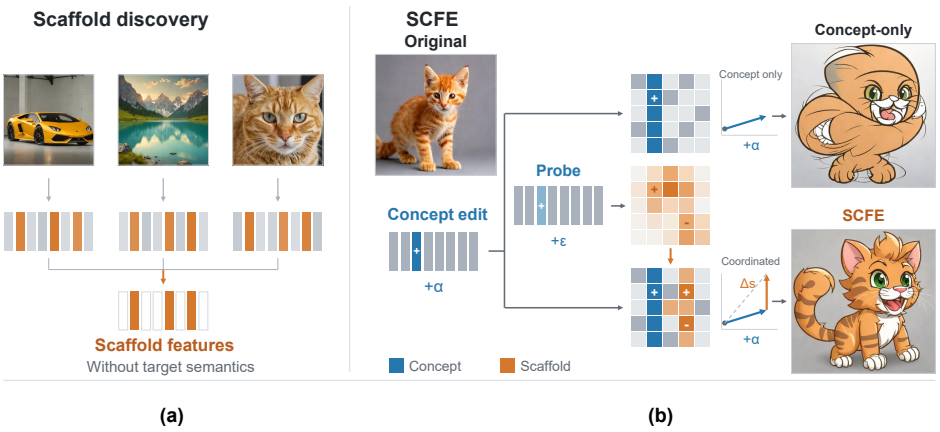


Figure 1: Discovery of scaffold features and Scaffold-Coordinated Feature Editing (SCFE). (a) Different generated content repeatedly invokes a shared set of features; scaffold-feature selection does not rely on target semantics. (b) SCFE uses a small concept intervention ε to estimate reference scaffold-feature responses for the current prompt, then adds constrained scaffold-feature updates Δs while retaining the concept-edit strength α. Blue denotes the target concept, and orange denotes scaffold features.

across different generated content. Figure 1(a) illustrates this pattern of cross-content reuse. These features have content-dependent effects on materials, layout, or local structure that are difficult to summarize with a single concept, yet they respond strongly to concept editing. Along both same-step and cross-step paths, scaffold-response blocking produces larger mean output changes than matched ordinary-response blocking. Image comparisons also show more pronounced structural and content damage. Cross-content reuse, content-dependent effects, and the causal-intervention experiments together support interpreting these features as important internal variables that carry shared generative context. We therefore call them scaffold features.

These observations lead us to reconsider the choice of SAE editing variables: features beyond the target are not only representations to protect from collateral damage, but can also serve as control variables for actively improving edits. Prior work constrains intervention directions or transforms activation distributions to control target concepts while reducing non-target effects (Rodríguez et al., 2025; Shi et al., 2026; Chrabąszcz et al., 2026). We propose Scaffold-Coordinated Feature Editing (SCFE), which builds on existing SAE-based concept editing to perform joint updates using scaffold responses to the current concept intervention. Given a pretrained diffusion model and its associated SAE, SCFE updates neither model's parameters and trains no additional modules. As shown in Figure 1(b), the method first uses small concept interventions to estimate spatial scaffold-feature responses for the current prompt. During the actual edit, it manipulates the target concept and adds constrained scaffold-feature updates along the corresponding directions. Using the SAEs released by Surkov et al. (2025), we define our baseline as editing only the target concept feature without adding scaffold updates, termed Concept-only. SCFE adds scaffold-response updates on top of this baseline. The anime-style editing comparison in the figure illustrates the effect of this joint update: at the same concept-edit strength, SCFE reduces damage to the subject's structure relative to concept-only editing and more clearly realizes the anime-style semantics. Thus, scaffold-feature identification does not depend on target semantics, whereas the update is determined by the response to the current edit. This allows us to use their content-dependent effects to improve editing without assigning them a single concept label.

Following prior editing work's joint attention to semantic control and content preservation (Surkov et al., 2025; Kim & Ghadiyaram, 2025; Kamenetsky et al., 2026), we evaluate both target-semantic change and overall image preservation. Experiments show that SCFE improves both under stronger edits; replacing features or disrupting response correspondence with the current prompt or spatial positions fails to reproduce comparable gains. Beyond SDXL-Turbo, we reidentify scaffolds and estimate responses in FLUX.1-schnell and its public SAEs (Shabalin et al., 2025). The new model also contains a scaffold population that is highly reused across prompts and participates in concept

editing. As Table 1 shows, SCFE improves both target semantics and image preservation over concept-only editing in this configuration. These results show that statistically identified shared features not only participate in generation but can also improve targeted editing across models; their value is not limited to explicit semantic labels.

2 Related Work

2.1 Sparse Feature Decompositions for Diffusion Models

Sparse autoencoders offer a feature-level alternative to individual neural coordinates, which can mix several behaviors (Huben et al., 2024). In text-to-image diffusion models, k-SAE and dictionary-learning analyses recover sparse directions from U-Net, diffusion-transformer, and text-conditioning activations. They use top-activating examples, layer–timestep comparisons, and causal perturbations to connect directions with visual behavior (Kim et al., 2025; Shabalin et al., 2025; Tinaz et al., 2025; Surkov et al., 2025). Temporal SAE work tracks feature structure across denoising steps. Circuit analyses reconstruct dependencies among sparse features (Huang et al., 2026; Yeung et al., 2026; Mazur et al., 2026).

Sparse representations also support diffusion-model audits. They expose concept-prevalence gaps and identify rare attributes. They also reveal non-target semantic drift that aggregate quality metrics can miss (Bohacek et al., 2025; Vadakkeeveetil Sreelatha et al., 2026; Laria et al., 2026). These studies often select a feature through an output behavior, a target concept, or a visually legible example set. Our recurrence screen selects a population before those choices. It separates cross-content recurrence from later role analysis.

2.2 SAE-Based Concept Editing and Unlearning

Sparse feature spaces enable inference-time control without retraining the diffusion model. At the text-conditioning interface, Concept Steerers, SAEdit, and ELROND control concepts, attributes, and stochastic visual realizations with sparse or decomposed directions (Kim & Ghadiyaram, 2025; Kamenetsky et al., 2026; Skiers et al., 2026). Interpret-then-Deactivate selects target-associated features in the text encoder for concept erasure (Tian et al., 2025). SNCE and SAE Debias construct localized concept or bias interventions in text-conditioning representations (He et al., 2025; Wu et al., 2025). Within the denoising network, DIFFLENS attributes and mitigates demographic bias (Shi et al., 2025); SAeUron selects target-associated features for concept erasure (Cywiński & Deja, 2025); SafeDIG selects position-aware safety interventions in diffusion transformers (Xue et al., 2026).

Several methods sharpen target selection through supervised or concept-aware feature separation. SAEmnesia concentrates a target concept into a controllable latent (Cassano et al., 2026b). SAEParate reduces feature sharing with a concept-aware objective (Kim et al., 2026). Supervised sparse feature models allocate structured concept subspaces for compositional editing (El Harzli et al., 2026). OrthoEraser protects a coupled benign subspace during target erasure (Shi et al., 2026). Scaffold selection identifies a recurrent population without using target semantics. We then test its role in target editing through concept interventions and response blocking.

2.3 Distribution-Preserving Editing and Feature Coupling

Activation Transport learns transformations between activation distributions (Rodríguez et al., 2025). Conditioned Activation Transport restricts nonlinear transport to unsafe activation regions (Chrabąszcz et al., 2026). Look But Don't Touch reports that direct SAE-latent interventions can create out-of-distribution activations. It uses SAE features to localize target regions before replacing patch embeddings with in-distribution alternatives (Cassano et al., 2026a). These methods distinguish target detection from a valid editing operation.

Feature coupling offers a local view of the same challenge. OrthoEraser protects a coupled benign subspace. Temporal and circuit analyses show that feature effects depend on layer, timestep, and interactions with other features (Shi et al., 2026; Huang et al., 2026; Yeung et al., 2026; Mazur et al., 2026). By comparing generations before and after model adaptation, DriftScope reveals non-target

concept drift that aggregate quality metrics may miss (Laria et al., 2026). We combine cross-prompt recurrence with response interventions to study how shared features participate in concept editing and use them for joint updates.

3 Discovery and Role of Scaffold Features

3.1 A Feature Population Reused Across Prompts

We identify shared features through the statistical property of cross-prompt reuse, without specifying a target concept in advance. Using the public pretrained SDXL-Turbo SAEs released by Surkov et al. (2025), we measure feature activations across four network modules and four denoising steps for 10,000 LAION-COCO prompts (Schuhmann et al., 2022). Appendix A details the SAE configuration, prompt sampling, and generation settings. Appendix B examines the content diversity of the prompts and generated images. Let $a_{i,\ell,t,k,s}$ denote the activation of feature k at spatial position s in module ℓ and denoising step t under prompt i. We define cross-prompt prevalence and the scaffold set as

$$\pi_{\ell,t,k} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\exists s : a_{i,\ell,t,k,s} \neq 0], \qquad H = \{(\ell,t,k) : \pi_{\ell,t,k} \geq 0.9\}. \tag{1}$$

Here, N is the number of prompts, and H retains feature instances that occur in at least 90% of prompts. Each instance corresponds to a combination of a module, a denoising step, and a feature index. We refer to this group of features repeatedly used across content as scaffold features.

As shown in Figure 2(a), scaffold features occur in every module and denoising step examined. The prevalence curve in Figure 2(b) further shows that some features occur in almost every prompt. This frequent reuse cannot be entirely attributed to responses to fixed inputs. Figure 2(c) shows the activations obtained by feeding zero vectors, global means, and position-wise means into the SAEs: the three inputs activate only part of the selected set, and some scaffold features remain inactive under all three. These fixed inputs are insufficient to explain cross-prompt reuse across the entire scaffold population. Appendix C provides supplementary analyses of the effects of spatial-coverage requirements and sample size, fixed-input responses, and spatial-location variation.

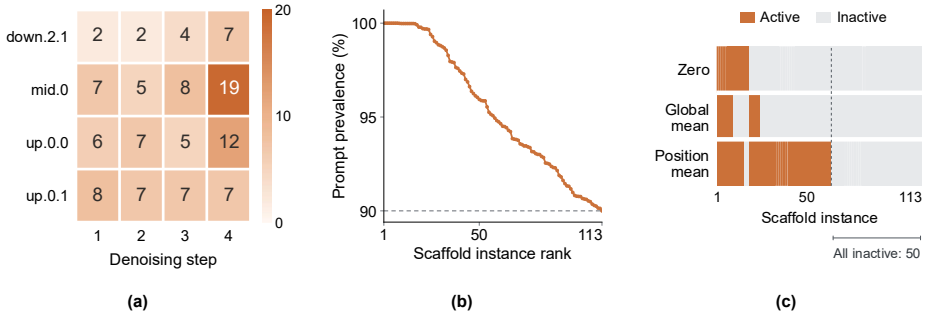


Figure 2: **Scaffold features reused across prompts and fixed-input controls.** (a) Number of selected instances in each module and denoising step. (b) Cross-prompt prevalence of the 113 selected instances, sorted in descending order; the dashed line marks the 90% selection threshold. (c) Activation of the same selected set under zero, global-mean, and position-mean inputs. Both means are computed from the natural SAE inputs for the prompts above. Columns are grouped by their activation patterns under the three inputs, so their order differs from (b). Orange indicates nonzero activation at one or more spatial positions; gray indicates no activation. Instances to the right of the dashed line are inactive under all three inputs.

3.2 Scaffold Features Are Difficult to Characterize with a Single Concept

Reuse across prompts does not mean that scaffold features always correspond to the same visual concept. Following the visualization method of Surkov et al. (2025), we upsample natural activation

maps and overlay them on generated images. As shown in Figure 3(b), the activation regions of the same feature involve objects, local structures, and backgrounds across different content. The interventions in Figure 3(c) further illustrate differences in their effects: although `up.0.1/f2727` exhibits a shared brightness trend, it also changes the surface appearance of interior furnishings and a backpack; `up.0.0/f1985` changes furniture and furnishings in an indoor scene, but local structures such as buckles and pockets in a backpack image. These features can have recognizable shared trends, yet their specific effects vary with the generated content and are difficult to fully characterize with a single concept.

We also compare the changes in image representations caused by enhancing or suppressing all 62 scaffold features across prompts. As shown in Figure 3(a), the directions of change for most features lack cross-prompt consistency in both CLIP and DINOv2 representations, while a few exhibit stronger shared trends. These population-level statistics complement the examples above, showing that cross-prompt reuse is not accompanied by generally consistent editing effects. Appendix D provides the computation details and per-feature results.

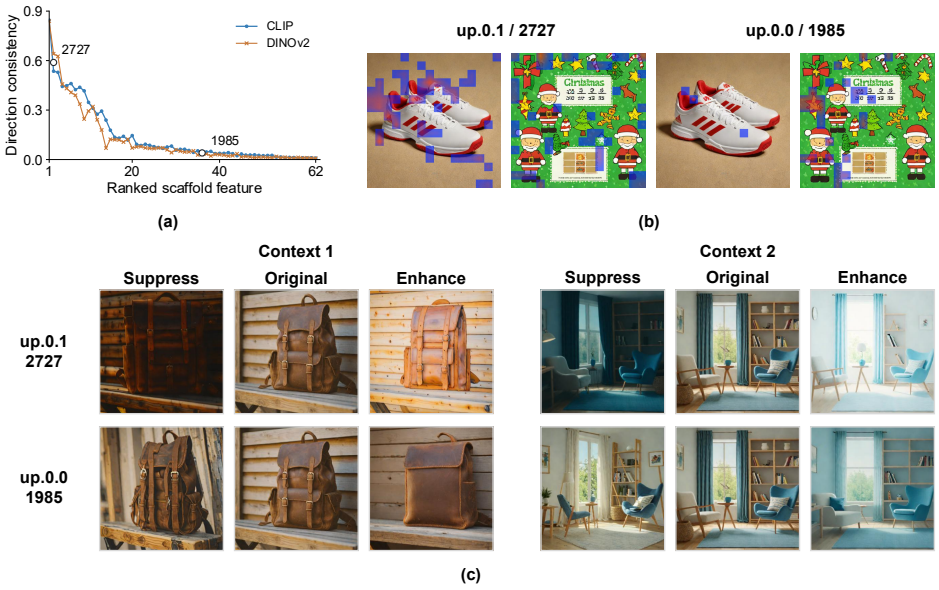


Figure 3: **Content-dependent effects of scaffold features.** (a) Cross-prompt edit-direction consistency for 62 module–feature pairs on 10,000 prompts. Directions are estimated from strength sweeps with interventions at all spatial positions and all four steps; consistency is the mean pairwise cosine similarity of normalized directions. The horizontal axis is sorted by the mean of CLIP and DINOv2 consistency. Open circles mark the two features shown below at their mean consistency. (b) Natural activation overlays for two features under different prompts, with colors normalized separately for each image. (c) Suppressed, unedited, and enhanced results for an indoor scene and a backpack; each comparison uses the same prompt and random seed. Panels (b) and (c) use different example prompts to show spatial activation and visible editing effects, respectively.

3.3 The Causal Role of Scaffold Responses in Concept Editing

We intervene on concept features across 10,000 prompts, normalize subsequent responses by their natural activation scales, and compare them with ordinary features in the same module and timestep. As shown in Figure 4(b), scaffold responses concentrate in the upper region of the ordinary-feature response distribution, indicating that changing the target concept also recruits these shared features.

We test the causal role of these responses using the restoration intervention in Figure 4(a). On the path from `down.2.1` to `mid.0` within the same denoising step, we keep the upstream concept intervention unchanged and restore either selected scaffold coordinates or matched ordinary coordinates to their unedited reference values, blocking the corresponding changes caused by the concept

intervention. Ordinary coordinates come from the same module and timestep and are matched in number and decoder-space response strength.

Using concept-only images as the reference, Figure 4(d) shows that blocking scaffold responses causes larger output changes than blocking matched ordinary responses, with a consistent average relationship across the three image-distance metrics. The cross-step path shows the same average trend; full results are provided in Appendix E. Thus, scaffold responses have a greater influence on editing results even after matching the number of coordinates and decoder-space response strength.

These effects are not limited to the strength of target-concept changes. Figure 4(c) shows damage to non-target structure: blocking scaffold responses detaches the left flower from its stem, whereas blocking matched ordinary responses preserves their connection. The semantic evaluation in Appendix E.4 further shows that blocking scaffold responses weakens target-semantic changes on average compared with blocking matched ordinary-feature responses. This indicates that scaffold responses not only help preserve image content but also contribute to realizing the target concept. Image comparisons also show cases where the target color change remains clear but the object's structure is damaged. Together with cross-prompt reuse and content-dependent effects, these results support interpreting scaffold features as internal variables that carry shared generative context: they not only respond to changes in the target concept, but also help preserve image structure and content relationships during editing.

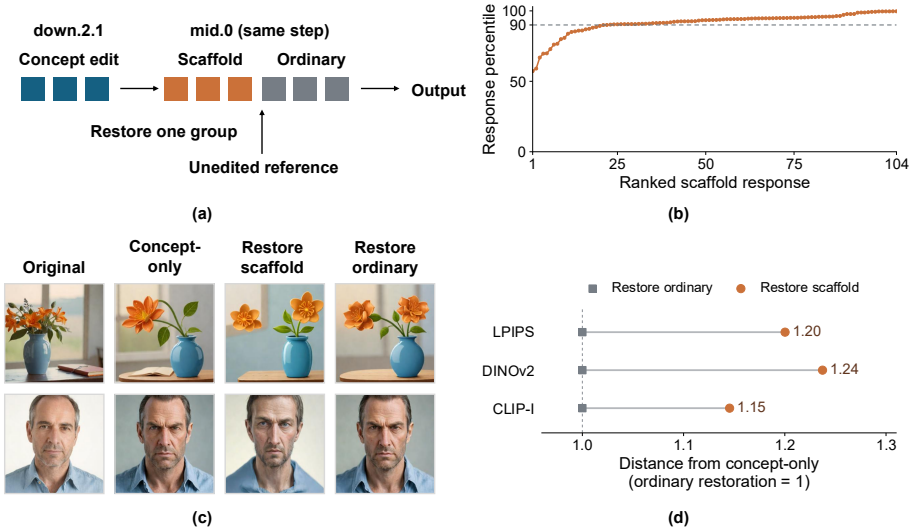


Figure 4: **Scaffold responses during concept editing and same-step blocking controls.** (a) Keeping the upstream concept intervention unchanged, scaffold or matched ordinary coordinates at a downstream position within the same step are restored to their unedited reference values; the number of colored blocks is schematic. (b) Response percentiles of 104 scaffold-response items relative to ordinary features for the same source feature and readout position, across 10,000 prompts and edit strengths $0 < |s| \leq 20$. Higher percentiles indicate stronger responses; the dashed line marks the 90th percentile. This panel includes both same-step and cross-step paths. (c) Separately generated vase and portrait editing examples. Each row shows the unedited, concept-only, scaffold-restored, and ordinary-restored results, in that order. (d) Aggregate comparison on the same-step path from `down.2.1` to `mid.0`, covering 10,000 prompts, two concept features, and the strength range above. Image distances use the concept-only result as the reference, with the mean distance under ordinary restoration normalized to 1.

4 Scaffold-Coordinated Feature Editing

Section 3 and Appendix E show that scaffold features respond strongly to concept edits. On both tested paths, blocking scaffold responses causes larger mean output changes than blocking ordinary responses matched in number and response strength. Thus, an edit is shaped not only by the

concept feature being manipulated, but also by the scaffold responses that follow it. We propose Scaffold-Coordinated Feature Editing (SCFE), which estimates the local mapping from concept edits to scaffold responses and constrains additional updates by natural activation scales. SCFE estimates responses and applies joint updates between SAE positions with a clear execution order within the same denoising step.

Let C be the set of concept-control coordinates to edit and H the scaffold-candidate set from Section 3.1. The concept edit and scaffold activations are denoted by $\Delta z_C \in \mathbb{R}^{d_C}$ and $z_H \in \mathbb{R}^{d_H}$, respectively. For compact notation, spatial positions, SAE positions, and injection times are included in scaffold-coordinate indices; each concept-control coordinate specifies a spatially uniform feature increment at a given timestep. For prompt p and concept coordinate c, we fix the initial noise, apply symmetric interventions of magnitude ϵ along the unedited reference trajectory, and estimate the downstream response by a central finite difference:

$$G_p[:, c] = \frac{z_{p,H}^{(+\epsilon e_c)} - z_{p,H}^{(-\epsilon e_c)}}{2\epsilon}, \qquad G_p \in \mathbb{R}^{d_H \times d_C}. \tag{2}$$

Here, e_c is the unit vector for concept coordinate c. Entry $G_p[h, c]$ approximates the response of scaffold coordinate h to a unit change in c near the reference generative state for the current prompt. As an intervention difference, it is distinct from ordinary correlation between concept and scaffold features.

Across P calibration prompts, we compute each candidate scaffold feature's spatially averaged unit response $s_{q,f}$ and its cross-prompt mean $\bar{s}_f = P^{-1} \sum_q s_{q,f}$. These averages are used only to select features for joint editing. For the requested concept edit Δz_C, the predicted scaffold response for the current prompt and the cross-prompt sign consistency used for selection are

$$r_p = G_p \Delta z_C, \qquad \rho_f = \left| \frac{1}{|\mathcal{V}_f|} \sum_{q \in \mathcal{V}_f} \operatorname{sign}(s_{q,f}) \right|. \tag{3}$$

Here, $\mathcal{V}_f$ contains valid calibration prompts after excluding near-zero responses, and ρ_f measures whether feature f responds in the same direction across prompts. SCFE retains features with $\rho_f \geq \tau_s$ and selects the K largest $|\bar{s}_f|$ values; their spatial coordinates form the accompanying set H_C. This rule selects reusable features from cross-prompt statistics without adding every scaffold feature indiscriminately. In the actual update, r_p retains the current prompt's complete spatial response, without averaging across prompts or space. Selection is performed separately for each concept and denoising step, as detailed in Appendix F.

After the concept edit propagates through the network, the current scaffold state at the specified downstream location is z_H^{pre}. Here, *pre* denotes the state before this scaffold update, which may include effects of scaffold updates at earlier timesteps. SCFE starts from this state. Rather than restoring a response absent from the trajectory, it applies a controllable gain along the probed response direction. Let $D_H = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_{d_H}^2)$, where σ_h is the activation scale of scaffold coordinate h in natural generations. SCFE scales the predicted response by λ and bounds the additional update using natural activation scales.

The constraints below use the nonnegative SDXL-Turbo SAEs released by Surkov et al. (2025) as the reference case. For the signed FLUX SAE, we remove only the activation nonnegativity constraint and retain feature selection and amplitude bounds; see Appendix H.

$$u^\star = \arg\min_{u \in \mathcal{U}} \|u - \lambda r_p\|^2_{D_H^{-1}}, \tag{4}$$

$$\mathcal{U} = \{u : \operatorname{supp}(u) \subseteq H_C,\ |u_h| \leq \kappa \sigma_h,\ z_H^{\text{pre}} + u \geq 0\}, \qquad z_H^{\text{joint}} = z_H^{\text{pre}} + u^\star. \tag{5}$$

The projection keeps the executable update closest to the predicted response under three constraints: it changes only selected accompanying scaffold coordinates, bounds each update by κ times its natural activation scale, and preserves nonnegative SAE activations. Under the diagonal-scale approximation, coordinatewise clipping solves the optimization without iteration. Setting $\lambda = 0$ recovers concept-only editing; positive λ adds updates along the current prompt's local response direction. Parameters κ, K, and τ_s control update amplitude, the number of jointly edited features, and cross-prompt stability, respectively.

In summary, given a pretrained diffusion model and SAE, SCFE updates neither model's parameters and trains no additional prediction module. Response maps must be estimated for the requested prompt before generation; they can be reused across edit strengths for the same prompt, initial noise, and target concept. During generation, constrained scaffold-coordinate updates are applied to the current state. SCFE maps concept edits into the downstream response directions identified in Section 3, then limits the intervention through sparse projection.

5 EXPERIMENTS

We evaluate whether joint editing improves concept steering and whether the improvement depends on the correct scaffold response. Our main experiments pair SDXL-Turbo with the public pretrained SDXL-Turbo SAEs released by Surkov et al. (2025) under identical prompts, initial noise, and concept-edit strengths. We also test transfer to FLUX.1-schnell with the SAEs of Shabalin et al. (2025). Specific concepts, intervention locations, and evaluation settings are detailed in Appendices G.1 and H.1.

Table 1: Editing results for two diffusion-model and SAE configurations. Within each configuration, methods share prompts, initial noise, and concept-injection strength, then average over the tested concepts, directions, and strengths. Preservation is relative to the unedited image; GPT-5.6-Sol scores target-directed change from image pairs. Bold marks the best listed result within each configuration. Appendix H.4 gives the full six-method FLUX comparison.

Method	LPIPS ↓	DINOv2 dist. ↓	CLIP-I ↑	Semantics ↑
SDXL-Turbo SAEs released by Surkov et al. (2025)				
Concept-only	0.341	0.444	0.869	6.817
Random-feature joint editing	0.572	0.816	0.708	4.844
Responsive non-scaffold editing	0.415	0.524	0.856	6.727
Prompt mismatch	0.625	0.747	0.613	4.224
Spatial shift	0.536	0.724	0.745	5.477
SCFE	**0.300**	**0.282**	**0.895**	**7.049**
FLUX.1-schnell SAEs released by Shabalin et al. (2025)				
Concept-only	0.300	0.209	0.868	5.449
SCFE	**0.265**	**0.179**	**0.880**	**5.934**

5.1 EFFECTIVENESS OF JOINT EDITING

Overall, SCFE improves image preservation and target-semantic scores at the same concept-injection strength. As Table 1 shows, SCFE achieves lower LPIPS and DINOv2 distance and higher CLIP-I similarity than Concept-only. The semantic improvement indicates that the preservation gains do not merely come from weakening the target edit.

SCFE's average gains are more pronounced for stronger edits. In Figure 5, the methods have similar mean semantic scores and CLIP-I similarities at low strengths; SCFE's joint gains in semantics and preservation become clearer as strength increases. The editing examples also show that joint updates can reduce the loss of subject parts, local structure, and existing content when changing style or a person's appearance.

This effect also occurs with a different diffusion architecture and SAE. For FLUX.1-schnell, we identify scaffolds anew using the same statistical rule and re-estimate concept responses rather than reusing SDXL feature indices. Table 1 shows improvements in target-semantic change and all three preservation metrics over concept-only editing in this configuration.

5.2 JOINT EDITING REQUIRES THE CORRECT FEATURE RESPONSE

The gains depend on feature identity and the correspondence between responses and current content. We use four controls: random-feature joint editing, responsive non-scaffold-feature joint editing, prompt mismatch, and spatial shift. The first two replace the updated features; the latter two retain the scaffold set but alter the prompt source or spatial positions of its response.

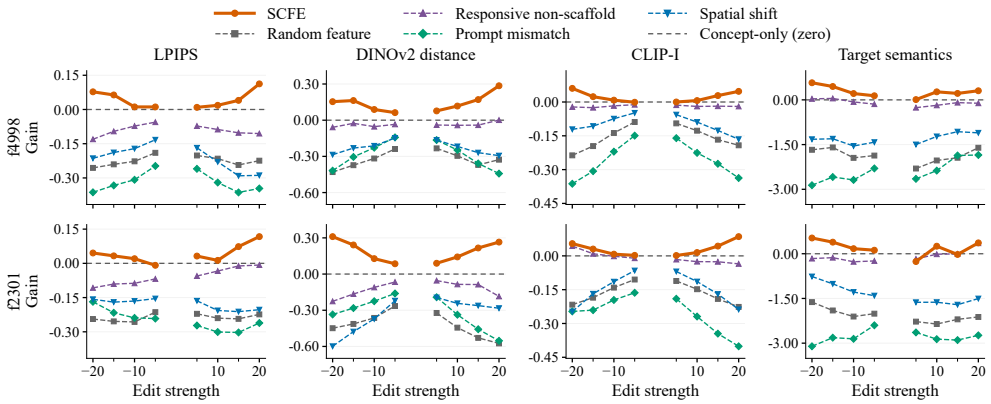


Figure 5: Editing gains over Concept-only. Rows correspond to `f4998` and `f2301`; columns show LPIPS, DINOv2 distance, CLIP-I, and target-semantic score. The first two report Concept-only minus method; the last two report method minus Concept-only, so positive values indicate improvement. Points average over prompts at each signed strength; dashed lines mark zero gain.

As shown in Figure 6, SCFE outperforms Concept-only and all controls on all three image-preservation metrics. Replacing scaffold features or disrupting the prompt or spatial correspondence of their responses not only removes SCFE's preservation gains but also yields worse average performance than Concept-only on all three metrics. These controls show that scaffold identity and spatial responses to the current content are key to the gains from joint editing. SCFE benefits from the targeted use of these shared features rather than arbitrary additional updates.

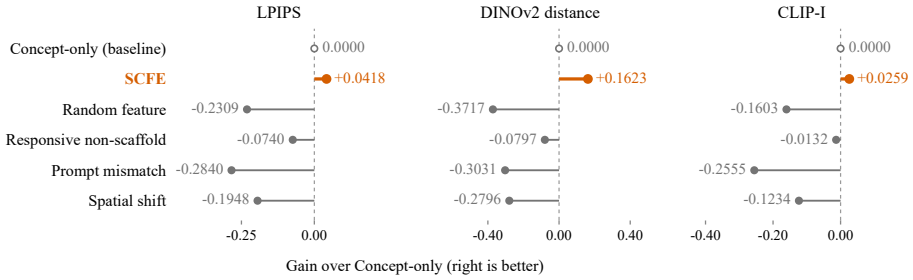


Figure 6: Mean preservation gains over Concept-only, averaged across all paired evaluation conditions. Gain definitions follow Figure 5. Columns show LPIPS, DINOv2 distance, and CLIP-I; positive values indicate improvement. Dots and labels report mean gains; horizontal segments connect them to zero.

6 Conclusion

This work shows that a clear concept label is not a prerequisite for a feature to have editing value. We identify scaffold features through cross-prompt reuse and find that, although difficult to characterize with a single concept, they participate in realizing target concepts; blocking their responses impairs editing results more than blocking matched ordinary-feature responses. SCFE turns this finding into a joint-editing strategy, using prompt-specific scaffold responses to strengthen target-semantic change while improving overall image preservation under stronger edits. Experiments across diffusion architectures and SAEs also validate the strategy's ability to generalize. These results broaden the focus of concept editing from target-semantic features alone to shared features that jointly participate in generation, showing how statistical discovery and response interventions can together reveal and exploit the editing value of internal representations.

AI USE STATEMENT

Generative AI assisted literature organization, English-language editing, and code development. We also used GPT-5.6-Sol to evaluate target-semantic changes, as described in the experimental sections and appendices. We reviewed all AI-generated content, including text and code, and take full responsibility for the paper's content and research results.

ETHICS STATEMENT

Feature-level control can support model inspection and safety work, and it can alter generated content.

REPRODUCIBILITY STATEMENT

The main text and appendices document the sources of the diffusion models and pretrained SAEs, prompt sampling procedures, scaffold-selection rules, intervention settings, and evaluation methods.

REFERENCES

Matyas Bohacek, Thomas Fel, Maneesh Agrawala, and Ekdeep Singh Lubana. Uncovering conceptual blindspots in generative image models using sparse autoencoders. *arXiv preprint arXiv:2506.19708*, 2025. URL https://arxiv.org/abs/2506.19708.

Enrico Cassano, Riccardo Renzulli, Rayyan Ahmed, Marco Grangetto, and Stephan Alaniz. Look but don't touch with sparse autoencoders for unlearning in diffusion models. *arXiv preprint arXiv:2606.31699*, 2026a. URL https://arxiv.org/abs/2606.31699. Accepted at ECCV 2026.

Enrico Cassano, Riccardo Renzulli, Marco Nurisso, Mirko Zaffaroni, Alan Perotti, and Marco Grangetto. SAEmnesia: Erasing concepts in diffusion models with supervised sparse autoencoders. *arXiv preprint arXiv:2509.21379*, 2026b. URL https://arxiv.org/abs/2509.21379. Accepted at ICML 2026.

Maciej Chrabąszcz, Aleksander Szymczyk, Jan Dubiński, Tomasz Trzciński, Franziska Boenisch, and Adam Dziedzic. Conditioned activation transport for T2I safety steering. *arXiv preprint arXiv:2603.03163*, 2026. URL https://arxiv.org/abs/2603.03163.

Bartosz Cywiński and Kamil Deja. SAeUron: Interpretable concept unlearning in diffusion models with sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025. URL https://arxiv.org/abs/2501.18052.

Ouns El Harzli, Hugo Wallner, Yoonsoo Nam, and Haixuan Xavier Tao. Supervised sparse autoencoders for interpretable and compositional representations. *arXiv preprint arXiv:2602.00924*, 2026. URL https://arxiv.org/abs/2602.00924.

Thomas Fel, Matthew Kowal, Mozes Jacobs, Dron Hazra, Usha Bhalla, Lee Sharkey, Lucius Bushnaq, Satchel Grant, Tal Haklay, Thomas Icard, Can Rager, Michael Pearce, Daniel Wurgaft, Aiden Swann, Fenil Doshi, Siddharth Boppana, Curt Tigges, Nick Cammarata, Thomas Serre, Vasudev Shyam, Owen Lewis, Thomas McGrath, Jack Merullo, Ekdeep Singh Lubana, and Atticus Geiger. Structuring sparsity: Block-sparse featurizers capture visual concept manifolds. *arXiv preprint arXiv:2606.25234*, 2026. doi: 10.48550/arXiv.2606.25234. URL https://arxiv.org/abs/2606.25234.

Dan Friedman and Adji Bousso Dieng. The Vendi Score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=g97OHbQyk1.

Qinqin He, Jiaqi Weng, Jialing Tao, and Hui Xue. A single neuron works: Precise concept erasure in text-to-image diffusion models. *arXiv preprint arXiv:2509.21008*, 2025. URL https://arxiv.org/abs/2509.21008.

Phu Gia Hoang, Anwoy Chatterjee, Tanmoy Chakraborty, Iryna Gurevych, and Subhabrata Dutta. Sparse autoencoders encode both concepts and functions: The downstream geometry of feature effects. *arXiv preprint arXiv:2607.24645*, 2026. doi: 10.48550/arXiv.2607.24645. URL https://arxiv.org/abs/2607.24645.

Victor Shea-Jay Huang, Le Zhuo, Yi Xin, Zhaokai Wang, Fu-Yun Wang, Yuchi Wang, Renrui Zhang, Peng Gao, and Hongsheng Li. TIDE: Temporal-aware sparse autoencoders for interpretable diffusion transformers in image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 435–443, 2026. doi: 10.1609/aaai.v40i1.37006. URL https://ojs.aaai.org/index.php/AAAI/article/view/37006.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1fa1ab11f4bd5f94b2ec20e794dbfa3b-Abstract-Conference.html.

Mohammad Jalali, Cheuk Ting Li, and Farzan Farnia. An information-theoretic evaluation of generative models in learning multi-modal distributions. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0434. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1f5c5cd01b864d53cc5fa0a3472e152e-Abstract.html.

Ronen Kamenetsky, Sara Dorfman, Daniel Garibi, Roni Paiss, Or Patashnik, and Daniel Cohen-Or. SAEdit: Token-level control for continuous image editing via sparse autoencoder. In *Computer Vision – ECCV 2026*, volume 17055 of *Lecture Notes in Computer Science*, pp. 359–377. Springer, 2026. doi: 10.1007/978-3-032-37261-1_21. URL https://link.springer.com/chapter/10.1007/978-3-032-37261-1_21.

Dahye Kim and Deepti Ghadiyaram. Concept steerers: Leveraging k-sparse autoencoders for test-time controllable generations. *arXiv preprint arXiv:2501.19066*, 2025. URL https://arxiv.org/abs/2501.19066.

Dahye Kim, Xavier Thomas, and Deepti Ghadiyaram. Revelio: Interpreting and leveraging semantic information in diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4659–4669, 2025. URL https://arxiv.org/abs/2411.16725.

Hyeonjin Kim, Hangyeol Jung, Heechan Yun, Sungjun Yun, and Dong-Jun Han. Disentangled sparse representations for concept-separated diffusion unlearning. *arXiv preprint arXiv:2605.12122*, 2026. URL https://arxiv.org/abs/2605.12122.

Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. ImagenHub: Standardizing the evaluation of conditional image generation models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/cbdeb402cd9adc39f5c50725c961b552-Paper-Conference.pdf.

Héctor Laria, Yiping Han, Julian D. Santamaria, Kai Wang, Bogdan Raducanu, Joost van de Weijer, and Alexandra Gomez-Villa. DriftScope: Measuring the hidden effects of diffusion model adaptation. In *Computer Vision – ECCV 2026*, volume 17024 of *Lecture Notes in Computer Science*, pp. 58–78. Springer, 2026. doi: 10.1007/978-3-032-37556-8_4. URL https://link.springer.com/chapter/10.1007/978-3-032-37556-8_4.

Artyom Mazur, Nina Konovalova, and Aibek Alanov. DifFRACT: Diffusion feature reconstruction and attribution for circuit tracing. *arXiv preprint arXiv:2606.15796*, 2026. URL https://arxiv.org/abs/2606.15796.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. DreamBench++: A human-aligned benchmark for personalized image generation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/71ad539a57b1fd49b19e5c80070cb8b9-Paper-Conference.pdf.

Pau Rodríguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Marco Cuturi, and Xavier Suau. Controlling language and diffusion models by transporting activations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=l2zFn6TIQi.

Christoph Schuhmann, Andreas Köpf, Richard Vencu, Theo Coombes, and Romain Beaumont. LAION-COCO: 600m synthetic captions from LAION2B-en. LAION dataset release, 2022. URL https://laion.ai/blog/laion-coco/.

Stepan Shabalin, Ayush Panda, Dmitrii Kharlapenko, Abdur Raheem Ali, Yixiong Hao, and Arthur Conmy. Interpreting large text-to-image diffusion models with dictionary learning. *arXiv preprint arXiv:2505.24360*, 2025. URL https://arxiv.org/abs/2505.24360.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu Edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8871–8879, 2024. URL https://arxiv.org/abs/2311.10089.

Chuancheng Shi, Wenhua Wu, Fei Shen, Xiaogang Zhu, Kun Hu, and Zhiyong Wang. OrthoEraser: Coupled-neuron orthogonal projection for concept erasure. *arXiv preprint arXiv:2603.11493*, 2026. URL https://arxiv.org/abs/2603.11493.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Pawel Skiers, Tomasz Trzcinski, and Kamil Deja. ELROND: Exploring and decomposing intrinsic capabilities of diffusion models. *arXiv preprint arXiv:2602.10216*, 2026. URL https://arxiv.org/abs/2602.10216.

Xiaoqing Sun, Alessandro Stolfo, Joshua Engels, Ben Wu, Senthooran Rajamanoharan, Mrinmaya Sachan, and Max Tegmark. Dense SAE latents are features, not bugs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 140252–140282, 2025. doi: 10.52202/085713-4686. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/cd091a4d8e97157d32940428f902c7b0-Paper-Conference.pdf.

Viacheslav Surkov, Chris Wendler, Antonio Mari, Mikhail Terekhov, Justin Deschenaux, Robert West, Caglar Gulcehre, and David Bau. One-step is enough: Sparse autoencoders for text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2410.22366.

Zhihua Tian, Sirun Nan, Ming Xu, Shengfang Zhai, Wenjie Qu, Jian Liu, Ruoxi Jia, and Jiaheng Zhang. Sparse autoencoder as a zero-shot classifier for concept erasing in text-to-image diffusion models. *arXiv preprint arXiv:2503.09446*, 2025. URL https://arxiv.org/abs/2503.09446.

Berk Tinaz, Zalan Fabian, and Mahdi Soltanolkotabi. Emergence and evolution of interpretable concepts in diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 166943–166986, 2025. doi: 10.52202/085713-5566. URL https://arxiv.org/abs/2504.15473.

Silpa Vadakkeeveetil Sreelatha, Dan Wang, Serge Belongie, Muhammad Awais, and Anjan Dutta. RAIGen: Rare attribute identification in text-to-image generative models. *arXiv preprint arXiv:2602.06806*, 2026. URL https://arxiv.org/abs/2602.06806. Accepted at ICML 2026.

Chris Wendler. SDXL-Turbo SAE Checkpoints. Hugging Face model card, 2025. URL https://huggingface.co/wendlerc/sdxl-turbo-saes.

Chao Wu, Zhenyi Wang, Kangxian Xie, Naresh Kumar Devulapally, Vishnu Suresh Lokhande, and Mingchen Gao. Model-agnostic gender bias control for text-to-image generation via sparse autoencoder. *arXiv preprint arXiv:2507.20973*, 2025. URL https://arxiv.org/abs/2507.20973.

Zihao Xue, Yan Wang, Zhen Bi, Long Ma, Zhonglong Zheng, Zeyu Yang, Bingyu Zhu, Longtao Huang, Jie Xiao, and Jungang Lou. Robust and generalizable safety steering for text-to-image diffusion transformers. *arXiv preprint arXiv:2605.30049*, 2026. URL https://arxiv.org/abs/2605.30049.

Calvin Yeung, Prathyush Poduval, Ali Zakeri, Zhuowen Zou, and Mohsen Imani. Residualized temporal sparse autoencoders for interpreting diffusion models. *arXiv preprint arXiv:2605.27813*, 2026. URL https://arxiv.org/abs/2605.27813.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Yuchang Zhu, Huizhe Zhang, Bingzhe Wu, Jintang Li, Zibin Zheng, Peilin Zhao, Liang Chen, and Yatao Bian. Measuring diversity in synthetic datasets. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80373–80397. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhu25ac.html.

A SDXL-TURBO SAE CONFIGURATION AND EXPERIMENTAL SETUP

Section 3 and its accompanying appendices use the public pretrained SDXL-Turbo SAEs released by Surkov et al. (2025), with checkpoints provided by Wendler (2025). These SAEs correspond to four U-Net modules: `down.2.1`, `mid.0`, `up.0.0`, and `up.0.1`. They encode each module's residual update, namely the difference between its output and input. Each SAE has an input width of 1,280, 5,120 dictionary directions, and TopK set to 10. All images are generated at 512×512 resolution with four denoising steps, `float32` precision, and `guidance_scale` set to 0. Denoising steps 0, 1, 2, and 3 correspond to scheduler timesteps 999, 749, 499, and 249, respectively. SAE activation maps are 16×16 at every module.

Prompts are sampled from LAION-COCO (Schuhmann et al., 2022). We read the text field, discard missing values, strip leading and trailing whitespace, and remove empty texts and exact duplicates. We then shuffle the remaining prompts with the fixed random seed 41872963 and take the first 10,000 as the experimental prompt set. Appendix B describes the content coverage of this set and its unedited generated images.

Table A.1 summarizes the experimental settings. Natural-activation statistics and downstream-response readouts encode residual updates without modifying model activations. Feature interventions add an update to the module output along the target SAE decoder direction, with an amplitude equal to the edit strength multiplied by the feature's reference mean activation. The full update is applied at every spatial position in each of the four denoising steps; the strength is not divided across steps. Response restoration uses an unedited generation with the same prompt and random seed as its reference and adds only the difference in the selected coordinates' decoder contributions to the module output. Neither intervention replaces the module output with a full SAE reconstruction. The selection rules, metric for cross-prompt consistency of changes in image representations, and response-restoration method are detailed in Appendices C, D, and E, respectively.

Table A.1: Experimental settings for Section 3 and its accompanying appendices. Primary strengths are $\{\pm 5, \pm 10, \pm 15, \pm 20\}$ and stress strengths are $\{\pm 40, \pm 60, \pm 80, \pm 100\}$; strength 0 denotes the unedited baseline. Response measurement and restoration use two paths: `down.2.1`→`mid.0` covers steps 0–3, and `up.0.1`→`down.2.1` covers steps 1–3. Edits and controls within each group share the same prompt and random seed.

Experiment	Prompts	Targets	Strengths	Intervention or readout
Scaffold-feature selection	10,000	All SAE features in four modules	None	Read natural activations at all four denoising steps to measure cross-prompt reuse and spatial coverage
Cross-prompt direction consistency	10,000	62 deduplicated scaffold features	Primary	Intervene at all denoising steps and spatial positions in the feature's module
Downstream response measurement	10,000	4 concept features and their downstream SAE features	Primary and stress	Intervene on concept features and read responses along same-step downstream or cross-step paths
Response-restoration intervention	10,000	Active scaffold responses and matched ordinary responses	Primary and stress	Retain concept intervention and separately restore selected coordinates in the two response groups

B Prompt and Generated Image Diversity

B.1 Data Correspondence and Representations

Since all subsequent experiments use a fixed set of 10,000 prompts, we need to analyze the diversity of their content coverage. Otherwise, in the extreme case where all 10,000 prompts are highly similar or even identical, a population of repeatedly activated features would be almost inevitable. Each prompt corresponds to one 512×512 image, generated using four denoising steps, guidance_scale$= 0$, and float32 precision. All prompts are unique, and none is truncated by the SDXL `text_encoder_2` tokenizer.

We represent prompts with the 1,280-dimensional pooled output from SDXL `text_encoder_2` and images with the 768-dimensional global features from DINOv2 ViT-B/14, and unit-normalize each type of vector separately. Let $X = \{x_i\}_{i=1}^{N}$ denote the embedding set for either modality, where $N = 10{,}000$.

B.2 Diversity Metrics

The four metrics in Table B.1 describe data diversity in terms of between-sample differences and the similarity spectrum. DCScore (Zhu et al., 2025) measures distinguishability through classification probabilities between samples and is computed as the sum of their self-assignment probabilities. The Vendi Score (Friedman & Dieng, 2023) measures effective diversity as the exponential of the spectral entropy of the normalized similarity matrix. For Rényi kernel entropy, we use its second-order exponential form (Jalali et al., 2023) to measure effective diversity at a given kernel scale. We also report the mean cosine distance between distinct samples to directly describe average differences in their representations. With $s_{ij} = x_i^\top x_j$ and temperature $\tau = 1$, DCScore is computed as

$$P_{ij} = \frac{\exp(s_{ij}/\tau)}{\sum_{k=1}^{N} \exp(s_{ik}/\tau)}, \qquad \mathrm{DCScore}(X) = \sum_{i=1}^{N} P_{ii}. \tag{1}$$

For Vendi, let $\lambda_1, \ldots, \lambda_r$ be the nonnegative eigenvalues of $XX^\top/N$; we compute

$$\mathrm{Vendi}(X) = \exp\left(-\sum_{j=1}^{r} \lambda_j \log \lambda_j \right). \tag{2}$$

Using a Gaussian kernel with bandwidth σ, the exponential form of second-order Rényi kernel entropy is

$$D_{\mathrm{RKE}}(X;\sigma) = \left[\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \exp\left(-\frac{\|x_i - x_j\|_2^2}{\sigma^2} \right) \right]^{-1}, \tag{3}$$

Here, D_{RKE} corresponds to the Rényi Kernel Entropy Mode Count (RKE-MC) in the original paper. The mean cosine distance is

$$\bar{D}(X) = \frac{2}{N(N-1)} \sum_{i<j} (1 - x_i^\top x_j). \tag{4}$$

Vendi's nonzero eigenvalues are computed from $X^\top X/N$, avoiding explicit construction of a $10{,}000 \times 10{,}000$ matrix. The prompt matrix produces 15 tiny negative eigenvalues, the smallest of which is -3.22×10^{-9}. These values arise from numerical roundoff and are set to zero before calculating spectral entropy.

RKE uses a squared Gaussian kernel and includes all ordered sample pairs and self-pairs. Mean cosine distance is calculated only over unordered pairs of distinct samples. We also recompute mean cosine distance with an equivalent vector-sum formula and obtain the same result.

At $N = 10{,}000$ and $\tau = 1$, DCScore describes the strength of self-matching relative to matching other samples, rather than the number of distinct samples. Vendi describes how dispersed the similarity spectrum is, whereas RKE describes effective diversity at a given kernel scale.

Table B.1: Diversity metrics for 10,000 prompts and their generated images. Prompts and images use the SDXL text encoder and DINOv2 representations, respectively; all metrics are computed in their own representation spaces.

Data	DCScore	Vendi	$D_{\mathrm{RKE}}(0.3)$	Mean cosine distance
Prompts	2.123	133.85	9899.2	0.755
Generated images	2.635	440.07	9530.7	0.971

Table B.2: RKE at different kernel bandwidths. The main experiment uses $\sigma = 0.3$; results at $\sigma = 0.2$ and 0.4 are also reported.

Object	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$
10,000 prompts	9992.3	9899.2	4053.6
10,000 images	9973.5	9530.7	7254.7

As shown in Table B.1, DCScore exceeds 1, the value for a set of identical samples, in both representations; mean cosine distance also indicates representational differences among samples. The Vendi results show that the similarity spectrum is distributed across multiple effective directions, rather than concentrated in very few shared directions. Differences in prompt wording thus coincide with changes in both text representations and generated images.

RKE complements these results at different similarity scales. Table B.2 shows that, at a smaller bandwidth, the effective number of modes approaches the total sample count, indicating that most samples have low mutual kernel similarity at this local scale. As the bandwidth increases, more samples are treated as similar, and the effective number of modes decreases. The data therefore retain differences among samples at local scales while also containing similar structures at larger scales. We use $\sigma = 0.3$ as the reference setting and retain the results at $\sigma = 0.2$ and 0.4 to show this scale dependence.

B.3 Two-Dimensional Distributions and Image Samples

To inspect the embedding distributions visually, we reduce prompt and image vectors to two dimensions separately. Each set is first reduced to 50 dimensions with non-whitened PCA and then to two dimensions with PaCMAP. PaCMAP uses 10 neighbors, `MN_ratio`$= 0.5$, and `FP_ratio`$= 2.0$.

We run HDBSCAN in each modality's 50-dimensional PCA space, with minimum cluster size 100 and minimum samples 20. Under these settings, prompts do not form qualifying clusters; images form three clusters, with another 6,829 images left unclustered.

As shown in Figure B.1(a), the prompt projections form a continuous distribution with varying density, rather than just a few compact groups. In Figure B.1(b), the image representations contain both a broad main region and more distinct local groups. The numbered markers link four prompt–image pairs to their respective projection positions, showing a surfing scene, jewelry, a botanical illustration, and an industrial scene. These examples connect the overall distributions to specific generated content.

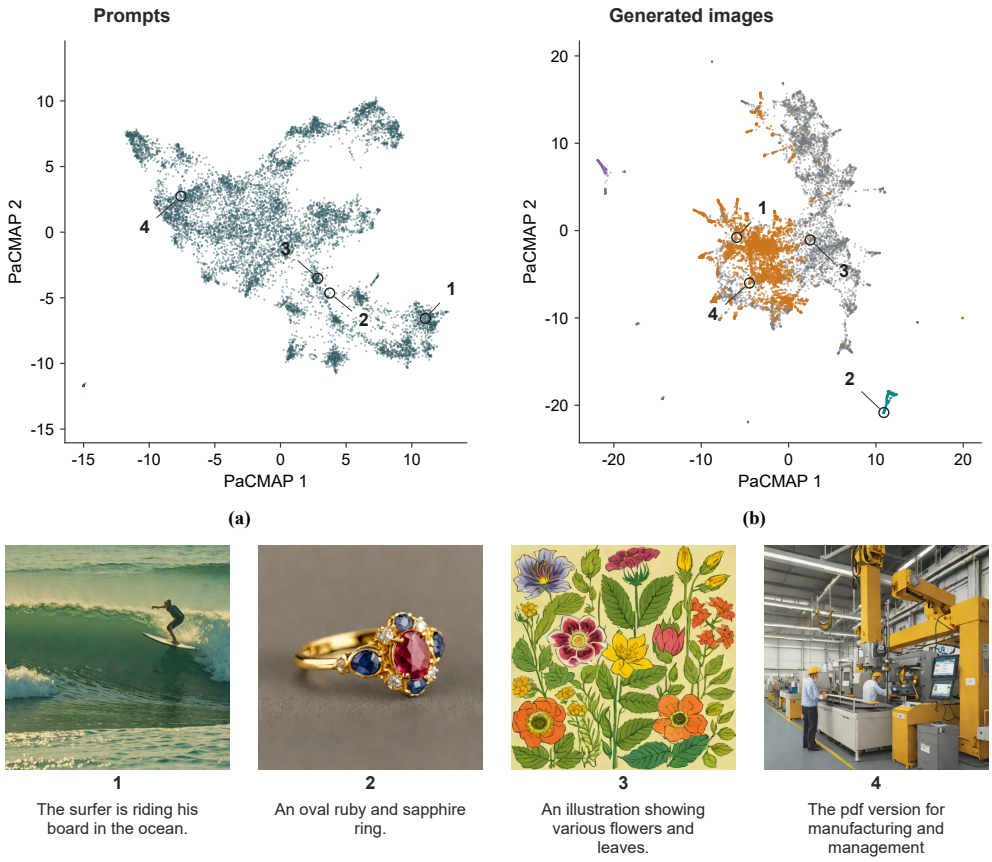


Figure B.1: Two-dimensional distributions and linked examples from the scaffold-selection data. (a) Projections of text representations for 10,000 prompts; (b) projections of visual representations for the corresponding generated images. Each set of representations is reduced with PCA and PaCMAP. In the image panel, orange, cyan, and purple denote the three clusters identified by HDBSCAN; gray denotes unclustered samples. Matching numbers in the two panels refer to the same prompt–image pair below, with circles and leader lines marking its projected positions.

Figure B.2 shows more original generated images. To cover different content, we group the existing DINOv2 representations into 48 groups with MiniBatch K-Means and take the four images nearest to each center. From these 192 candidates, we select 36 images based on differences in objects, scenes, and image types, not on editing effects. The sample wall includes human activities, products, buildings, indoor and outdoor scenes, as well as illustrations, maps, technical graphics, and interfaces. Together with the representation statistics in Table B.1 and the overall distributions in Figure B.1, these examples show that scaffold selection is not limited to repeated prompts or a single image type, but measures which features recur across diverse generated content.

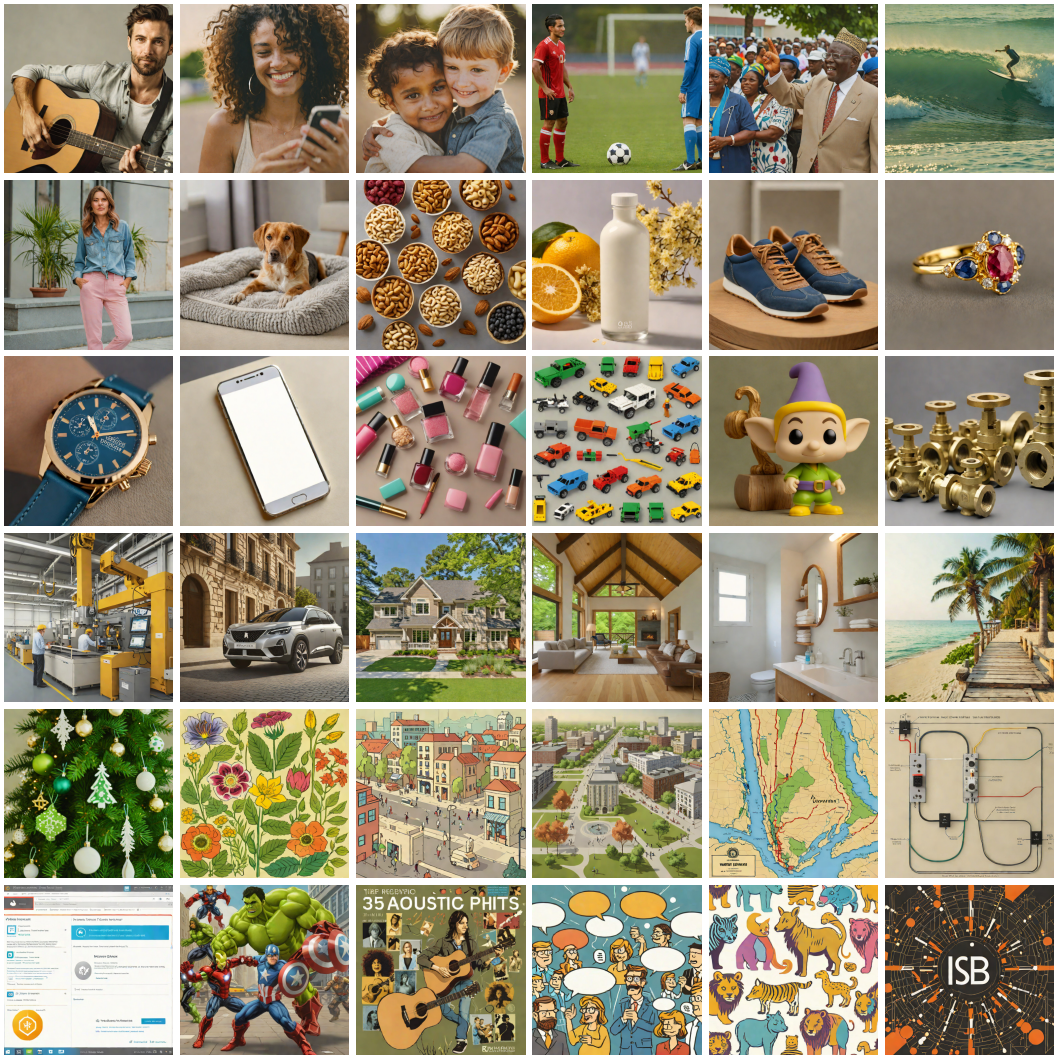


Figure B.2: Generated content covered by the scaffold-selection data. The 36 images are selected from unedited SDXL-Turbo outputs for the same 10,000 prompts and their random seeds used in Section 3.1, covering different objects, scenes, and image types. All images retain their original content.

C SELECTION CRITERIA AND SIMPLE-SOURCE CHECKS

This appendix supplements the screening experiments in Section 3. We examine the effects of spatial-coverage requirements and prompt sample size, and analyze scaffold features' responses to fixed inputs and variation in their spatial locations.

C.1 EFFECTS OF SPATIAL COVERAGE AND SAMPLE SIZE

The main text measures cross-prompt prevalence by checking whether a feature activates at one or more positions in an image. To examine the effect of spatial-coverage requirements, we generalize this criterion as

$$b_{i,\ell,t,k}(q) = \mathbf{1}\left[\sum_{s=1}^{256} \mathbf{1}(a_{i,\ell,t,k,s} \neq 0) \geq m(q)\right], \tag{5}$$

where $m(0^+) = 1$ and $m(q) = \lceil 256q \rceil$ for $q > 0$. We retain a module–timestep–feature instance (ℓ, t, k) if it satisfies this condition in at least 90% of prompts. All counts below use these instances as the unit: the same feature is counted separately at different timesteps.

Figure C.1(a) shows the number and module composition of selected instances under different spatial-coverage requirements on 10,000 prompts. Raising the requirement from at least one active position to 5% reduces the selected population from 113 to 11 instances, all from `up.0.1`. Thus, stricter spatial coverage changes not only the population size but also its module composition. Cross-prompt reuse and widespread activation within a single image are distinct properties: a feature can recur across different content while activating only in a local region of each image. The main text uses the at-least-one-position rule to retain features with sparse spatial support but high cross-prompt prevalence.

With $q = 0^+$ and the 90% cross-prompt threshold fixed, Figure C.1(b) compares the selected sets across sample sizes. We use the selected set obtained from 1,000,000 prompts as the reference and compute Jaccard similarity as the size of the set intersection divided by the size of the union. This million-prompt sample is a subset of LAION-COCO, not the full dataset. Increasing the sample size from 1,000 to 10,000 changes the selected count from 118 to 113 and raises the similarity from 0.8934 to 1. The selected set remains unchanged when the sample size increases further to 100,000 and 1,000,000 prompts. The 10,000 prompts used in the main text therefore yield the same screening result as these larger samples.

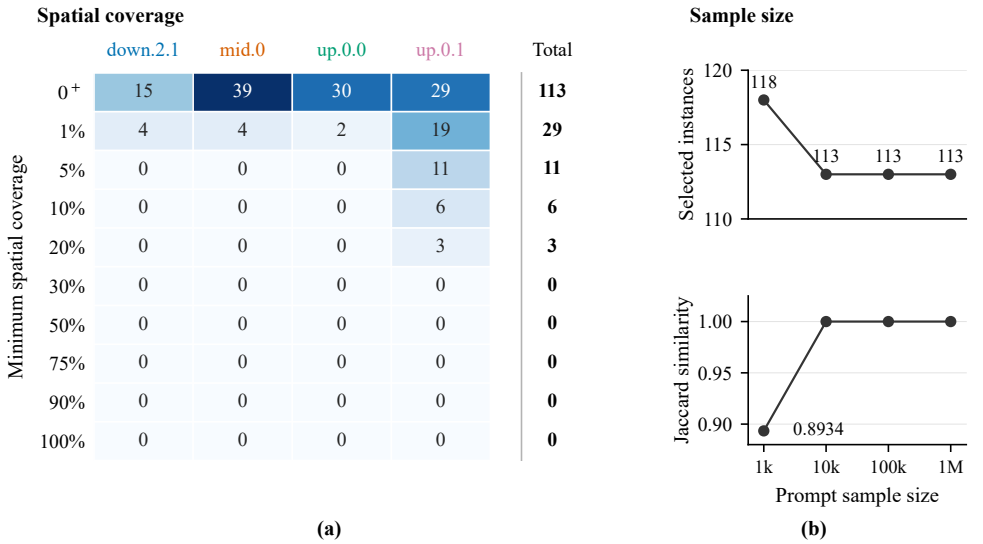


Figure C.1: Effects of spatial coverage and sample size on scaffold selection. (a) Selected instance counts by module as the minimum within-image spatial coverage varies on 10,000 prompts; the rightmost column gives the totals. The cross-prompt prevalence threshold is fixed at 90%, and the same feature is counted separately at different timesteps. $q = 0^+$ denotes activation at one or more positions. **(b)** With $q = 0^+$ fixed, separate panels show the selected counts at four prompt sample sizes and the Jaccard similarity of each selected set to the set selected from 1,000,000 prompts.

C.2 Fixed-Input Responses and Spatial-Location Variation

To distinguish cross-prompt reuse from responses to simple fixed inputs, we construct zero, global-mean, and position-mean inputs. Let $x_{i,\ell,t,s}$ denote the activation vector received by the SAE during natural generation. Using $N = 10{,}000$ prompts, we compute

$$\mu_{\ell,t,s} = \frac{1}{N}\sum_{i=1}^{N} x_{i,\ell,t,s}, \qquad \bar{\mu}_{\ell,t} = \frac{1}{256}\sum_{s=1}^{256} \mu_{\ell,t,s}. \tag{6}$$

Zero input uses an all-zero vector at every position; global-mean input repeats $\bar{\mu}_{\ell,t}$ at all positions; position-mean input uses $\mu_{\ell,t,s}$ at position s. All three inputs are encoded by the corresponding SAE. A feature is counted as responding to an input if it has a nonzero activation at any position.

Figure 2(c) in the main text shows the fixed-input responses of individual scaffold instances, while Figure C.2(a) breaks down these response combinations by module. The modules differ markedly: most instances in `down.2.1` activate under position-mean input, whereas most instances in `up.0.1` activate under none of the three fixed inputs. The scaffold population therefore does not share a uniform fixed-input response pattern, and its cross-prompt reuse cannot be uniformly attributed to responses to zero or mean inputs.

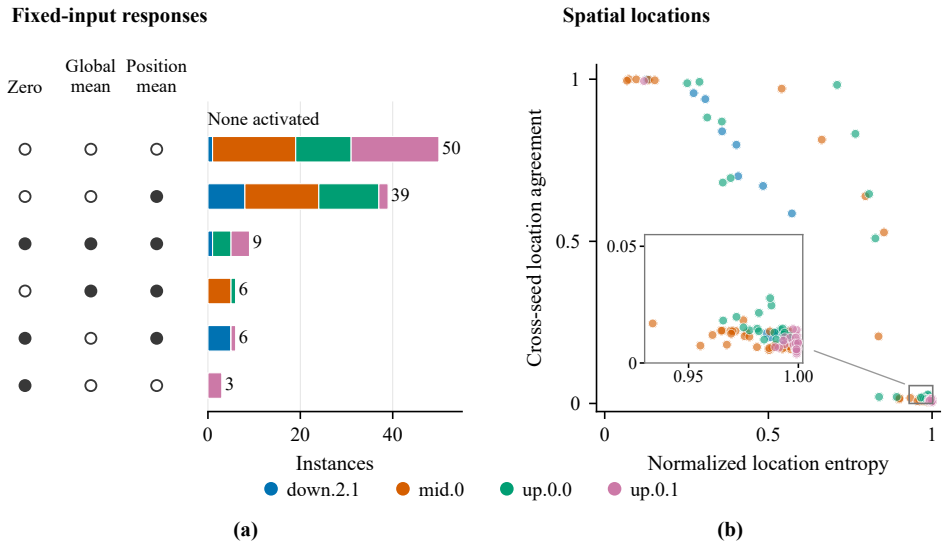


Figure C.2: Fixed-input responses and spatial-location variation of scaffold features. (a) Three filled or open markers indicate whether a feature responds to zero, global-mean, and position-mean input, respectively. Horizontal stacked bars show instance counts for each response combination by module and include all 113 scaffold instances. **(b)** Each point represents one scaffold instance, colored by module. The horizontal axis is normalized location entropy computed from 10,000 prompts; the vertical axis is agreement in the maximum-activation location under the same prompt with different random seeds. The cross-seed experiment uses the same 10,000 prompts as the main text, with four seeds per prompt, and includes only seed pairs in which the feature is active in both generations. The inset enlarges the high-entropy, low-agreement region.

We also examine whether frequent occurrence is accompanied by fixed activation locations. For each scaffold instance f, we aggregate its nonzero activation counts during natural generation by spatial position and normalize them to obtain a location distribution $p_{f,s}$. We then compute normalized location entropy as

$$H_{\mathrm{loc}}(f) = -\frac{1}{\log 256}\sum_{s=1}^{256} p_{f,s}\log p_{f,s}. \tag{7}$$

Higher location entropy indicates a more dispersed activation-location distribution aggregated across prompts. This measure differs from within-image spatial coverage, since a locally active feature can still have high location entropy when its activation location varies with content.

To examine location variation under the same prompt, we reuse the 10,000 prompts from the main text, generate images with four random seeds per prompt, and compare the six seed pairs for each

prompt. Conditional on the feature being active in both generations, we record the proportion of pairs with exactly the same maximum-activation location as the cross-seed location agreement.

As shown in Figure C.2(b), instances in `down.2.1` generally have more concentrated location distributions and higher cross-seed location agreement. Most instances in the middle and upsampling modules instead have higher location entropy and lower location agreement. These results complement the natural activation examples in Figure 3(b) in the main text and show that features can participate repeatedly in generation across prompts without being tied to the same image region.

D Effects of Scaffold Features Across Content

D.1 Cross-Prompt Consistency of Changes in Image Representations

This section supplements the computation, per-feature results, and comparison between image encoders for Figure 3(a) in the main text. To investigate whether scaffold features can be characterized by a single concept, we use causal interventions to examine whether the image changes caused by the same scaffold feature across prompts share a common direction. Here, direction refers to changes in the CLIP or DINOv2 image-representation space, not to the SAE decoder direction. The experiment includes all 62 distinct module–feature pairs and uses a fixed set of 10,000 prompts; occurrences of the same feature at different denoising steps are not counted separately.

Each intervention edits only one feature by adding its SAE decoder direction at all 16×16 spatial positions in all four denoising steps, scaled by that feature's mean activation. In addition to the unedited image, we use strengths $\mathcal{S} = \{-20, -15, -10, -5, 5, 10, 15, 20\}$. We encode images with CLIP ViT-L/14 and DINOv2 ViT-B/14 and measure edit directions independently in their respective representation spaces.

For prompt p, feature f, edit strength s, and image encoder M, let $e^M_{p,f,s}$ denote the unit-normalized image embedding. We estimate the signed edit direction for that prompt from the positive and negative strength sequence:

$$b^M_{p,f} = \frac{\sum_{s\in\mathcal{S}} s\left(e^M_{p,f,s} - e^M_{p,f,0}\right)}{\sum_{s\in\mathcal{S}} s^2}, \qquad \widehat{d}^M_{p,f} = \frac{b^M_{p,f}}{\|b^M_{p,f}\|_2}. \tag{8}$$

Here, $e^M_{p,f,0}$ is the embedding of the unedited image, $b^M_{p,f}$ is the direction of change fitted from the strength sequence, and $\widehat{d}^M_{p,f}$ is its unit vector. This estimate uses the complete strength sweep rather than comparing edits at only one strength.

The cross-prompt direction consistency of feature f is defined as

$$C^M_f = \frac{1}{N(N-1)} \sum_{p\neq p'} \left\langle \widehat{d}^M_{p,f}, \widehat{d}^M_{p',f} \right\rangle, \qquad N = 10{,}000. \tag{9}$$

This metric is the mean cosine similarity between the edit directions produced by the same feature under different prompts. Higher values indicate a stronger direction of change shared across prompts, while values close to zero indicate that the directions lack a common orientation. Because each per-prompt direction is unit-normalized, the metric measures directional consistency.

Figure D.1(a) shows the cumulative distribution of consistency averaged across the two encoders. Most scaffold features lie in the low-consistency region, while a small subset has a stronger shared direction of change. The per-feature results in Table D.1 also reveal differences between modules. Higher-consistency features occur mainly in `up.0.1`, whereas most features in `mid.0` and `up.0.0` have lower consistency; `down.2.1` contains both higher- and lower-consistency features.

Figure D.1(b) compares the two encoders' measurements of the same features. CLIP and DINOv2 have a Pearson correlation of 0.971 and a Spearman correlation of 0.964, indicating similar assessments of which features have higher or lower consistency. The distribution, with a low-consistency majority and a small high-consistency subset, is therefore not specific to one encoder. The next section uses individual interventions to show the image changes corresponding to these differences in representation-space directions.

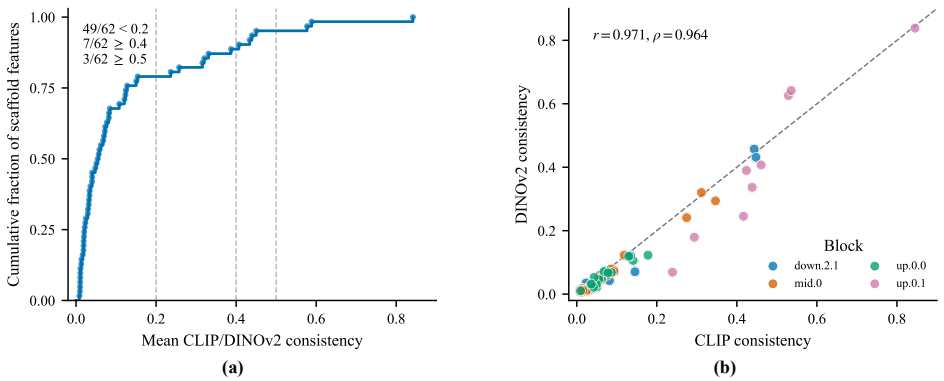


Figure D.1: Changes in image representations induced by the full scaffold population and their cross-prompt consistency. (a) The empirical cumulative distribution for 62 scaffold features, with the arithmetic mean of CLIP and DINOv2 consistency on the horizontal axis. **(b)** Per-feature consistency measured by the two encoders. Each point represents a module–feature pair, colored by module; the dashed line indicates equal consistency values for the two encoders. Both panels use the same fixed set of 10,000 prompts and strength sweep.

Table D.1: Cross-prompt consistency of changes in image representations induced by individual scaffold features. Cross-prompt direction consistency in CLIP and DINOv2 representations is reported for all 62 module–feature pairs, sorted in descending order of the mean consistency across the two encoders, reading each row from left to right.

Module	Feature	CLIP C	DINOv2 C	Module	Feature	CLIP C	DINOv2 C
up.0.1	879	0.845	0.838	up.0.1	2727	0.536	0.642
up.0.1	500	0.529	0.626	down.2.1	4781	0.443	0.457
down.2.1	4843	0.448	0.431	up.0.1	1910	0.461	0.407
up.0.1	4163	0.424	0.390	up.0.1	362	0.438	0.337
up.0.1	3936	0.417	0.246	mid.0	473	0.347	0.294
mid.0	2825	0.311	0.320	mid.0	876	0.275	0.241
up.0.1	2164	0.294	0.179	up.0.1	3078	0.239	0.069
up.0.0	4512	0.178	0.123	up.0.0	2221	0.135	0.121
up.0.0	2229	0.130	0.119	up.0.0	1172	0.141	0.106
mid.0	961	0.118	0.123	down.2.1	1456	0.145	0.070
mid.0	2691	0.092	0.077	mid.0	3042	0.085	0.079
mid.0	3983	0.093	0.071	up.0.0	1667	0.086	0.068
up.0.0	2533	0.078	0.067	up.0.0	573	0.073	0.071
up.0.0	801	0.068	0.071	mid.0	532	0.076	0.060
down.2.1	989	0.082	0.042	mid.0	3952	0.057	0.060
up.0.0	244	0.062	0.051	mid.0	3018	0.063	0.046
up.0.0	1623	0.055	0.047	up.0.0	4939	0.043	0.053
mid.0	836	0.053	0.028	up.0.0	1985	0.049	0.031
up.0.0	1604	0.046	0.034	up.0.0	1728	0.050	0.022
up.0.0	5036	0.036	0.032	up.0.0	2565	0.038	0.030
mid.0	843	0.038	0.027	up.0.0	4722	0.042	0.022
mid.0	2282	0.036	0.024	down.2.1	320	0.023	0.035
down.2.1	1714	0.034	0.016	down.2.1	1857	0.027	0.019
mid.0	419	0.027	0.016	mid.0	2928	0.029	0.012
mid.0	3955	0.025	0.014	mid.0	3406	0.024	0.015
mid.0	4143	0.025	0.013	mid.0	3150	0.025	0.012
mid.0	982	0.015	0.018	mid.0	4906	0.015	0.012
mid.0	3251	0.012	0.014	mid.0	904	0.013	0.009
mid.0	3397	0.012	0.009	mid.0	2139	0.008	0.013
mid.0	296	0.010	0.010	up.0.0	3234	0.010	0.011
mid.0	2308	0.010	0.009	mid.0	2756	0.009	0.008

D.2 Cross-Content Intervention Examples and Natural Activation Visualizations

We select `up.0.0/f1985`, `mid.0/f2756`, and `up.0.0/f4512` to provide additional intervention examples across different generated content. The examples use six prompts covering pottery, a bicycle, a portrait, architecture, a plant, and an illustration, with the random seed fixed at 42. Each trajectory keeps the prompt and seed fixed and shows nine strengths from -20 to $+20$ in increments of 5. Generation and intervention settings follow Section D.1.

Figure D.2 shows the cross-content effects of `up.0.0/f1985`. For the teapot, intervention changes the body proportions, handle, and lid decoration, with the handle disappearing under stronger positive intervention. Changes to the cottage involve roofing materials, doors, windows, and the entrance structure. The bicycle's rack, fenders, and frame details, as well as the plant's petal arrangement and flower-head shape, also vary with strength. Positive intervention has a shared tendency to simplify structure across these examples, but the particular parts changed or removed depend on the generated object.

The effects of `mid.0/f2756` primarily involve viewpoint, orientation, and spatial composition, as shown in Figure D.3. The cottage shifts from a nearly frontal presentation to a view with a more visible side, changing the relative arrangement of doors, windows, and the path. The tram shifts from a flat side view to one showing its front, while the street and building composition also changes. The bicycle's viewing angle and the plant's flower-head orientation and local shape likewise vary with intervention. These examples show that lower representation-space direction consistency can coexist with a recognizable visual tendency whose specific manifestation differs across content.

Figure D.4 shows the complete trajectories of `up.0.0/f4512`. Positive intervention gradually produces stronger contours and flatter color regions, creating an illustrative tendency across different content. The changes extend beyond surface style. The teapot's spout, handle, and decoration change; facial features, hairstyle, and clothing appearance change together in the portrait; and components and background structures are also altered in the bicycle and tram scenes. Stronger positive intervention additionally pushes some images toward line drawings.

The natural-activation visualization in Figure 3(b) follows *One-Step Is Enough: Sparse Autoencoders for Text-to-Image Diffusion Models* (Surkov et al., 2025). During unedited generation, we extract the 16×16 feature activations at a designated reference timestep, bilinearly upsample the activation values to image resolution, and retain active regions with a nearest-neighbor-upsampled nonzero mask. Colors are normalized to each image's nonzero activation range and overlaid on the corresponding generated image to show relative activation strength within that image.

These examples show that shared tendencies and content-dependent effects can coexist. Interventions along the same feature direction produce different structural, appearance, and spatial changes across generated content. Scaffold features can have a recognizable dominant tendency, but a single concept label cannot readily capture their full range of cross-content effects.

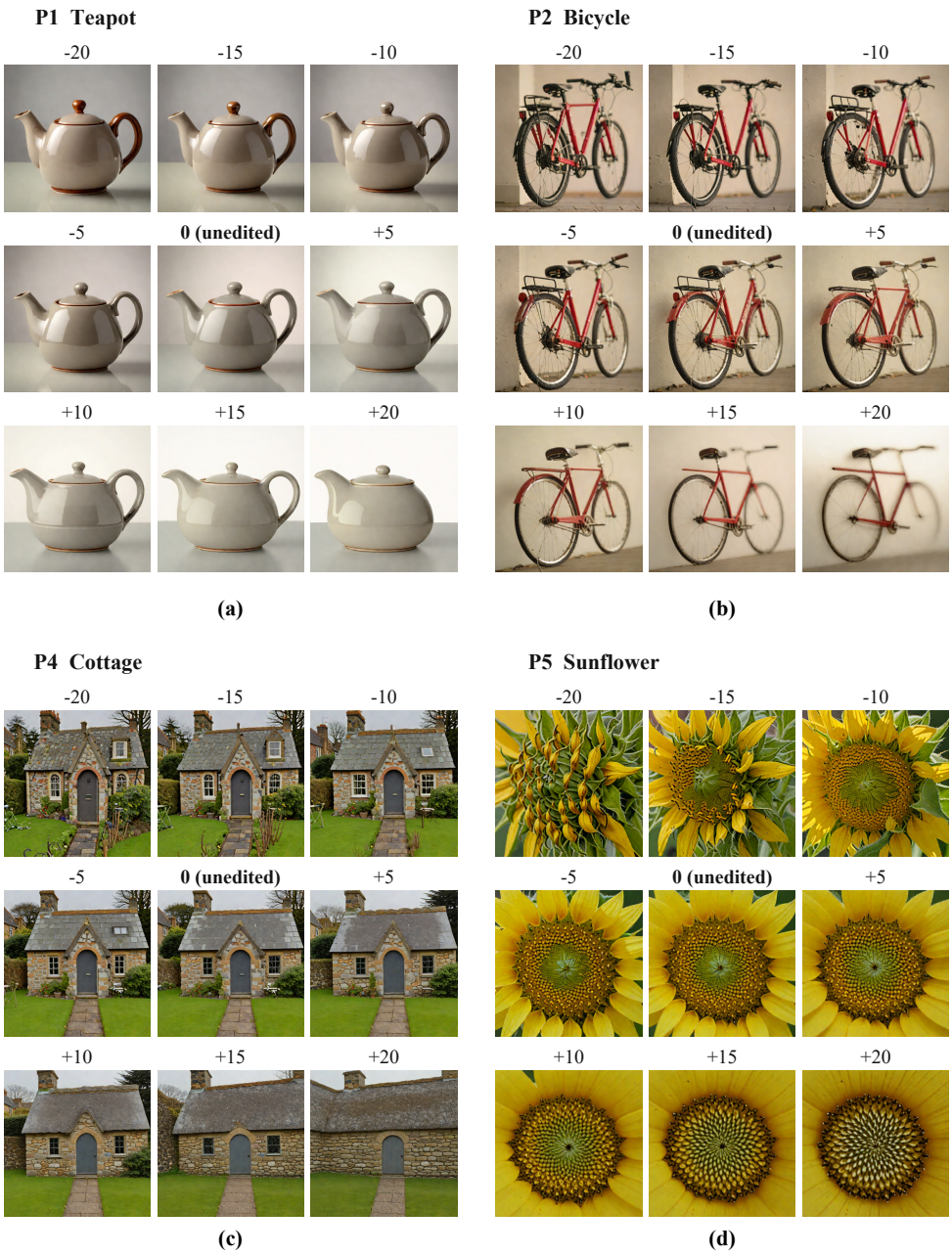


Figure D.2: Cross-content edit trajectories of scaffold feature `up.0.0/f1985`. The panels show interventions on a teapot, bicycle, cottage, and plant. Each group fixes the prompt and seed and shows strengths from -20 to $+20$ in increments of 5, ordered left to right and then top to bottom; zero strength is the unedited image. Nonzero interventions act at all spatial positions in all four denoising steps. The particular components, materials, and local shapes changed by the feature depend on the generated content.

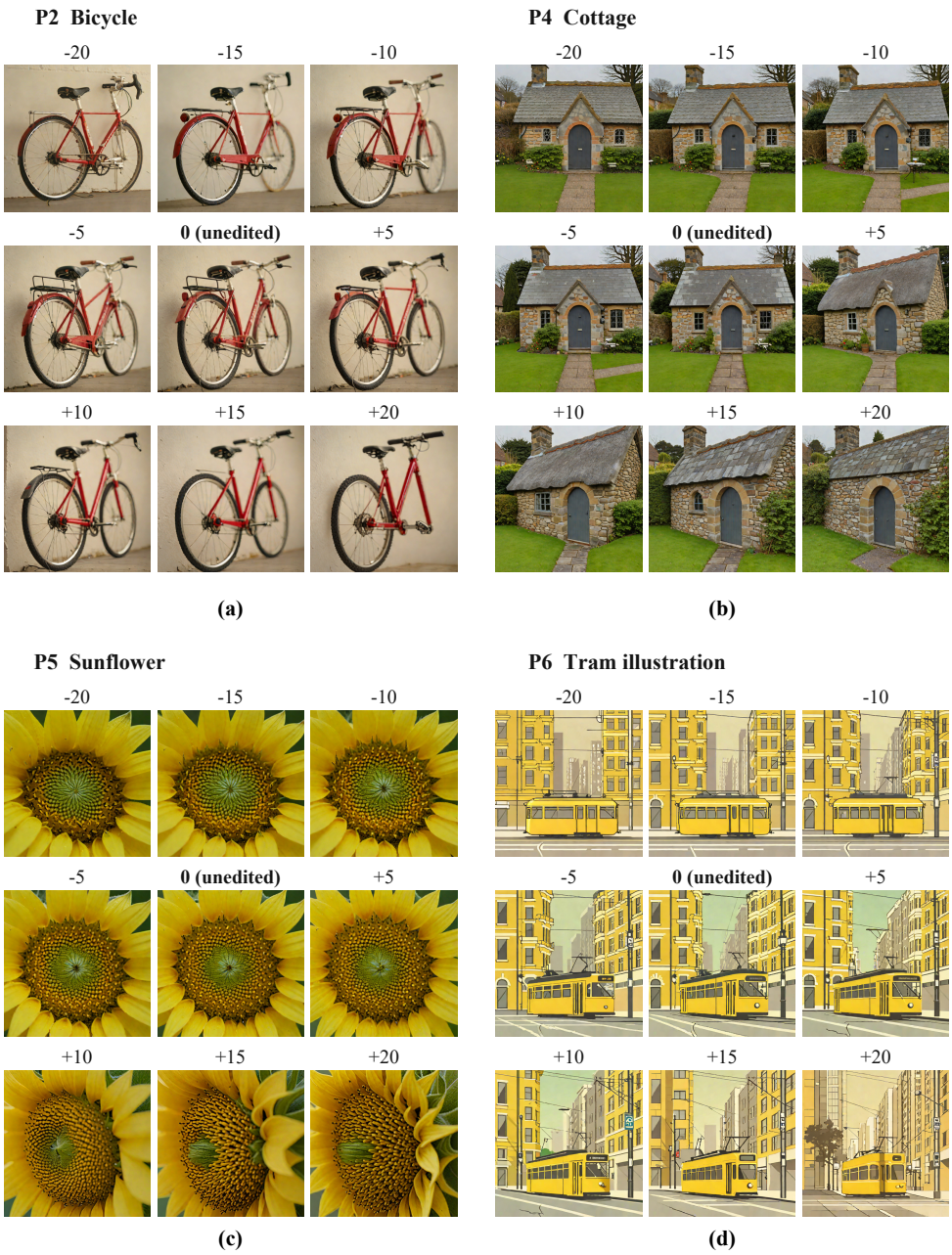


Figure D.3: Cross-content edit trajectories of scaffold feature `mid.0/f2756`. The panels show interventions on a bicycle, cottage, plant, and tram illustration. Each group fixes the prompt and seed and contains nine strengths from -20 to $+20$, ordered left to right and then top to bottom, with editing at all spatial positions in all four denoising steps. The feature changes viewing angle, subject orientation, and spatial composition across content.

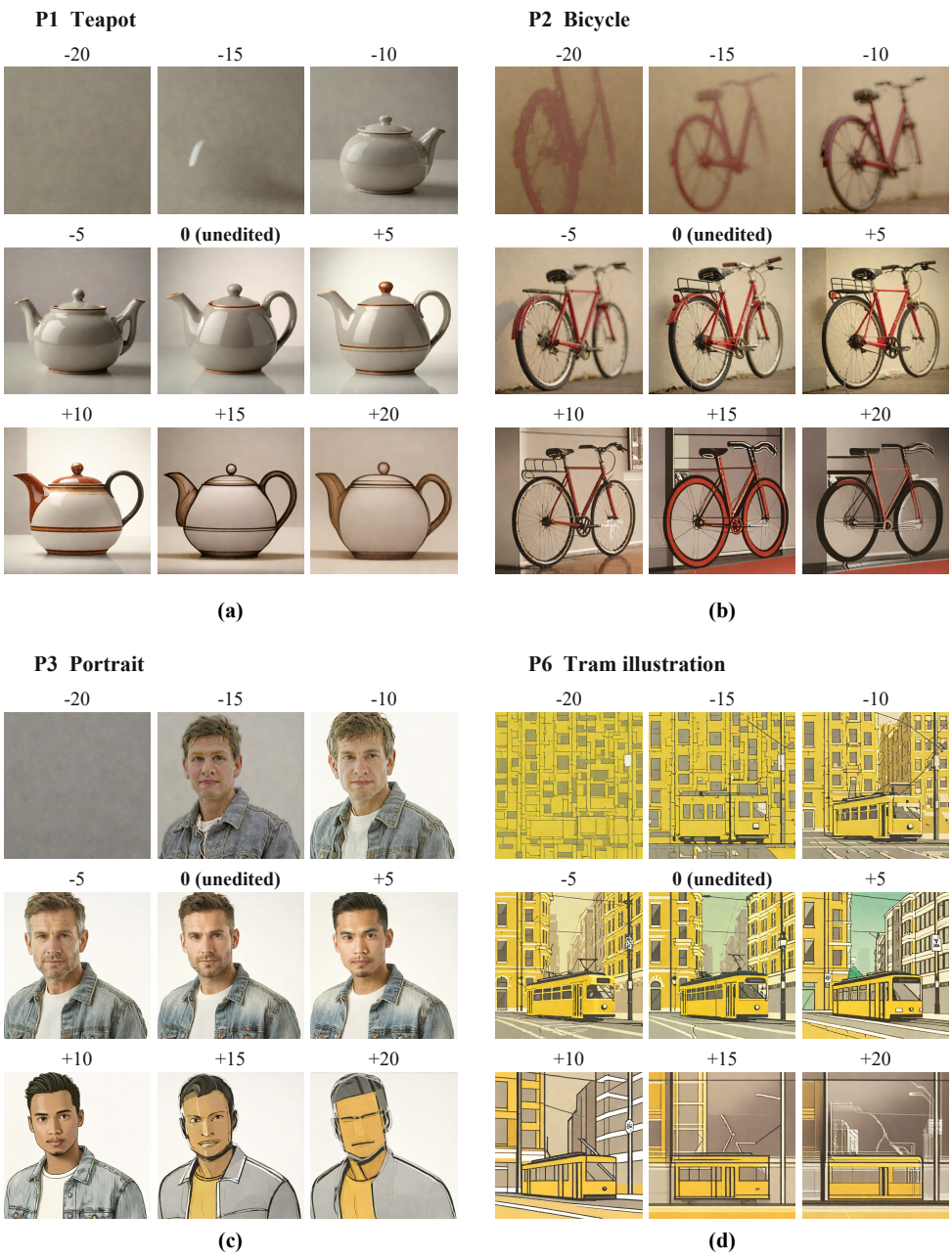


Figure D.4: Cross-content edit trajectories of scaffold feature `up.0.0/f4512`. The panels show interventions on a teapot, bicycle, portrait, and tram illustration. Each group fixes the prompt and seed and contains nine strengths from -20 to $+20$, ordered left to right and then top to bottom, with editing at all spatial positions in all four denoising steps. Positive intervention gradually produces stronger outlines and an illustrative tendency while also changing subject appearance, components, and scene structure across content.

E SCAFFOLD-RESPONSE RECRUITMENT AND CAUSAL EFFECTS

This appendix supplements the response-measurement and causal-intervention experiments in Section 3.3. We first measure downstream feature responses to concept editing, then restore selected coordinates to their unedited activations to block their editing responses. Comparing the effects of blocking scaffold and matched ordinary responses allows us to examine how these responses influence the final image.

E.1 RESPONSE MEASUREMENT

The experiments cover four concept features: `up.0.1/f3997`, `up.0.1/f1635`, `down.2.1/f4998`, and `down.2.1/f2301`. They are used to edit subject color, background color, cartoon-like rendering, and a threatening appearance, respectively. The `down.2.1`→`mid.0` path reads same-step downstream responses at steps 0–3, while the `up.0.1`→`down.2.1` path reads cross-step responses at steps 1–3. Prompt counts, editing strengths, and generation settings are given in Appendix A.

After editing concept feature g at strength a, the response of coordinate j is defined as

$$r^{\mathrm{raw}}_{p,c,t,j\leftarrow g}(a) = \left[\frac{1}{256}\sum_{x}\left(z^{a}_{p,c,t,x,j\leftarrow g} - z^{0}_{p,c,t,x,j}\right)^2\right]^{1/2}, \qquad r^{\mathrm{std}} = \frac{r^{\mathrm{raw}}}{\sigma_{c,t,j} + 10^{-8}}. \tag{10}$$

Here, p denotes the prompt, c the response module, t the denoising step, and x the spatial position; z^a and z^0 denote edited and unedited activations, respectively. The natural scale $\sigma_{c,t,j}$ is the root-mean-square activation of that coordinate across prompts and spatial positions in the unedited generations used for selection. Standardization allows comparison between features with different natural activation scales.

E.2 DISTRIBUTION OF SCAFFOLD RESPONSES

To locate scaffold responses within the ordinary-feature distribution, we first average standardized responses over prompts and a specified strength set A, then compute their percentiles:

$$\begin{aligned} \bar{r}^{A}_{g,c,t,j} &= \frac{1}{N|A|}\sum_{p}\sum_{a\in A} r^{\mathrm{std}}_{p,c,t,j\leftarrow g}(a), \\ q^{A}_{g,c,t,h} &= \frac{1}{|\mathcal{O}_{c,t}|}\sum_{o\in\mathcal{O}_{c,t}} \mathbf{1}\left[\bar{r}^{A}_{g,c,t,o} \le \bar{r}^{A}_{g,c,t,h}\right]. \end{aligned} \tag{11}$$

Here, N is the number of prompts, h is a scaffold feature, and $\mathcal{O}_{c,t}$ is the set of valid ordinary features at the same response location. A value of $q^A \ge 0.9$ places the scaffold response in the upper 10% of ordinary responses.

As shown in Figure E.1, scaffold responses are broadly distributed in the upper tail of ordinary responses, rather than being driven by a few extreme coordinates. A majority of scaffold responses enter this upper tail for each of the four concepts at primary strengths, indicating that the pattern is not specific to a single concept. Table E.1 breaks down the response locations: upper-tail concentration is more pronounced at later denoising steps, while early locations retain greater variation. Figure E.2 presents the complete results by concept and signed strength, allowing comparison of the same scaffold's response percentile across editing conditions.

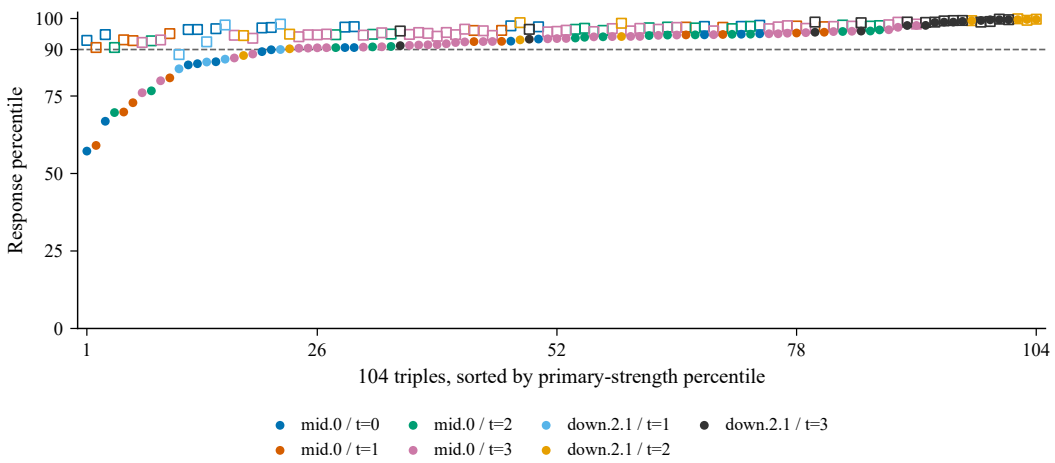


Figure E.1: Ranking scaffold responses among ordinary features. All 104 triples of concept feature, response location, and scaffold feature are shown, with responses averaged over 10,000 prompts and the corresponding strength set. Triples are sorted by their primary-strength percentile, with results over all nonzero strengths also indicated; the dashed line marks the upper-10% threshold of ordinary responses. Colors identify response locations; filled circles and open squares denote primary and all nonzero strengths, respectively.

Table E.1: Scaffold-response statistics by response location. Counts include all triples and those in the upper 10% of ordinary responses at primary and all nonzero strengths. Responses are averaged over 10,000 prompts and the corresponding strength set.

Response location	Triples	Primary top 10%	All nonzero top 10%
`down.2.1`/t=1	4	1	3
`down.2.1`/t=2	8	7	8
`down.2.1`/t=3	14	14	14
`mid.0`/t=0	14	7	14
`mid.0`/t=1	10	6	10
`mid.0`/t=2	16	14	16
`mid.0`/t=3	38	34	38

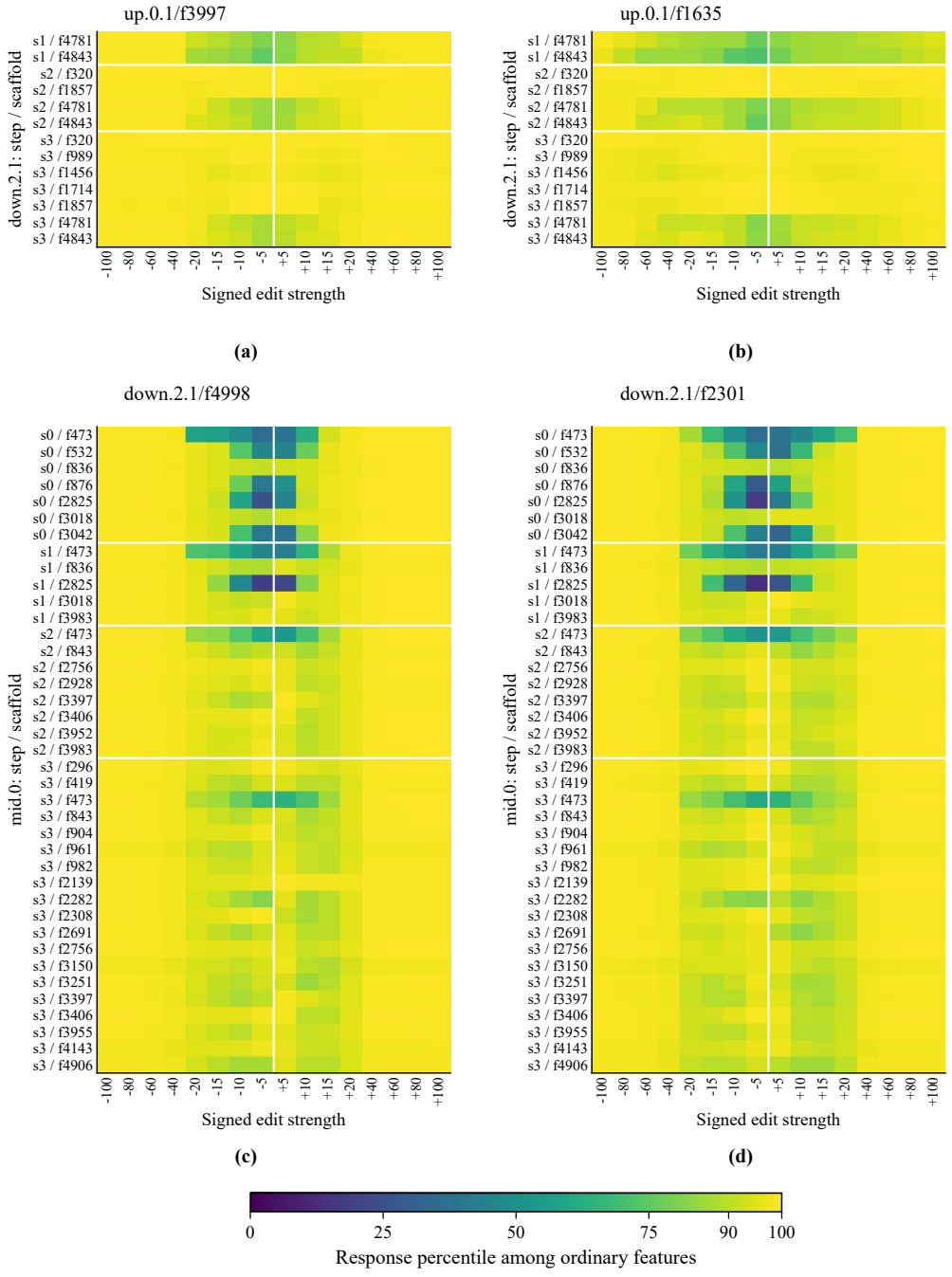


Figure E.2: Scaffold responses by concept and strength. Each panel corresponds to one concept, with nonzero signed strengths on the horizontal axis and response locations and scaffold features on the vertical axis. Color indicates the scaffold's response percentile among ordinary features after averaging over 10,000 prompts at the given strength; all panels share the same color scale. Each row represents one triple, with white lines separating denoising-step groups.

E.3 Matched Restoration and Path Comparison

We mark scaffold responses above the null-intervention noise threshold as active and match them to ordinary coordinates at the same response location. Matching uses decoder-space response strength $E_j = \|d_{c,j}\|_2 r_j^{\text{raw}}$, where $d_{c,j}$ is the SAE decoder direction. One-to-one matching minimizes the total difference, and the two groups restore the same number of coordinates. Figure E.3 shows the resulting matches.

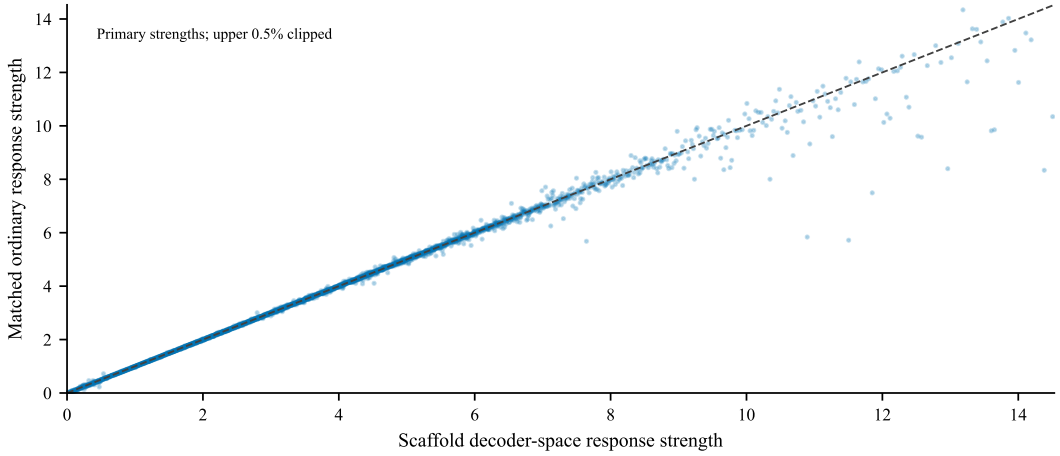


Figure E.3: Strength matching between scaffold and ordinary responses. Each point represents a matched coordinate pair at primary strengths. The horizontal and vertical axes show their decoder-space response strengths; the dashed line denotes equality.

During restoration, we encode the current intervened trajectory and use unedited activations from the same prompt and seed as the reference, adding only the difference in the selected coordinates' decoder contributions to the module output. The concept intervention remains unchanged, and the module output is not replaced by a full SAE reconstruction. Let C, H, and O denote the images from concept-only editing, scaffold-response blocking, and matched ordinary-response blocking, respectively. We define

$$\Delta_M = \mathbb{E}[d_M(C, H) - d_M(C, O)], \tag{12}$$

where d_M is the LPIPS, DINOv2, or CLIP-I image distance; CLIP-I distance in this appendix is one minus cosine similarity. Positive values indicate that blocking scaffold responses causes greater deviation from the concept-only image.

Figure E.4 and Table E.2 show positive mean differences for all three metrics on both the same-step and cross-step paths at primary strengths. Blocking scaffold responses causes greater deviation from concept-only images than blocking matched ordinary responses; this relationship is not limited to the same-step path. The concept-wise results in Figure E.5 retain this average trend for all four concepts, although their effect sizes differ.

These distance comparisons measure the impact of blocking on editing results. To distinguish whether these changes merely reflect weaker target editing, the next section examines their specific manifestations using target-semantic scores and image comparisons.

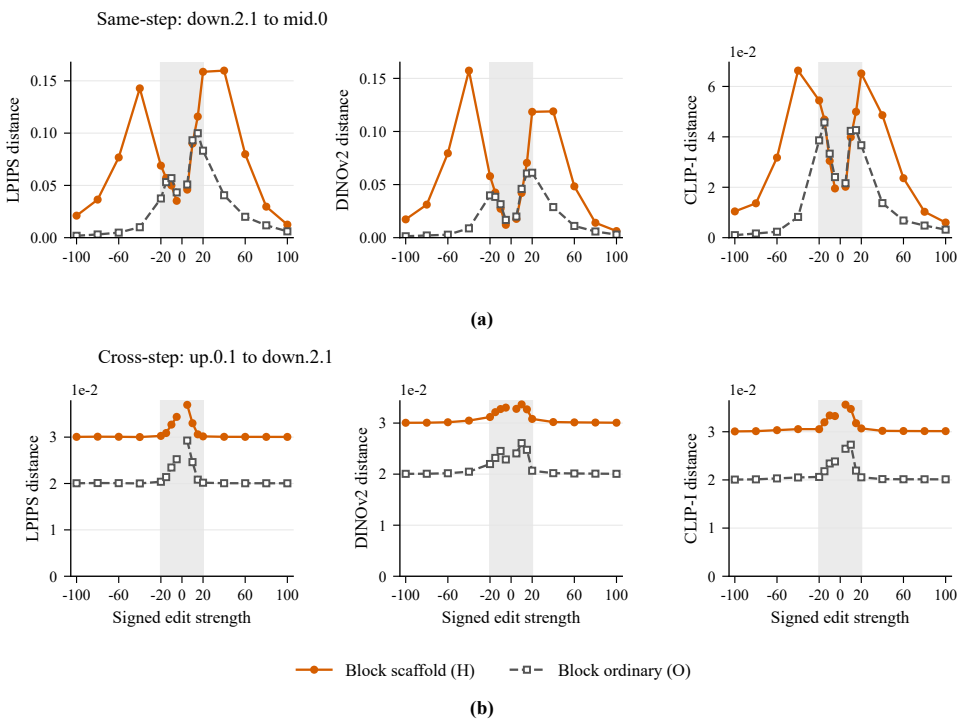


Figure E.4: Path-averaged effects of restoration interventions. LPIPS, DINOv2, and CLIP-I distances from the concept-only image are shown for both blocking conditions. Same-step and cross-step paths are shown separately; each point averages over 10,000 prompts and the two concepts on its path. Gray shading marks primary strengths.

Table E.2: Path-resolved results at primary strengths. Over 10,000 prompts, we report mean distances from the concept-only image under both blocking conditions, the paired difference Δ_M, and the ratio of the two mean distances. Primary strengths are $\pm 5, \pm 10, \pm 15, \pm 20$.

Metric	Block scaffold	Block ordinary	Δ_M	Ratio
`down.2.1`→`mid.0` (same-step)				
LPIPS	0.0777	0.0648	+0.0130	1.200
DINOv2	0.0486	0.0392	+0.0093	1.237
CLIP-I	0.0408	0.0356	+0.0052	1.145
`up.0.1`→`down.2.1` (cross-step)				
LPIPS	0.0324	0.0232	+0.0092	1.397
DINOv2	0.0324	0.0235	+0.0089	1.377
CLIP-I	0.0327	0.0232	+0.0095	1.409

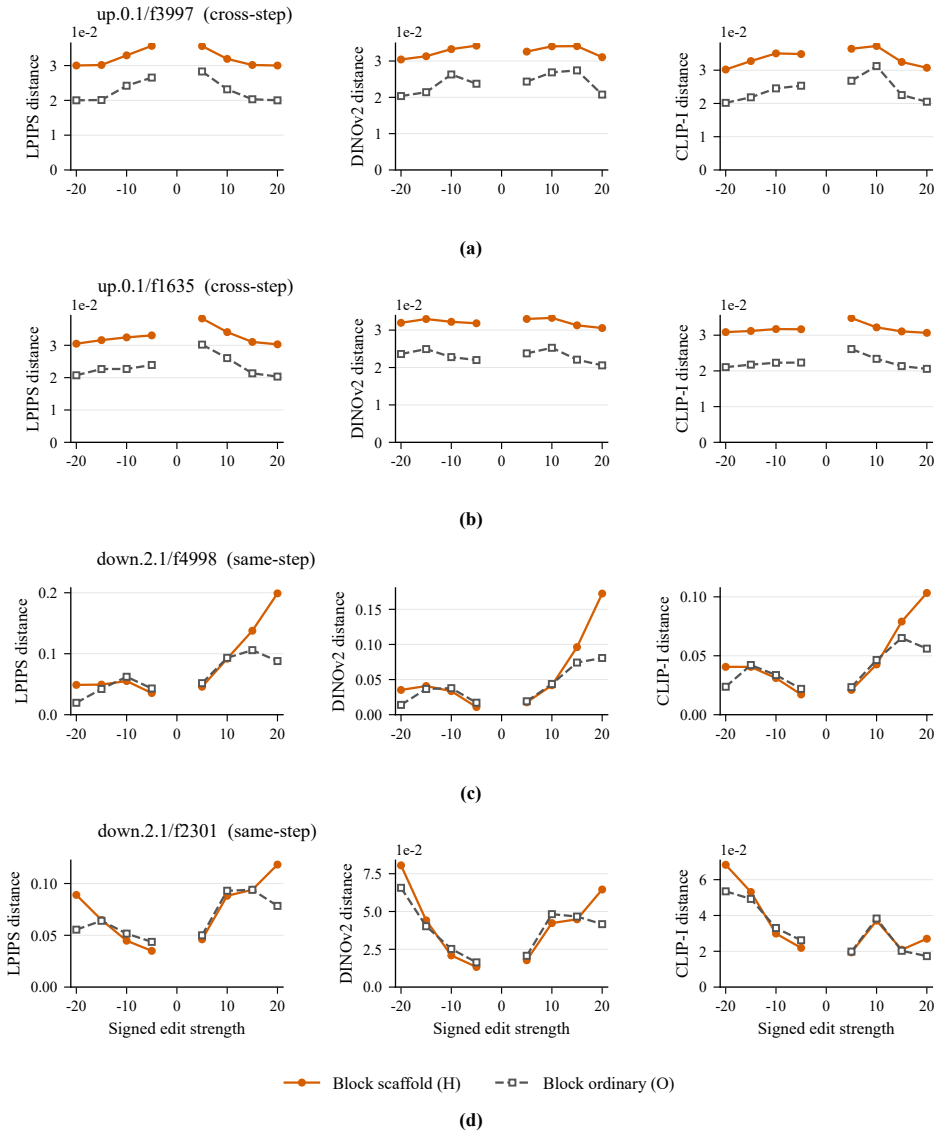


Figure E.5: Restoration-intervention effects for each concept. Both blocking distances are shown separately for the four concepts, using the same metrics as Figure E.4. This page shows primary strengths; the continuation shows stress strengths. Each point averages over 10,000 prompts for the same concept and strength.

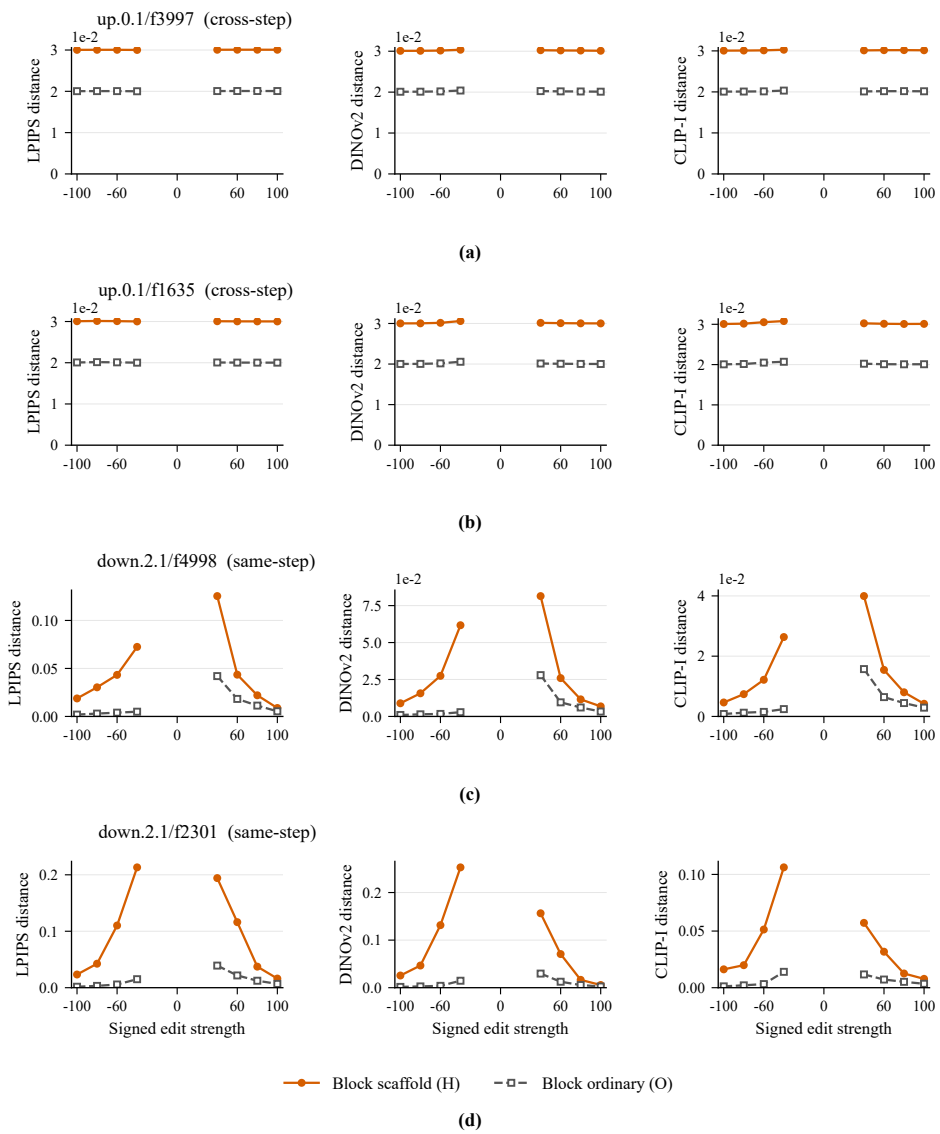


Figure E.5 (continued): Concept-wise restoration-intervention effects at stress strengths. Concept order, metrics, and blocking conditions match the preceding page.

E.4 Target-Semantic and Non-Target Content Changes After Blocking

To examine whether the image changes caused by blocking merely reflect weaker target editing, we combine target-semantic scores with image comparisons. The image comparisons show that changes in target attributes can be separated from the integrity of image content. For example, the ceramic pitcher in Figure E.8 remains blue after scaffold-response blocking, but its handle and attachment points develop abnormalities. Below, we report target-semantic changes and visible content damage.

Target-semantic evaluation uses `GPT-5.6-Sol`. The evaluator receives the unedited image, edited image, and specified editing direction, and scores the extent of target-direction change from 0 to 10 rather than overall image quality. The intervention condition is not provided to the evaluator. Let S_C, S_H, and S_O denote the scores for concept-only editing, scaffold-response blocking, and matched ordinary-response blocking, respectively. We define

$$D = (S_C - S_H) - (S_C - S_O) = S_O - S_H. \tag{13}$$

A positive D means that target-direction change is weaker after blocking scaffold responses than after blocking ordinary responses.

Figure E.6 shows semantic changes across concepts and editing strengths, and Table E.3 summarizes them by concept, direction, and strength range. At primary strengths, both directions of all four concepts have lower mean semantic scores after scaffold-response blocking than after matched ordinary-response blocking. Scaffold blocking therefore not only causes greater image changes but also weakens target-semantic changes relative to ordinary blocking.

Table E.3: Target-semantic evaluation by concept, direction, and strength range. Mean scores for the three intervention conditions and the paired contrast $D = S_O - S_H$ are reported over 10,000 prompts. (a) Summaries for positive, negative, and combined directions. (b) Overall means across concepts. A positive D means that target change is weaker after scaffold-response blocking than after ordinary-response blocking.

(a) Grouped means

Concept feature	Range	Direction	S_C	S_H	S_O	D
`up.0.1/f3997`	Primary	Negative	6.9897	6.9688	7.1450	+0.1762
		Positive	6.7000	6.9650	6.9900	+0.0250
		Both	6.8449	6.9669	7.0675	+0.1006
	Stress	Negative	3.2775	3.2237	3.2250	+0.0013
		Positive	5.3262	5.3063	5.3800	+0.0737
		Both	4.3019	4.2650	4.3025	+0.0375
`up.0.1/f1635`	Primary	Negative	7.3620	7.3537	7.3637	+0.0100
		Positive	9.8142	9.7587	9.7875	+0.0288
		Both	8.5881	8.5563	8.5756	+0.0194
	Stress	Negative	10.0000	10.0000	10.0000	+0.0000
		Positive	10.0000	10.0000	10.0000	+0.0000
		Both	10.0000	10.0000	10.0000	+0.0000
`down.2.1/f4998`	Primary	Negative	6.3177	6.2479	6.3204	+0.0725
		Positive	7.5962	7.4032	7.5119	+0.1088
		Both	6.9569	6.8256	6.9162	+0.0906
	Stress	Negative	0.0500	0.0275	0.0275	+0.0000
		Positive	3.9417	3.8400	3.9475	+0.1075
		Both	1.9959	1.9338	1.9875	+0.0537
`down.2.1/f2301`	Primary	Negative	6.3614	6.3674	6.4349	+0.0675
		Positive	6.9916	6.9466	7.1204	+0.1737
		Both	6.6765	6.6570	6.7777	+0.1206
	Stress	Negative	0.0000	0.0000	0.0250	+0.0250
		Positive	0.0000	0.0000	0.0000	+0.0000
		Both	0.0000	0.0000	0.0125	+0.0125

(b) Overall summary

Range	Condition	Semantic score	D
Primary	Concept-only (C)	7.2666	+0.0828
	Block scaffold (H)	7.2514	
	Block ordinary (O)	7.3342	
Stress	Concept-only (C)	4.0744	+0.0259
	Block scaffold (H)	4.0497	
	Block ordinary (O)	4.0756	

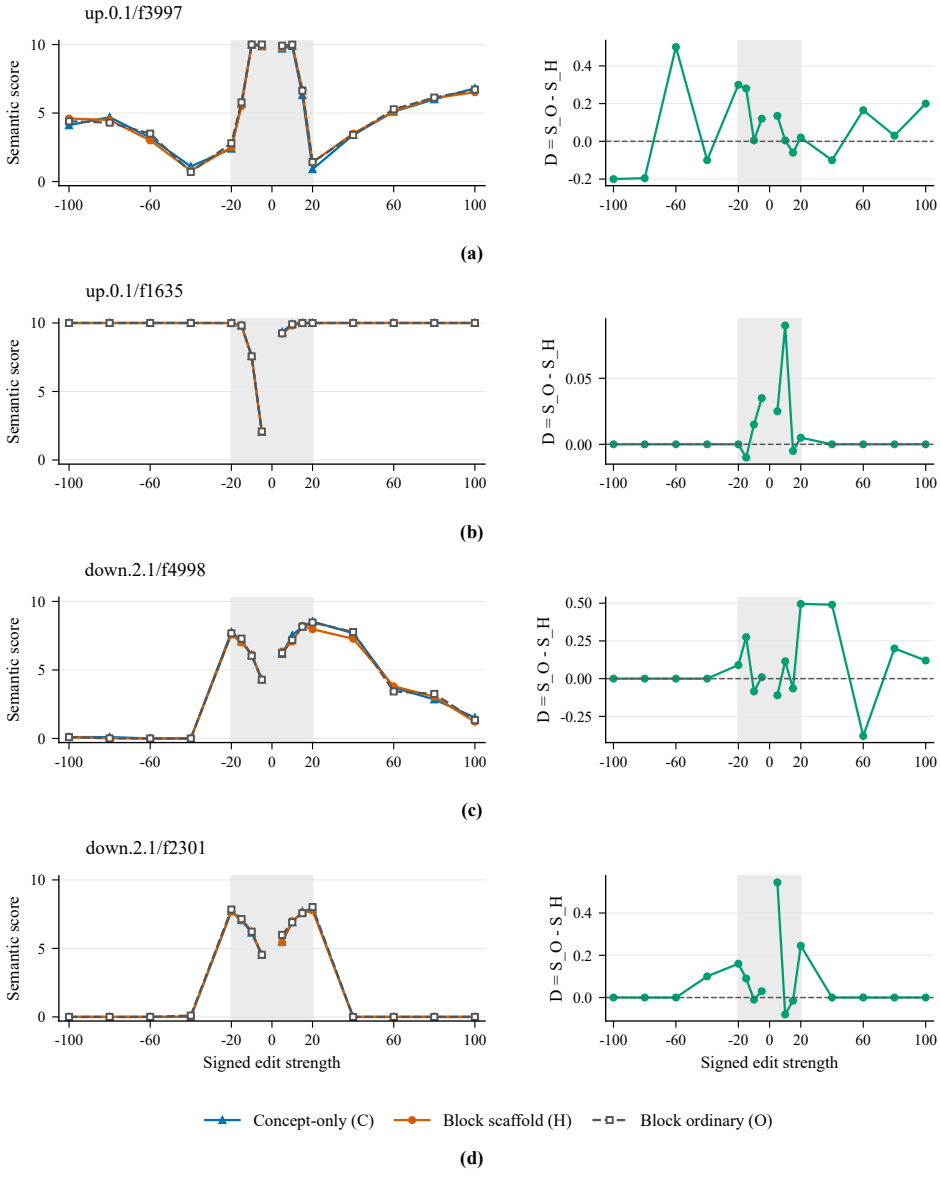


Figure E.6: Target semantics across editing strengths. Each row corresponds to a concept feature and shows scores S_C, S_H, S_O for concept-only editing, scaffold-response blocking, and matched ordinary-response blocking, together with the paired contrast $D = S_O - S_H$. Scores are assigned by `GPT-5.6-Sol`; each point averages 10,000 paired prompts at the same concept and strength. Gray shading marks the primary-strength range. Positive and negative directions are blue or cyan versus orange-red for subject color, green versus pink or magenta for background color, cartoon or illustrative versus photographic rendering for `f4998`, and more versus less threatening appearance for `f2301`.



Figure E.7: Response blocking along the same-step path. Columns show the unedited image, concept-only edit, scaffold-response blocking, and matched ordinary-response blocking, respectively. From top to bottom, rows show cartoon-like editing of a camera and stone cottage, photographic rendering of a teapot illustration, increased threat in a knight, and reduced threat in a stern character. The first three rows use concept feature `down.2.1/f4998`, and the last two use `down.2.1/f2301`; blocking takes place at the same-step downstream module `mid.0`. Selected coordinates are restored to their unedited activations while the target concept intervention is retained. Scaffold-response blocking introduces more pronounced structural, material, or non-target content changes. Random seeds are 123, 123, 42, 8, and 42 from top to bottom.

Same-step image comparisons are shown in Figure E.7. Concept-only editing produces the intended changes in style or expression, and blocking matched ordinary responses yields similar results. Blocking scaffold responses instead introduces pronounced non-target changes. The camera's lens and body contours become disorganized, and the teapot handle turns into a spout-like component. Vegetation encroaches on the cottage facade, obscuring windows and the path. The two portrait examples exhibit a deformed head contour and a carved facial texture, respectively, compromising the original facial structure and photorealistic appearance.

Such damage is also visible along the cross-step path, as shown in Figure E.8. The ceramic pitcher remains clearly blue under all three editing conditions, but blocking scaffold responses produces abnormalities in its handle and attachment points, while the surface loses its smooth, continuous form. In the fountain example, concept-only editing and ordinary-response blocking retain a relatively clear fountain structure. Scaffold-response blocking instead reduces the background to layered textures, with substantial loss of fountain ornamentation and water-flow details.

Figures E.7 and E.8 show that a clear target change does not imply an intact editing result. The pitcher remains blue after scaffold-response blocking, yet develops abnormalities in its handle and surface; other examples also lose structure, materials, or non-target content. Together, semantic scores and image comparisons indicate that scaffold responses affect not only the strength of target attributes, but also the preservation of image content as the concept changes.

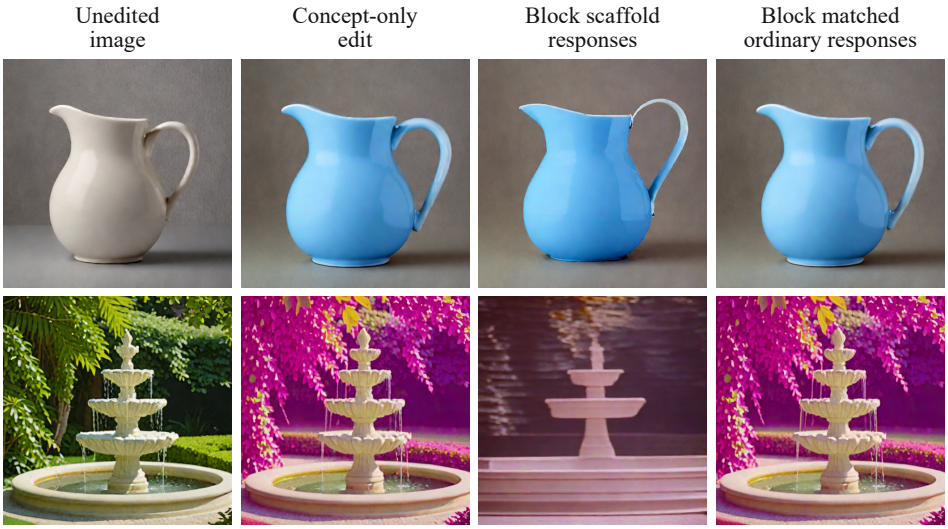


Figure E.8: Response blocking along the cross-step path. Column order matches Figure E.7. The top row uses `up.0.1/f3997` to increase the pitcher's blue coloration, and the bottom row uses `up.0.1/f1635` to shift the background toward pink. Blocking takes place at `down.2.1` in subsequent denoising steps. Scaffold-response blocking does not eliminate the blue change in the pitcher, but damages its handle and surface; in the fountain example, it causes more pronounced loss of structure and detail. Blocking matched ordinary responses yields results closer to concept-only editing. Random seeds are 42 and 7 for the top and bottom rows, respectively.

F Local Analysis and Implementation of SCFE

This appendix analyzes the local response estimate, constrained projection, and editing effects used by SCFE, and gives the complete algorithm. We estimate a separate response map for every directed pair from a concept-feature location to a scaffold-feature location. For compact notation, spatial positions, SAE positions, and denoising steps are included in scaffold-coordinate indices; concept-control coordinates follow the definition in Section 4.

The projection below includes the nonnegativity constraint of the SDXL-Turbo SAEs released by Surkov et al. (2025). For the signed FLUX SAE, we omit this constraint and use the interval $[-\kappa\sigma_h, \kappa\sigma_h]$ in Proposition 2; response estimation and amplitude bounds are unchanged.

F.1 Local Response Estimation

For prompt p and fixed initial noise, let

$$F_p : \mathbb{R}^{d_C} \to \mathbb{R}^{d_H} \tag{14}$$

map a concept intervention to later scaffold activations. Its input $\eta \in \mathbb{R}^{d_C}$ is an increment at the concept-control coordinates, and $F_p(\eta)$ is the scaffold state when the trajectory reaches the specified downstream location. Probes use the unedited trajectory as reference and apply concept interventions only at the specified timestep, without simultaneous scaffold updates. Hence,

$$F_p(0) = z^0_{p,H}, \qquad F_p(\Delta z_C) = z^{(\Delta z_C)}_{p,H}. \tag{15}$$

SCFE estimates the Jacobian of F_p near the unedited state with a central finite difference. For concept coordinate c,

$$G_p[:, c] = \frac{F_p(\epsilon e_c) - F_p(-\epsilon e_c)}{2\epsilon}, \tag{16}$$

where e_c is the cth unit vector and $G_p[:, c] \in \mathbb{R}^{d_H}$ records the local response of candidate scaffold coordinates to c. Probes are run separately for each concept and timestep; ϵ denotes the corresponding probe magnitude in actual activation units.

Proposition 1 (Central-difference error). Assume that the SAE Top-K support and nonnegative activation state remain fixed on $\{te_c : |t| \le \epsilon\}$ and that F_p is three times continuously differentiable there. If

$$\left\| D^3 F_p(te_c)[e_c, e_c, e_c] \right\|_2 \le M_{p,c}, \qquad |t| \le \epsilon, \tag{17}$$

then

$$\|G_p[:, c] - J_p(0)e_c\|_2 \le \frac{M_{p,c}}{6}\epsilon^2, \tag{18}$$

where $J_p(0)$ is the Jacobian of F_p at the unedited state.

Proof. Taylor expansion of $F_p(\epsilon e_c)$ and $F_p(-\epsilon e_c)$ along e_c gives

$$F_p(\epsilon e_c) = F_p(0) + \epsilon J_p(0)e_c + \frac{\epsilon^2}{2} D^2 F_p(0)[e_c, e_c] + \frac{\epsilon^3}{6} R_+, \tag{19}$$

$$F_p(-\epsilon e_c) = F_p(0) - \epsilon J_p(0)e_c + \frac{\epsilon^2}{2} D^2 F_p(0)[e_c, e_c] - \frac{\epsilon^3}{6} R_-, \tag{20}$$

where $\|R_+\|_2, \|R_-\|_2 \le M_{p,c}$. Subtracting the two equations and dividing by 2ϵ cancels the even-order terms:

$$G_p[:, c] = J_p(0)e_c + \frac{\epsilon^2}{12}(R_+ + R_-). \tag{21}$$

The triangle inequality proves the result.

The proposition shows that the response-estimation error decreases as ϵ^2 within a local region where the SAE sparse support does not change.

If J_p satisfies the local Lipschitz condition

$$\|J_p(\eta) - J_p(0)\|_2 \le L_p \|\eta\|_2, \tag{22}$$

then a concept edit Δz_C within that local region satisfies

$$\|F_p(\Delta z_C) - F_p(0) - G_p \Delta z_C\|_2 \le \frac{L_p}{2} \|\Delta z_C\|_2^2 + \frac{\epsilon^2}{6} \sum_{c \in C} M_{p,c} |\Delta z_c|. \tag{23}$$

The first term is due to nonlinearity in the generation network and the second is the finite-difference error. Thus, $G_p \Delta z_C$ approximates the scaffold response induced by the concept edit near the reference state, while clipping limits the actual update applied along that direction.

F.2 Cross-Prompt Stable Responses

Across P calibration prompts, let G_p^{cal} denote the finite-difference responses used for selection and define their mean matrix as

$$\bar{G} = \frac{1}{P}\sum_{p=1}^{P} G_p^{\mathrm{cal}}. \tag{24}$$

This average is used to select features for joint editing, not directly as the update response for the requested prompt. For a fixed concept c and denoising step, let Ω_f contain all spatial coordinates of candidate scaffold feature f. The spatially averaged unit response for each calibration prompt and its cross-prompt mean are

$$s_{p,f} = \frac{1}{|\Omega_f|}\sum_{h\in\Omega_f} G_p^{\mathrm{cal}}[h,c], \qquad \bar{s}_f = \frac{1}{P}\sum_{p=1}^{P} s_{p,f}. \tag{25}$$

Here, $\bar{s}_f$ is the mean unit response of candidate feature f, equivalently obtained by averaging the corresponding spatial coordinates of $\bar{G}[:,c]$. To prevent near-zero responses from affecting direction statistics, define the valid calibration-prompt set

$$\mathcal{V}_f = \left\{ p : \frac{|s_{p,f}|}{\sigma_f + 10^{-8}} > 10^{-6} \right\}, \tag{26}$$

where σ_f is the feature's natural activation scale at the corresponding timestep. We measure cross-prompt directional consistency by

$$\rho_f = \left| \frac{1}{|\mathcal{V}_f|}\sum_{p\in\mathcal{V}_f} \mathrm{sign}(s_{p,f}) \right|. \tag{27}$$

If $\mathcal{V}_f$ is empty, we set $\rho_f = 0$. Near-zero responses do not enter the denominator of this statistic. Within the valid calibration-prompt set, the majority direction occupies a fraction $(1+\rho_f)/2$. For example, $\rho_f = 0.8$ means that 90% of valid calibration prompts share one response direction. SCFE retains features with $\rho_f \geq \tau_s$ and selects the K largest $|\bar{s}_f|$ values:

$$\mathcal{F}_C = \mathrm{TopK}_{f:\rho_f\geq\tau_s} |\bar{s}_f|, \qquad H_C = \bigcup_{f\in\mathcal{F}_C} \Omega_f. \tag{28}$$

If fewer than K features qualify, all are retained. Cross-prompt averaging selects reusable features, the sign threshold removes features whose response direction repeatedly changes, and Top-K sets the number of jointly edited features. For the requested prompt p, we separately estimate the selected features' spatial responses and use $r_p = G_p \Delta z_C$ for the update.

F.3 Closed-Form Constrained Projection

Let z_H^{pre} denote the current trajectory's scaffold state at the specified downstream location before this scaffold update. It may include effects of joint updates at earlier timesteps. SCFE does not replace this state, but finds the executable update nearest the predicted gain λr_p:

$$u^\star = \arg\min_{u\in\mathcal{U}} \|u - \lambda r_p\|^2_{D_H^{-1}}, \tag{29}$$

where

$$D_H = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_{d_H}^2) \tag{30}$$

is defined on a coordinate space restricted to features whose natural scales exceed the numerical tolerance. For $h \in \Omega_f$, we set $\sigma_h = \sigma_f$, sharing each feature's root-mean-square activation scale from unedited data across its spatial coordinates. Thus D_H is positive diagonal. The feasible set is

$$\mathcal{U} = \left\{ u : \mathrm{supp}(u) \subseteq H_C,\ |u_h| \leq \kappa\sigma_h,\ z_{H,h}^{\mathrm{pre}} + u_h \geq 0 \right\}. \tag{31}$$

The three constraints restrict the joint-coordinate set, update amplitude, and nonnegativity of SAE activations, respectively.

Proposition 2 (Coordinatewise projection). For $h \in H_C$, define

$$\ell_h = \max\{-\kappa\sigma_h, -z_{H,h}^{\mathrm{pre}}\}, \qquad b_h = \kappa\sigma_h. \tag{32}$$

The optimization has the unique solution

$$u_h^\star = \begin{cases} \min\{b_h, \max\{\ell_h, \lambda r_{p,h}\}\}, & h \in H_C, \\ 0, & h \notin H_C. \end{cases} \tag{33}$$

Proof. Because D_H^{-1} is positive diagonal, the objective is

$$\|u - \lambda r_p\|^2_{D_H^{-1}} = \sum_{h=1}^{d_H} \frac{(u_h - \lambda r_{p,h})^2}{\sigma_h^2}. \tag{34}$$

The feasible set is also a product of coordinatewise intervals. The support constraint gives $u_h = 0$ for $h \notin H_C$. For $h \in H_C$, the amplitude and nonnegativity constraints give $u_h \in [\ell_h, b_h]$. The minimizer of a one-dimensional quadratic over this interval is the projection of $\lambda r_{p,h}$ onto it, yielding the stated form.

The joint state is $z_H^{\text{joint}} = z_H^{\text{pre}} + u^\star$. SCFE therefore needs no numerical optimizer: coordinatewise clipping implements the constrained projection exactly rather than approximately. The implementation writes only the selected coordinates' decoder increments back to the current network output, retaining its SAE reconstruction residual.

F.4 Sufficient Conditions for Local Editing Improvement

Let $T(z_H)$ be the target-semantic effect, with larger values denoting a more complete requested edit, and let $D(z_H)$ be non-target content change or image-quality cost, with smaller values preferred. Both are compared to the state z_H^{pre} before this scaffold update, holding the preceding trajectory and all other update settings fixed.

The nonnegativity constraint can clip negative responses when $z_{H,h}^{\text{pre}} = 0$. Define the feasible direction of the projection at $\lambda = 0$ as

$$v_h = \begin{cases} r_{p,h}, & h \in H_C,\ z_{H,h}^{\text{pre}} > 0, \\ \max\{r_{p,h}, 0\}, & h \in H_C,\ z_{H,h}^{\text{pre}} = 0, \\ 0, & h \notin H_C. \end{cases} \tag{35}$$

For sufficiently small positive λ, the natural-amplitude bound is inactive and $u^\star(\lambda) = \lambda v$.

Proposition 3 (Local Pareto improvement). Assume that T and D are differentiable at z_H^{pre}. If

$$\nabla_H T(z_H^{\text{pre}})^\top v > 0, \qquad \nabla_H D(z_H^{\text{pre}})^\top v < 0, \tag{36}$$

then there exists $\delta > 0$ such that every $0 < \lambda < \delta$ satisfies

$$T\big(z_H^{\text{pre}} + u^\star(\lambda)\big) > T\big(z_H^{\text{pre}}\big), \tag{37}$$

$$D\big(z_H^{\text{pre}} + u^\star(\lambda)\big) < D\big(z_H^{\text{pre}}\big). \tag{38}$$

Proof. Differentiability and $u^\star(\lambda) = \lambda v$ give

$$T\big(z_H^{\text{pre}} + \lambda v\big) - T\big(z_H^{\text{pre}}\big) = \lambda \nabla_H T^\top v + o(\lambda), \tag{39}$$

$$D\big(z_H^{\text{pre}} + \lambda v\big) - D\big(z_H^{\text{pre}}\big) = \lambda \nabla_H D^\top v + o(\lambda). \tag{40}$$

The first directional derivative is strictly positive and the second strictly negative. Since $o(\lambda)/\lambda \to 0$, there is a sufficiently small $\delta > 0$ for which the first-order terms determine both signs.

When $\nabla_H D^\top v = 0$, if D is twice differentiable and

$$v^\top \nabla_H^2 D(z_H^{\text{pre}}) v < 0, \tag{41}$$

a second-order expansion likewise gives a local cost reduction along v.

Proposition 3 gives local sufficient conditions for scaffold updates to strengthen target semantics and reduce image cost simultaneously. Section 5 compares target semantics and image preservation between SCFE and concept-only editing at the same concept-edit strength.

F.5 Complete Algorithm and Computational Cost

Algorithm 1: Scaffold-Coordinated Feature Editing.

Input: a frozen diffusion model and SAE; concept-coordinate set C; scaffold-candidate set H; calibration prompts; the requested prompt and initial noise; concept edit Δz_C; probe magnitudes for calibration and the current prompt; and τ_s, K, λ, and κ.

1. For every calibration prompt and concept coordinate, apply small positive and negative interventions and record scaffold activations at the specified same-step downstream SAE location.
2. Compute G_p^{cal} with central differences and obtain each candidate feature's spatially averaged unit response $s_{p,f}$.
3. Compute the cross-prompt mean $\bar{s}_f$ and sign consistency ρ_f over the valid calibration-prompt set.
4. Among features satisfying $\rho_f \geq \tau_s$, select up to K with the largest $|\bar{s}_f|$ and form H_C from their spatial coordinates.
5. Fix the initial noise for the requested prompt. Probe each concept and timestep symmetrically, record the selected scaffolds' complete spatial responses, and compute G_p and $r_p = G_p \Delta z_C$.
6. Start the actual concept edit from the same initial noise and read z_H^{pre} when the current trajectory reaches the specified downstream location.
7. Compute $u^\star$ using the closed form in Proposition 2. With $z_H^{\mathrm{pre}} + u^\star$ as the joint activation, write the selected coordinates' decoder increments back to the network.
8. Repeat current-state readout and joint updates at subsequent specified timesteps, complete the remaining denoising process, and output the image.

In the SDXL-Turbo experiments, calibration and current-prompt probes use the same finite-difference form but different actual magnitudes. Selection probes use ± 5 raw activation units, current-prompt probes use $\pm 5 m_c$, and the actual concept increment is $\Delta z_c = \alpha m_c$. Here, m_c is the mean concept activation from the statistics supplied with the SAE, and α is edit strength. The FLUX experiments use native latent units, with $\epsilon = 1$ for both calibration and current-prompt probes; see Appendix H. Each central-difference denominator is the difference between the actual positive and negative probe increments.

Both calibration and current-prompt response estimation are performed separately for each concept and denoising step. Every prompt–concept–timestep combination requiring a response estimate needs two probes, one positive and one negative. The current implementation runs a complete denoising trajectory for each probe and reads the response at the specified location, rather than executing only a local forward segment from concept injection to scaffold readout.

G SUPPLEMENTARY EXPERIMENTS ON JOINT EDITING

G.1 EXPERIMENTAL SETTINGS AND CONTROL DEFINITIONS

This appendix supplements the SDXL-Turbo joint-editing experiments. Quantitative evaluation uses the same 10,000 prompts and initial noise as the main experiments, testing cartoonization and photorealistic editing with `down.2.1/f4998`, and increases and decreases in threatening appearance with `down.2.1/f2301`. Concept interventions are applied at all denoising steps with signed strengths $\{-20, -15, -10, -5, 5, 10, 15, 20\}$. These experiments use the `down.2.1`→`mid.0` path, with joint scaffold updates at the downstream `mid.0` in the same step. All methods retain the same concept injection and differ only in their additional feature updates, as summarized in Table G.1.

For concepts that require particular objects to be present, we separately select prompts containing those objects and provide qualitative comparisons of dial details, plush texture, knitted texture, bunting, brooch ornamentation, ceiling-fan shape, and collar shape in Appendix G.5. These additional examples are not included in Table 1, Figures 5 and 6, or the existing quantitative statistics. We explain the reasons in detail in Appendix I.7.

The main quantitative experiments on SDXL-Turbo use $P = 10{,}000$ prompts for scaffold calibration. Joint updates use $\lambda = 1$ and $\kappa = 2$, with a cross-prompt sign-consistency threshold of $\tau_s = 0.6$. For each concept and denoising step, we select at most $K = 2$ scaffold features from candidates meeting the threshold, ranked by the absolute value of their mean response. If fewer than two qualify, all qualifying features are retained; if none qualify, no additional update is applied. Calibration probes use ± 5 raw activation units, while response probes for the current prompt use $\pm 5 m_c$, where m_c is the concept feature's mean activation from the statistics released with the SAE. The scaffold set is fixed after calibration, and editing uses the current prompt's complete spatial response. Prompt mismatch uses another prompt's response while retaining the receiving prompt's generation conditions. Spatial shift circularly translates the 16×16 response map by eight positions along each axis. These controls test prompt source and spatial correspondence, respectively. Responses used for updates are estimated through a small concept intervention, not remeasured from SCFE outputs. Section 4 and Appendix F define response estimation and the update.

Image preservation is measured by LPIPS, DINOv2 distance, and CLIP-I similarity. LPIPS uses AlexNet v0.1 weights; DINOv2 distance is the cosine distance between ViT-B/14 global image embeddings; CLIP-I is the cosine similarity between CLIP ViT-L/14 image embeddings. Each edited output is compared with the unedited image generated from the same prompt and initial noise.

GPT-5.6-Sol evaluates target semantics from image pairs. Given the unedited image, edited image, and predefined direction, it scores target-directed change rather than the absolute strength of the attribute in the edited image, on a 0–10 scale. Positive and negative directions correspond to cartoonization and photorealism for `f4998`, and increased and decreased threatening appearance for `f2301`. Unrelated distortion or disappearance of the subject does not count as successful editing. Figure G.1 presents semantic scores alongside preservation.

Table G.1: Experimental roles of the controls. All six methods retain the same target-concept intervention. The four functional controls change feature identity, response source, or spatial positions in the joint update.

Method	Additional update	Question tested
Concept-only	None; only the target concept feature is edited	Concept-only baseline
SCFE	Scaffold response for the current prompt	Can scaffold responses improve editing?
Random-feature joint editing	Randomly selected features replace scaffolds	Are arbitrary additional features equally effective?
Responsive non-scaffold editing	Non-scaffolds that respond to concept intervention	Is responsiveness sufficient to replace scaffold identity?
Prompt mismatch	Scaffold response from another prompt	Must the response match current content?
Spatial shift	Spatially translated current scaffold response	Must the response retain its spatial correspondence?

G.2 Results Across Editing Conditions

Figure G.1 gives the full strength scans for all methods, while Figure 5 in the main text shows gains over Concept-only. Averaged over both concepts and positive and negative directions, SCFE improves target-semantic scores and all three preservation metrics at each of the four strength magnitudes, with mean gains increasing with strength. Semantic and CLIP-I differences are small at low strengths, and the joint gains become clearer for stronger edits.

SCFE's target-semantic scores increase with strength for both concepts and both editing directions, but its gains over Concept-only vary across conditions. For example, when increasing threatening appearance, its semantic score is lower than Concept-only at strengths $+5$ and $+15$, but higher at $+10$ and $+20$. At an absolute strength of 20, both concepts in both directions show gains in semantics and all three preservation metrics.

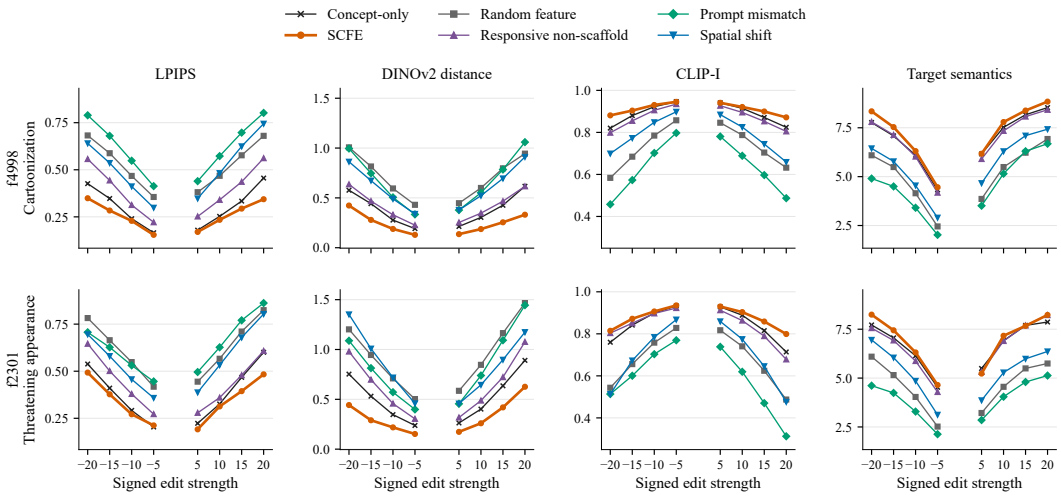


Figure G.1: Full strength scans for two concepts on SDXL-Turbo. Rows correspond to `f4998` and `f2301`; columns report LPIPS, DINOv2 distance, CLIP-I similarity, and target-semantic score. Each point averages over the 10,000 evaluation prompts in the corresponding condition. Positive and negative directions are connected separately.

G.3 Condition-Specific Functional Controls

Feature replacement and response mismatch do not reproduce SCFE's overall gains. In Figure G.2, replacing joint updates with random or responsive non-scaffold features does not provide the same preservation improvement. Responding to a concept intervention alone is insufficient to replace the role of the scaffold population.

Responses must also correspond to the current generated content. Prompt mismatch changes their content source; spatial shift retains response values but changes where they act. Both controls increase image deviation overall without yielding higher mean target-semantic scores. This differs from SCFE's joint improvement in semantics and preservation. Effective joint editing requires both appropriate features and information about how they respond to the current intervention.

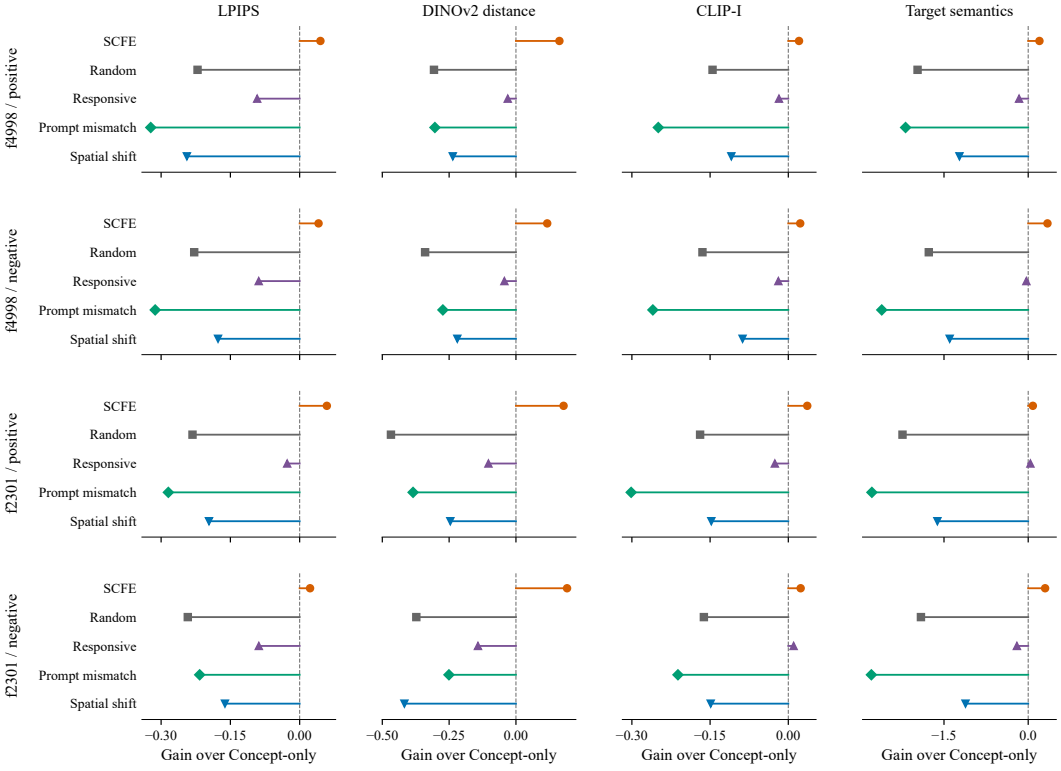


Figure G.2: Functional controls by concept and editing direction. Rows are positive and negative `f4998`, followed by positive and negative `f2301`. Columns report three preservation gains and the target-semantic gain over Concept-only. Within each concept and direction, results average over four strengths and the 10,000 evaluation prompts. Positive values indicate improvement in every column; zero denotes equality with Concept-only.

G.4 Additional Editing Examples

Figure G.3 shows cartoonization and photorealistic editing. For the coffee machine and scooter, style changes under concept-only editing are accompanied by the loss of local components and their connections, which SCFE preserves better. The teapot-and-cup example shows another difference: concept-only editing loses the cup while changing material appearance, whereas SCFE retains both objects. Joint updates therefore preserve not only colors or textures but also non-target content.

Figure G.4 supplements increases and decreases in threatening appearance. The positive examples reveal incidental changes to the eyes and brows, nose and mouth, and head contours. Negative examples facilitate comparisons of appearance, hairstyle, and the original rendering style. SCFE can change the target expression without necessarily replacing these other attributes.

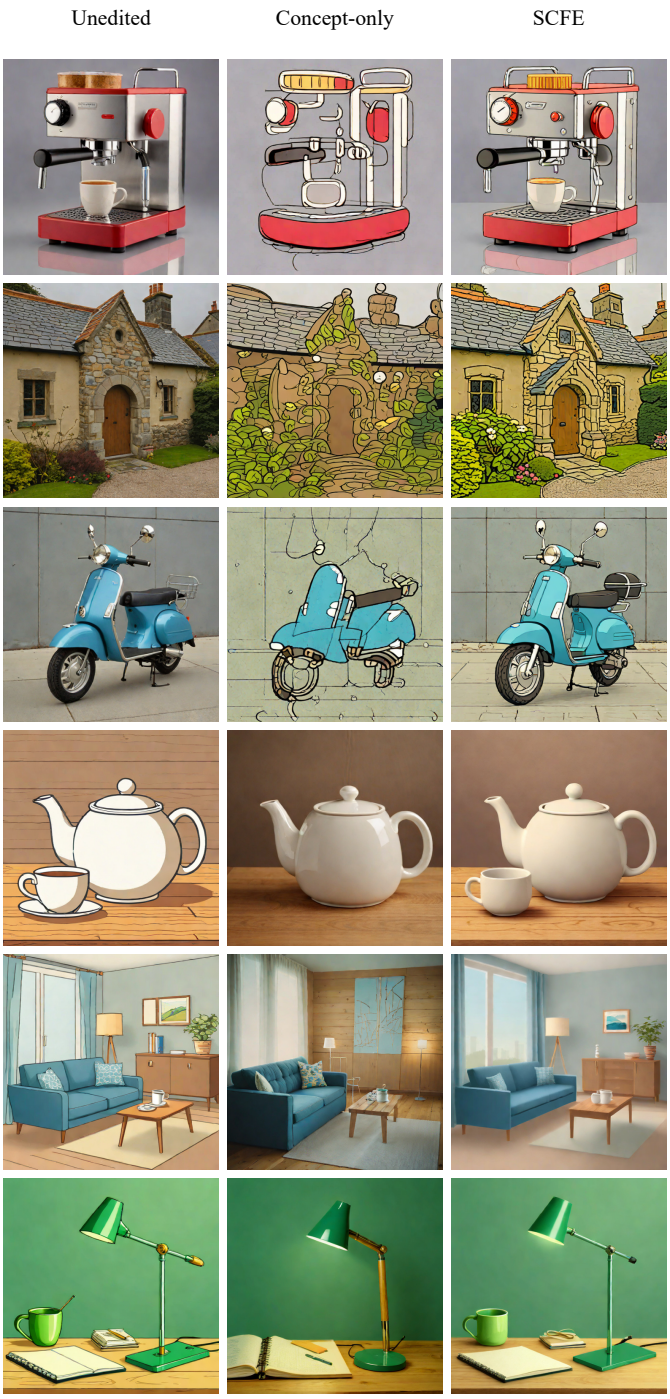


Figure G.3: Additional cartoonization and photorealistic-editing examples. Columns show the unedited image, Concept-only, and SCFE. The first three rows cartoonize a coffee machine, cottage, and scooter; the last three make a teapot and cup, an interior, and desktop objects more photorealistic. The images compare target-style changes and accompanying content preservation.



Figure G.4: Additional threatening-appearance edits. Columns show the unedited image, Concept-only, and SCFE. The first three rows increase threatening appearance and the last three decrease it, showing changes to expression, facial structure, and other aspects of appearance.

G.5 Editing Object-Dependent Concepts

We select prompts containing the relevant objects to provide eight qualitative examples covering seven target concepts. Figure G.5 shows edits to knitted texture, plush texture, collars, and brooches; Figure G.6 shows edits to bunting, a station clock, a ceiling fan, and a wristwatch. Each row compares the unedited image, Concept-only, and SCFE.

These examples extend the comparison from global style and a person's appearance to local materials, decoration, and object shape. Concept-only editing can also change facial appearance, the main object, or scene layout. While changing the target attribute, SCFE better preserves personal appearance, relationships between objects and their surroundings, and existing scene content.



Figure G.5: Editing material and appearance attributes. Each row shows the unedited image, Concept-only, and SCFE. From top to bottom: reducing knitted texture (`I2-16`), reducing plush texture (`I2-12`), adjusting collar shape (`I2-35`), and simplifying brooch ornamentation (`I2-28`). The images compare local attribute changes and preservation of people, objects, and surrounding content.

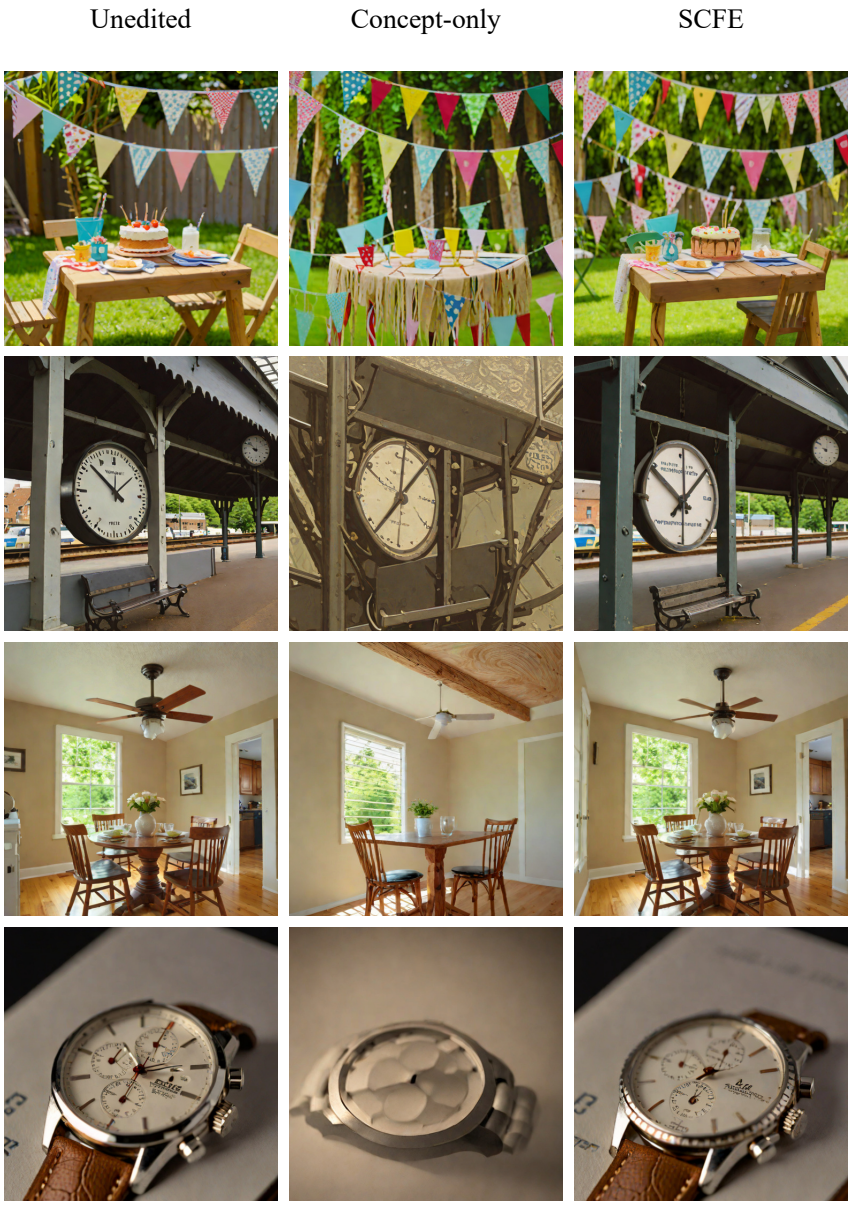


Figure G.6: Editing object details and scene decoration. Column order matches Figure G.5. From top to bottom: edits to bunting (`I2-23`), a station-clock dial (`I2-06`), ceiling-fan shape (`I2-33`), and a wristwatch dial (`I2-07`). Each row shows changes to the target details together with changes to object structure, surrounding objects, and scene layout.

H TRANSFER ACROSS DIFFUSION MODELS AND SAES

We test scaffold features and SCFE on FLUX.1-schnell with the public SAEs of Shabalin et al. (2025). We recompute cross-prompt reuse, identify scaffolds, and estimate concept responses in the new SAEs. The experiments examine the scaffold population, concept responses and blocking, and the benefits of joint editing in turn.

H.1 MODEL, SAES, AND TRANSFER SETTINGS

We use the FLUX SAEs released with the dictionary-learning work of Shabalin et al. (2025). Concept interventions occur at double-stream block 18, while responses and joint updates are located downstream at single-stream block 18 in the same step, abbreviated as `double18` and `single18`. Both SAEs have 65,536 features and TopK=64.

Scaffold discovery uses 10,000 prompts. Response, blocking, and joint-editing evaluations use the fixed 10,000 prompts and seeds from the preceding experiments. The 32 calibration prompts used to select joint-update features are separate from this evaluation set. Table H.1 summarizes the settings.

The concept features `double18/f18624` and `double18/f17869` increase or decrease anime style and Christmas elements, respectively. Signed strengths are $\{-60, -30, -10, 10, 30, 60\}$. These are additive amplitudes in the new SAE's native latent space, not strengths directly comparable to those used for SDXL. Concept interventions and downstream updates occur at steps 0–4; the last two denoising steps proceed normally.

SCFE follows the response estimation and constrained update in Section 4. Natural activation scales are full-sample RMS values from discovery, including zero activations. Calibration selects two scaffold features per concept and step by response-direction consistency. Diffusion-model and SAE weights remain fixed.

Table H.1: Main FLUX transfer settings. Scaffold sets and responses are obtained anew in the target model. Discovery, calibration, and formal evaluation prompt counts are listed separately.

Item	Setting
Diffusion model	FLUX.1-schnell
SAE locations	`double18`, `single18`; zero-based layers and steps
SAE features / TopK	65,536 / 64
Image resolution / denoising steps	256×256 / 7
Concept intervention $\rightarrow$ response and joint update	`double18` $\rightarrow$ `single18`, downstream in the same step
Intervention steps	0, 1, 2, 3, 4
Discovery / evaluation / calibration prompts	10,000 / 10,000 / 32
Cross-prompt reuse threshold	Nonzero spatial activation in at least 90% of discovery prompts
Concept features	`f18624`: anime style; `f17869`: Christmas elements
Signed concept strengths	$\pm 10, \pm 30, \pm 60$
Small response probe	$\epsilon = 1$
Calibration direction-consistency threshold / features per step	0.6 / 2
Joint-update parameters	$\lambda = 1$, $\kappa = 2$
Concept region / downstream response region	Exclude one boundary-token ring / all image tokens

H.2 Scaffold Features in the New Model

A highly reused cross-prompt population also exists in the FLUX SAEs. Table H.2 and Figure H.1 show that both `double18` and `single18` contain features meeting the reuse threshold at all five steps. They occupy only a small fraction of the dictionary but repeatedly participate in generating different content. This population is identified statistically, without reference to target concepts such as anime style or Christmas elements.

Fixed inputs do not explain all of this high reuse. We feed zero, global-mean, and spatial-position-mean inputs through the same SAEs. As Figure H.1 shows, most selected features are inactive under these inputs, while only some remain responsive.

Table H.2: Feature counts meeting the 90% reuse threshold across 10,000 prompts. Each step is counted separately. The last column is the union of selected features across five steps within the same SAE.

SAE location	Step 0	Step 1	Step 2	Step 3	Step 4	Step union
`double18`	338	388	358	359	348	602
`single18`	300	358	362	374	379	591

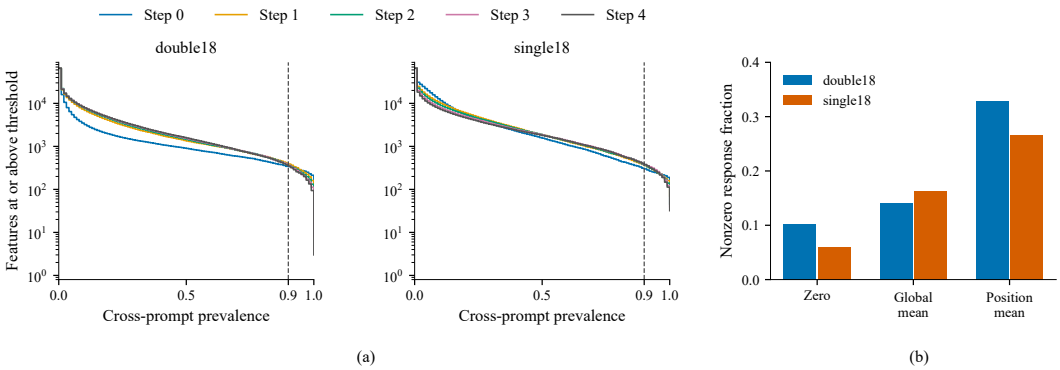
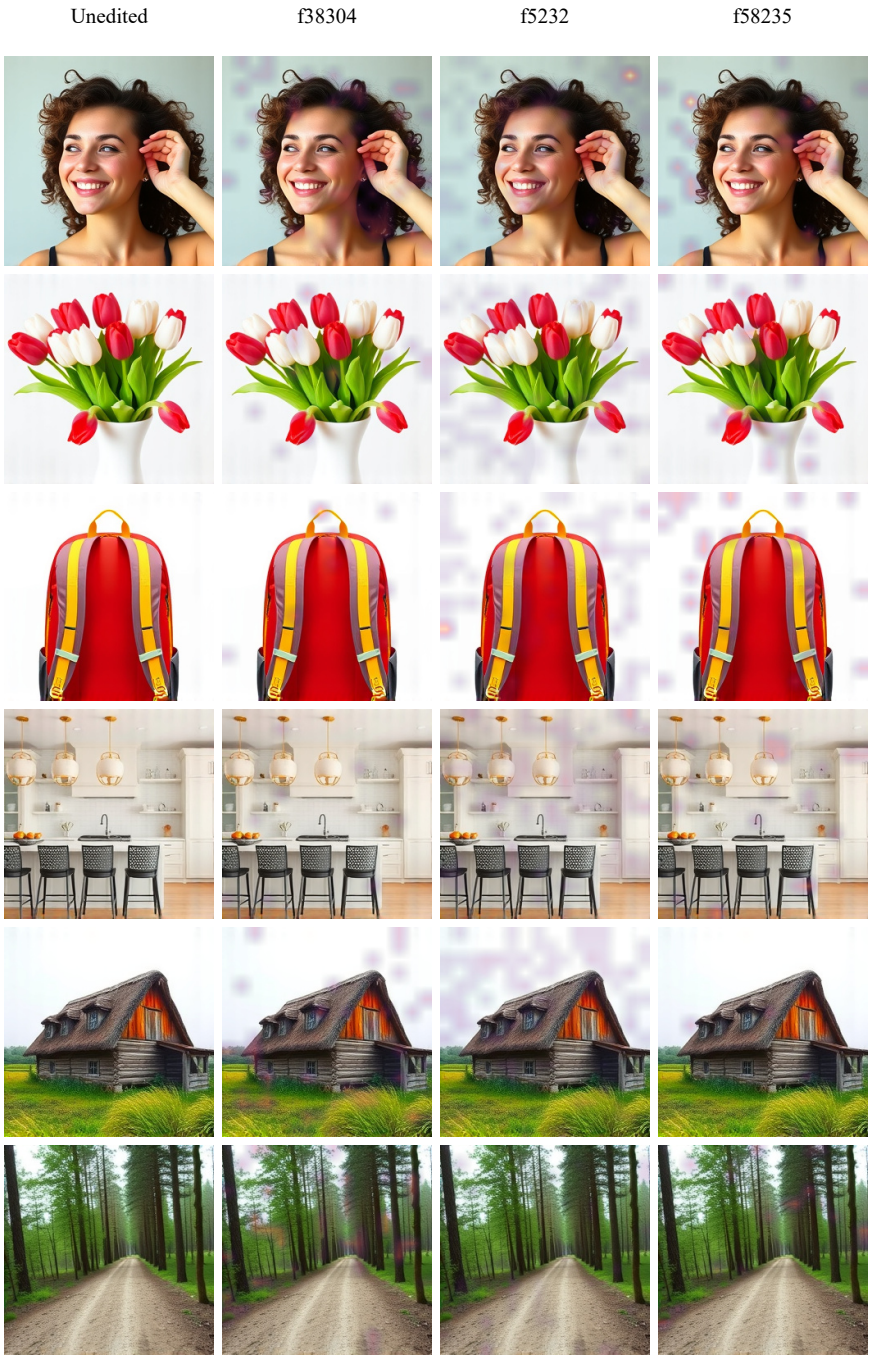


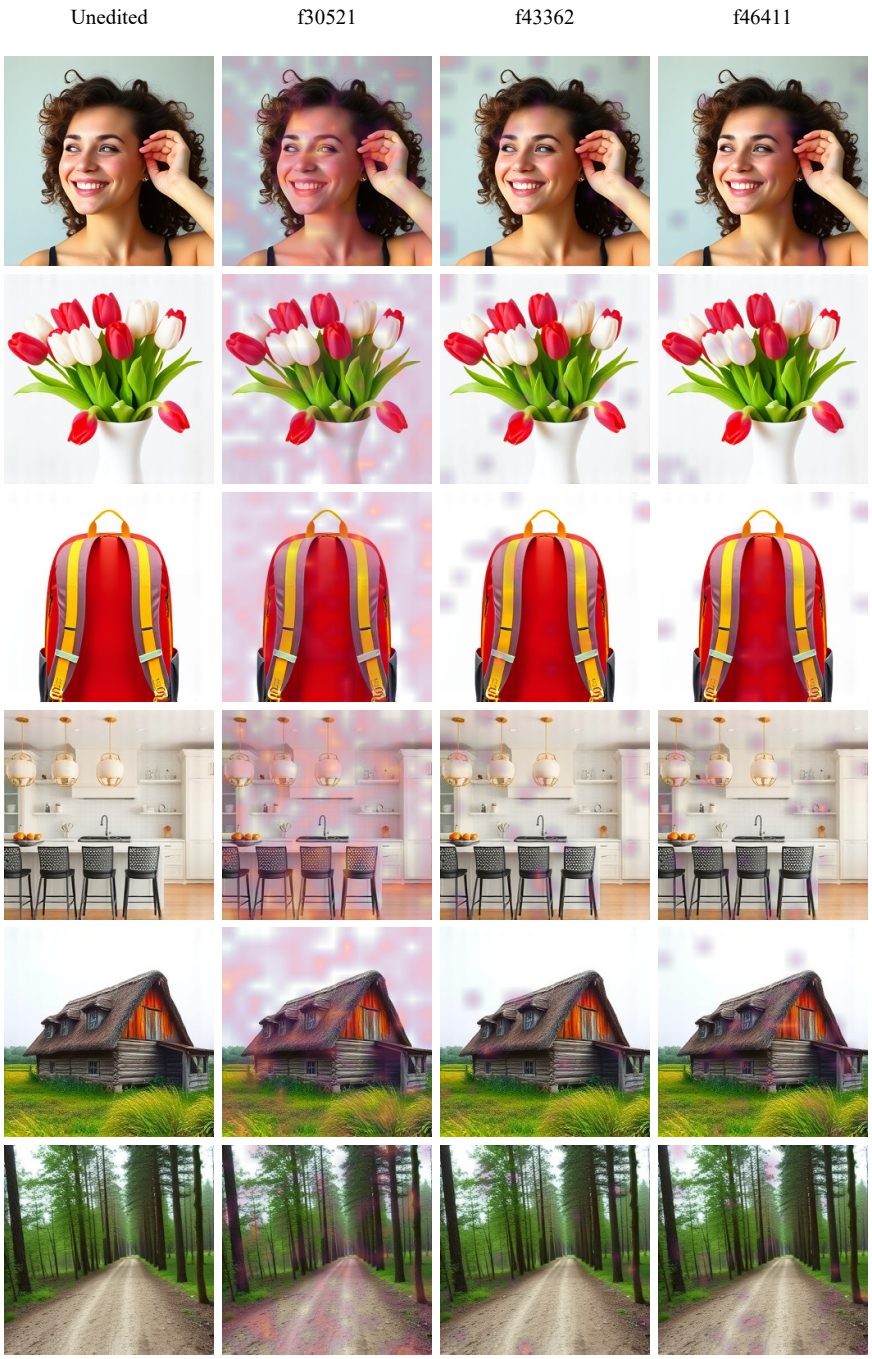
Figure H.1: Cross-prompt reuse and fixed-input controls in FLUX SAEs. (a) Reuse distributions across steps at both SAE locations, expressed as the number of features meeting each prevalence threshold; the scaffold-selection threshold is marked. (b) Fractions of selected features with nonzero responses to zero, global-mean, and position-mean inputs. Fixed-input units are feature–step pairs, grouped by SAE location. Both panels use the 10,000 discovery prompts.

Figure H.2 shows natural activations across content. Spatial activation magnitudes are upsampled and overlaid on the corresponding image using a shared color scale for each feature. High-response regions occur across people, objects, interiors, and natural scenes rather than consistently corresponding to one object category. Figure H.3 supplements this with scaffold interventions that can alter personal appearance, object structure, furnishings, and scene organization. These examples show the population's roles across different content.



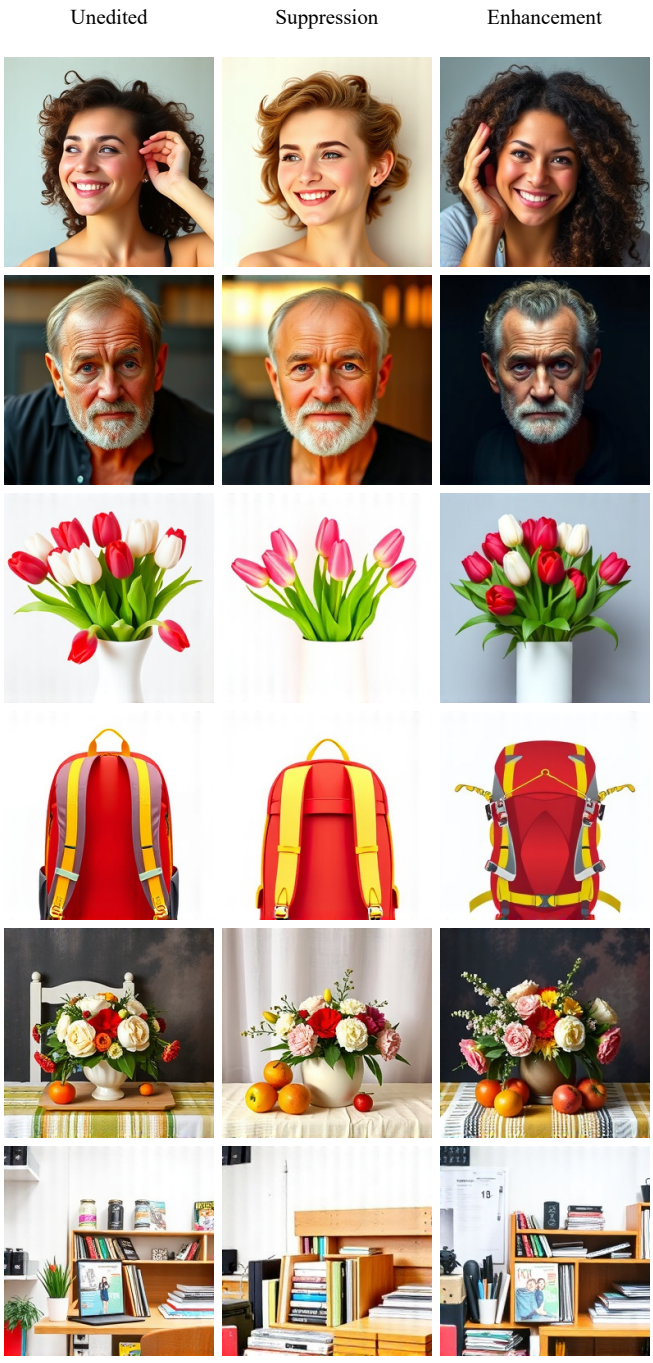
(a) double18

Figure H.2: Natural spatial activations of scaffold features across content. (a) `double18`; continued panel (b) shows `single18`. Each group contains three features and their paired originals, all at step 4. Colors show absolute activation magnitude with the same scale across prompts for a given feature.



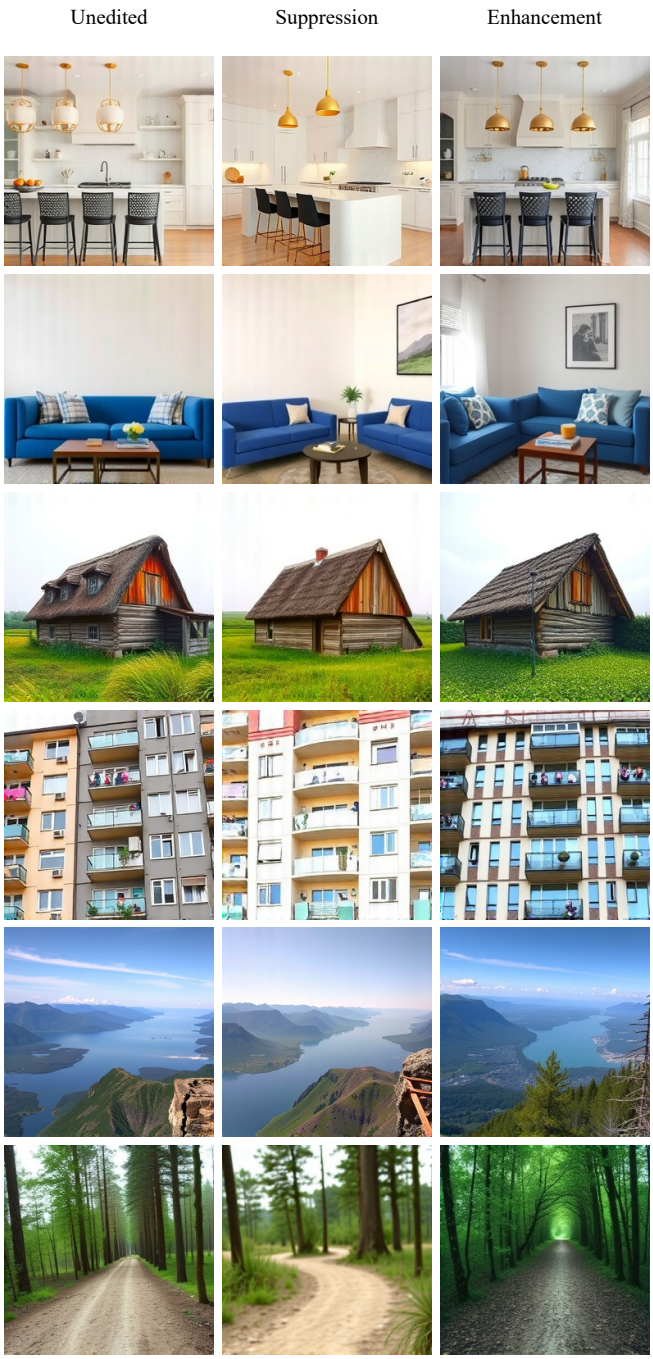
(b) single18

Figure H.2 (continued): (b) Three `single18` scaffold features, with the same definitions of paired originals, reference step, and shared color scales as on the preceding page.



(a)

Figure H.3: Scaffold-feature interventions across content. Columns show unedited images, suppression, and enhancement, covering portraits, objects, interiors, architecture, and natural scenes. Examples are organized by content to illustrate the image factors scaffold interventions can change; six further examples appear on the next page.



(b)

Figure H.3 (continued): Further scaffold interventions with the same column order. The specific changes depend on the generated content.

H.3 Concept Responses and Blocking Results

The newly identified scaffold population responds strongly to concept interventions. We compare `single18` latents along concept-edited and unedited trajectories, normalizing differences by natural activation RMS. In Figure H.4, mean normalized scaffold responses exceed ordinary-feature responses for both concepts, signs, all strengths, and every tested step. Highly reused features are thus not only active during natural generation but also recruited by target-concept changes.

To test their role, we retain the upstream concept edit and separately restore the unedited coefficients of downstream scaffolds and matched ordinary features. Ordinary features are matched one-to-one by response amplitude in the original residual space, with equal counts; this amplitude includes the decoder direction's inverse PCA and scale transforms. Matching uses unblocked concept-edit responses and is fixed before generating blocked outputs.

Blocking scaffolds changes the normal concept-edit result more than blocking matched ordinary features. Figure H.4 shows positive paired LPIPS and DINOv2 distance differences for every concept and signed strength. Distances are referenced to concept-only images, measuring how much blocking changes the editing result, not preservation relative to the original image. Figure H.5 illustrates what these changes look like.

In Figure H.5, scaffold blocking often weakens the target change. Anime-style outputs move back toward their unedited appearance, and photorealistic edits regain flat color regions and outlines. Added Christmas elements are suppressed, while edits that reduce them leave more original decorations behind. Matched ordinary blocking remains closer to normal concept editing. Scaffold responses therefore affect final images and participate in realizing the intended semantic change.

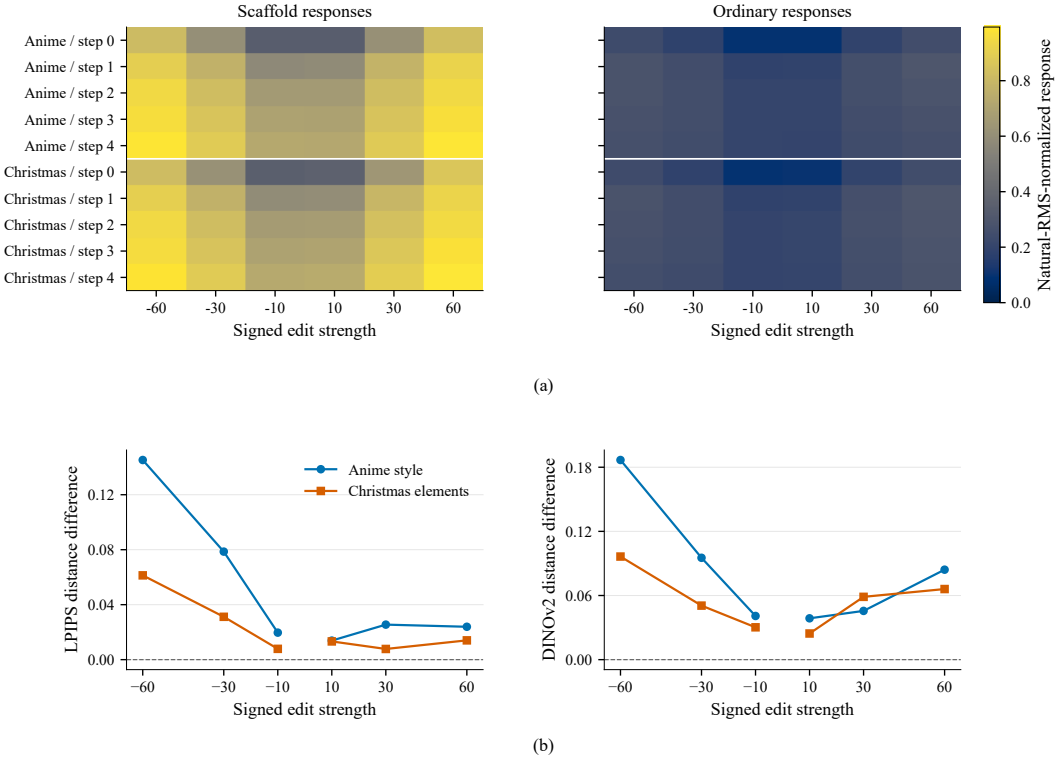
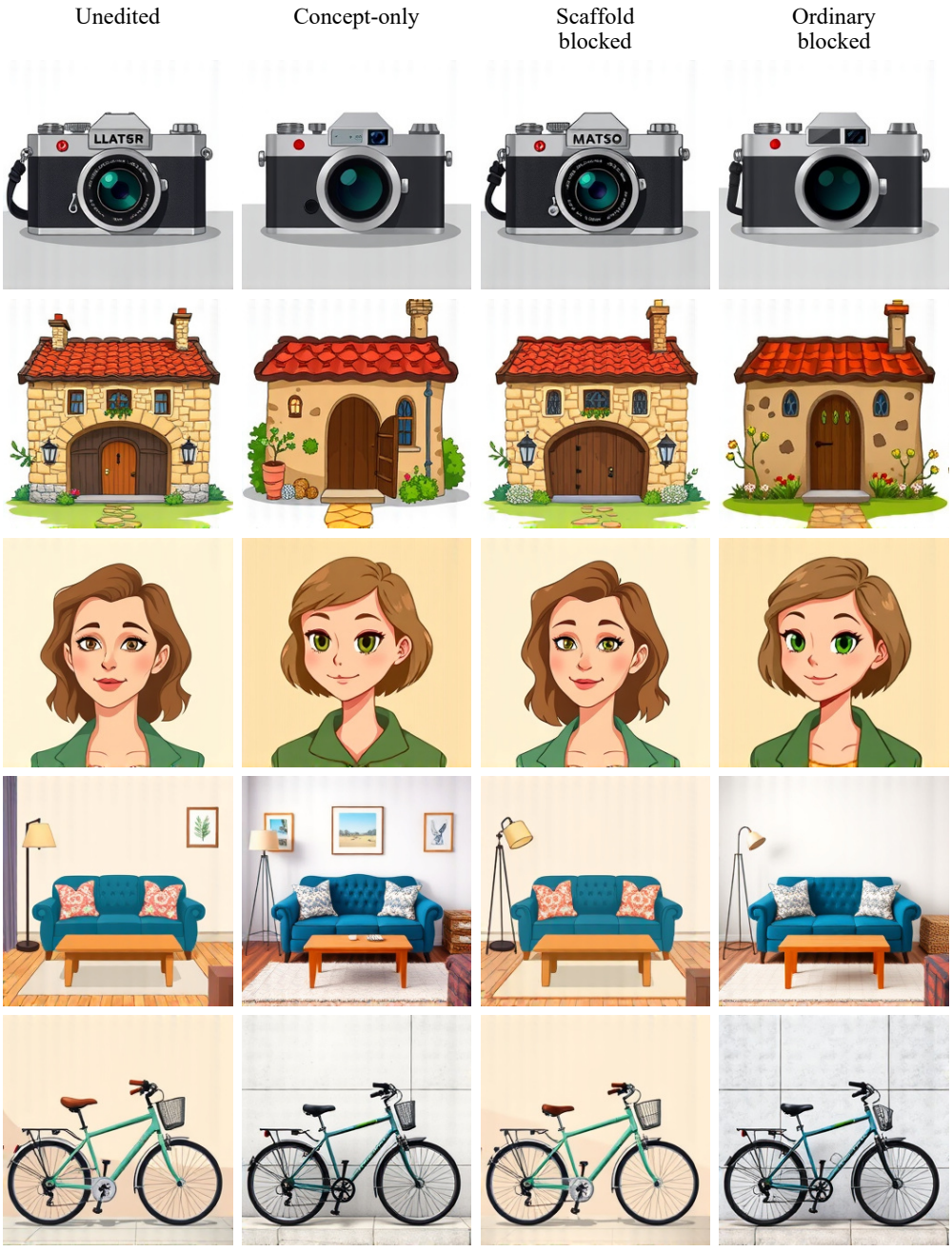


Figure H.4: Quantitative concept responses and matched blocking. (a) Mean normalized scaffold and ordinary-feature responses by concept, strength, and step, covering 60 conditions. (b) Distance from scaffold-blocked to concept-only images minus distance from ordinary-blocked to concept-only images, measured by LPIPS and DINOv2 across 12 concept–strength conditions. Positive values indicate larger output changes from scaffold blocking. Each condition averages over the fixed 10,000 evaluation prompts.



(a) Anime style

Figure H.5: Matched blocking examples on FLUX. Columns show the unedited image, concept-only editing, scaffold-response blocking, and matched ordinary-response blocking. In (a), the first three rows increase anime style and the last two decrease it; in continued panel (b), the first two rows add Christmas elements and the last three reduce them. The four columns compare the target edit and both blocking conditions against the same reference.



(b) Christmas elements

Figure H.5 (continued): (b) Christmas-element edits with the same column order. The first two rows add Christmas elements and the last three reduce them.

H.4 Full SCFE Transfer Results

SCFE also improves target semantics and preservation in the new model. In Table H.3, it obtains lower LPIPS and DINOv2 distance, higher CLIP-I similarity, and higher target-semantic scores than concept-only editing. Figure H.6 expands all signed strengths for both concepts: differences are smaller at low strengths, with clearer gains at medium and high strengths.

These gains do not arise from arbitrary additional updates. The random-feature control samples features with valid natural RMS outside the five-step scaffold union in `single18`, selecting two per concept and step to match SCFE's count, and updates them using their own responses. The responsive non-scaffold control selects ordinary features under the same calibration criteria. Prompt mismatch uses a fixed prompt permutation; spatial shift circularly translates responses by half the image-token grid along both axes. All four retain the native concept intervention.

In Table H.3, random-feature joint editing remains close to concept-only editing overall. Responsive non-scaffold features give some improvement, but less preservation and semantic benefit than SCFE. Prompt mismatch and spatial shift have slightly higher mean semantic scores than concept-only editing while increasing image deviation. Correct scaffold responses offer a better combination of semantics and preservation than these controls.

Table H.3: Joint editing and functional controls on FLUX.1-schnell. Each method includes 10,000 prompts, two concepts, and six signed strengths, with conditions weighted equally in the mean. Preservation is relative to paired unedited images; target semantics follow the image-pair scoring principles in Appendix G.1. Bold indicates the best result.

Method	LPIPS ↓	DINOv2 dist. ↓	CLIP-I ↑	Semantics ↑
Concept-only	0.2998	0.2089	0.8675	5.449
Random-feature joint editing	0.2997	0.2090	0.8677	5.449
Responsive non-scaffold editing	0.2934	0.2068	0.8715	5.549
Prompt mismatch	0.3023	0.2107	0.8665	5.529
Spatial shift	0.3020	0.2109	0.8666	5.480
SCFE	**0.2646**	**0.1791**	**0.8802**	**5.934**

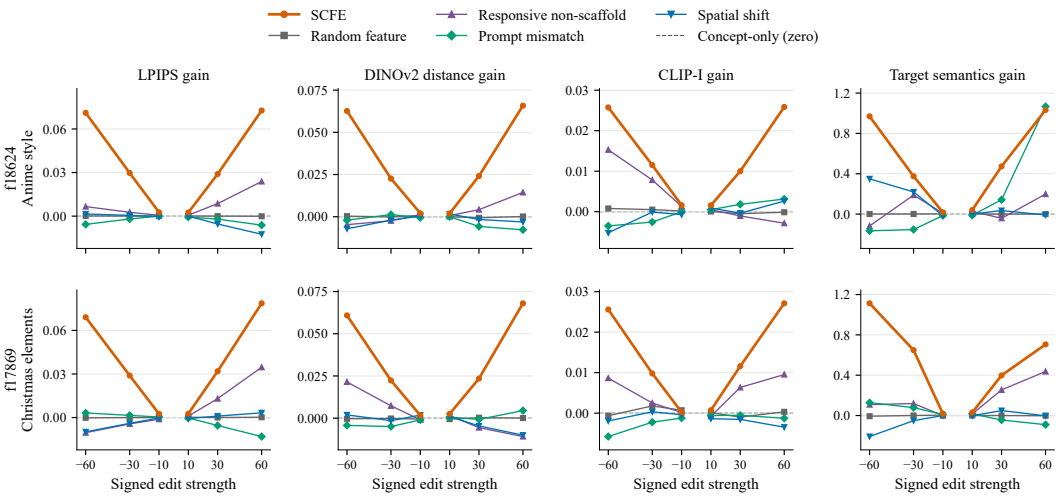


Figure H.6: FLUX editing results by concept and signed strength. Rows correspond to anime style and Christmas elements; columns report three preservation metrics and target-semantic scores. Curves give gains over concept-only editing, with positive values indicating improvement throughout. Each point averages over 10,000 evaluation prompts. Positive and negative branches are connected separately.

Figures H.7 and H.8 show edits for both concepts. In the photorealistic teapot-and-cup example, SCFE preserves the cup lost by concept-only editing; the desktop example shows fewer incidental

changes to background and other objects. In Christmas-element editing, SCFE adjusts decorations while better retaining doors and windows, furniture, and scene layout. Negative edits reduce Christmas decorations, though some remain in certain examples.

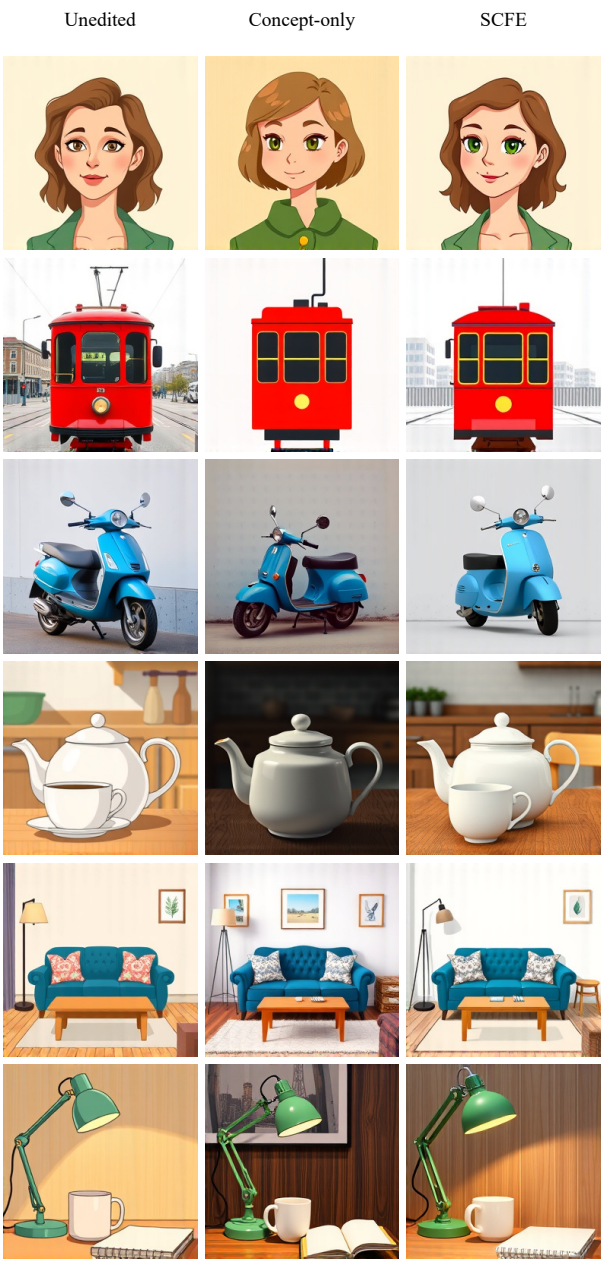


Figure H.7: Additional anime-style edits. Columns show the unedited image, concept-only editing, and SCFE. The first three rows increase anime style and the last three decrease it, illustrating style changes and preservation of subjects, parts, and scene content.



Figure H.8: Additional Christmas-element edits. Columns show the unedited image, concept-only editing, and SCFE. The first three rows add Christmas elements and the last two reduce them, showing decoration changes and preservation of doors and windows, furniture, displays, and scene layout.

I METHOD POSITIONING AND FURTHER DISCUSSION

This appendix addresses questions readers may have about the positioning of our method, the principles of comparison, and its relationship to prior work.

I.1 IS SCFE A STANDALONE EDITING METHOD OR AN ENHANCEMENT TO EXISTING METHODS?

SCFE does not address which feature encodes a target concept. It addresses how to improve an edit once the target feature has been identified. Existing SAE editing methods provide the target features and their edit magnitudes; SCFE uses associated scaffold responses to assist that edit. It is therefore best positioned as an editing enhancement built on an existing SAE feature interface, rather than another method for discovering target features.

We retain this division deliberately. SCFE does not require redefining existing concept features or replacing the SAE. Given a target concept intervention, it uses small probes to estimate scaffold responses for the current prompt, then adds constrained updates during the actual edit. Section 5 uses *One-Step Is Enough: Sparse Autoencoders for Text-to-Image Diffusion Models* (Surkov et al., 2025) to implement the Concept-only baseline through feature interventions. The comparison tests what adding SCFE changes while retaining exactly the same target concept injection.

I.2 HOW SHOULD COMPARISONS BE MADE AFTER CHANGING THE SAE OR DIFFUSION MODEL?

The relevant comparison asks whether adding SCFE helps within the same base editing system. Comparing final scores from two different SAEs, or even two different diffusion models, cannot isolate SCFE's contribution. Differences may reflect the base model's generation capabilities, the quality of the feature decomposition, or how readily the target concept can be controlled.

This does not make cross-SAE experiments uninformative. An appropriate design compares base editing with its SCFE-enhanced version separately within each configuration, then asks whether the gains persist. Changing the SAE requires identifying scaffold features and concept features again and estimating the response scales, rather than transferring feature indices and editing coefficients directly. Changing the diffusion model likewise requires establishing these correspondences in the new generative process. This is precisely what we do in Appendix H.

Thus, the paired comparison in Table 1 asks whether SCFE improves an existing edit. Experiments with other SAEs or diffusion models ask whether this improvement extends to new configurations. The objective is to evaluate the enhancement strategy's applicability, not to rank SAE architectures using SCFE's results.

I.3 HOW DOES SCFE DIFFER FROM CONCEPT-MANIFOLD EDITING?

Prior work argues that a concept need not correspond to an isolated direction, but can occupy a low-dimensional region with internal structure. Block-Sparse Featurizers (BSFs) learn multidimensional feature blocks and control generation in SDXL by moving along their internal geometry (Fel et al., 2026, Section 6). The concern that arbitrarily changing a direction can damage an image is therefore not new to this work.

However, this is not equivalent to concluding that failing to jointly update neighboring co-active features necessarily takes an edit off the data manifold. BSFs study how concepts are represented by multidimensional blocks and how to move within those blocks. Our scaffold set is not a multi-dimensional expansion of a particular target concept. Cross-prompt reuse determines the candidate set first; we then examine which candidates respond to the current concept intervention.

SCFE can reuse an existing SAE without first learning concept blocks or fitting a complete manifold. It uses the downstream responses to small concept interventions to determine the direction and spatial locations of additional updates, then bounds their magnitudes using natural activation scales. One possible explanation, however, is that coordinated updates to target-concept and scaffold features bring the edited internal representations closer to the representation manifold of natural

generation. This offers a geometric perspective on SCFE's editing gains, but it lies beyond the scope of this work and requires further research to confirm.

I.4 What is new given prior work on highly reused features in LLMs?

Prior LLM research shows that frequently active latents can support functions such as context binding rather than merely reflect training noise (Sun et al., 2025). FEGA selects pointer-like candidates based on high reuse across examples and prompt families among correctly answered examples of predefined tasks such as copying and rule application, then studies their task contributions and the geometry of their output effects (Hoang et al., 2026).

FEGA's pointer-like candidate selection centers on predefined tasks such as copying and rule application: it considers only correctly answered examples and requires high reuse across examples and prompt families at the final token before prediction (Hoang et al., 2026, Section 4.2). We instead start from image-generation prompts with open-ended content, without relying on target semantics or correctness on a predefined task. Our screening measures cross-prompt occurrence rates by network location and denoising step, without requiring features to cover the entire image or occupy fixed spatial positions. As Figures 2 and 3 show, the resulting scaffold population is shared across generated content, but its activation locations and intervention effects vary with content and are difficult to characterize with a single concept.

FEGA tests task contributions through population zero-ablation and matched ablation controls, and analyzes the geometry of output-logit changes induced by single-feature ablation (Hoang et al., 2026, Sections 4.5 and 5). We instead retain the upstream concept edit, restore selected downstream coordinates to their unedited reference values, and compare this intervention with blocking matched ordinary responses to test how concept-induced scaffold responses participate in editing. SCFE then uses the current prompt's spatial responses for constrained joint updates, turning these shared features from objects of functional analysis into control variables that improve target semantics and image preservation.

I.5 Do small gains at low strengths make SCFE unnecessary?

At low editing strengths, SCFE and the Concept-only baseline implemented with feature interventions from *One-Step Is Enough: Sparse Autoencoders for Text-to-Image Diffusion Models* (Surkov et al., 2025) have similar mean semantic scores and CLIP-I similarities, but SCFE already improves LPIPS and DINOv2 distance. As Figure 5 shows, SCFE's mean semantic and preservation gains both increase with intervention strength. Similarity on some metrics at low strengths therefore does not imply that joint updates have no effect.

A more direct question is whether low strength can always meet the editing need. For some images, it does. In Figure I.1(a), the camera already changes from an outlined illustration to a photographic appearance at low strength, and the chef's expression is visibly more relaxed. The explorer's cartoon-like appearance and the handbag's color change can likewise be achieved at low strength. For these examples, a further increase is unnecessary to establish a clear target change. This does not mean that other content is unchanged; for example, the handbag's background also becomes blue.

The same low strength does not meet editing needs across all content. In Figure I.1(b), the lighthouse remains photographic at low strength, while the espresso machine retains pronounced illustrative outlines. Higher strength produces a clearer cartoon style and photographic appearance, respectively. The bespectacled portrait already changes facial proportions at low strength, but clearer drawn outlines and shading emerge only at higher strength. Expression editing shows a similar difference. The person in the leather jacket still looks tense at low strength; increasing it produces more relaxed brows and a slight smile. The strength needed for a given feature to produce the target change depends on the generated content.

These examples show that low strength can suffice, but cannot meet every editing need. SCFE is not intended to apply stronger interventions to every image. It aims to balance target-semantic change and overall image preservation when stronger concept editing is needed.



(a) Low strength already produces a clear target change

Figure I.1: Editing-strength requirements across generated content. Columns show unedited, low-strength, and higher-strength images from Concept-only feature interventions as described in *One-Step Is Enough: Sparse Autoencoders for Text-to-Image Diffusion Models* (Surkov et al., 2025). Each row fixes the prompt, initial noise, and generation configuration; only the feature coefficient changes. Absolute strengths are 5 and 10; signed coefficients, features, prompt IDs, and seeds appear above each row. **(a)** Low strength already produces a clear target change; further amplification is unnecessary and need not improve the result. **(b, continued)** Higher strength produces a clearer style or expression change.



(b) The target change remains insufficient at low strength

Figure I.1 (continued): Editing-strength requirements across generated content. Column order, base method, and controls match the preceding page. Rows show cartoon-like editing of a bespectacled portrait and lighthouse, photographic rendering of an espresso machine, and a more relaxed expression in a person wearing a leather jacket. Absolute strengths remain 5 and 10; signs, features, prompt IDs, and seeds appear above each row.

I.6 Are these metrics sufficient to evaluate editing quality?

Evaluating concept editing is not equivalent to checking whether a class label is correct. The same target edit can have multiple reasonable outcomes, and target change, preservation of existing content, and visual quality do not always agree. We therefore do not treat any single metric as a definitive criterion for editing success. Prior work has also identified this issue. ImagenHub finds substantial gaps between automatic metrics and human judgments across several conditional image-generation and editing tasks (Ku et al., 2024).

The image metrics used here serve different purposes. LPIPS measures perceptual differences, while DINOv2 distance and CLIP-I measure differences and similarity in their respective image representations. They describe closeness to the original content, rather than directly determining whether the target edit is complete. This use of perceptual distances and representation similarity has precedents. LPIPS was developed for perceptual-similarity evaluation (Zhang et al., 2018), and Emu Edit uses DINO and CLIP image similarity to assess preservation between input and edited images (Sheynin et al., 2024).

Being closer to the original does not itself mean a better edit. An almost unchanged image can score well on preservation without fulfilling the editing objective. Conversely, changing a photograph into a cartoon necessarily changes texture, color, and shading, and these intended changes can also increase image distances.

To address this, target-semantic evaluation uses `GPT-5.6-Sol`. The evaluator receives the unedited image, edited image, and specified direction to judge whether and how much the target change occurs, rather than merely judging visual appeal. Table 1 and Figure 5 in Section 5 present semantic scores alongside preservation to distinguish better target editing from simply weaker editing. Multimodal-model-assisted evaluation has precedents. DreamBench++ uses GPT-4o to evaluate personalized generation and studies the role of evaluation prompts (Peng et al., 2025).

Image comparisons in the main text and appendices serve an additional, indispensable purpose: they let readers inspect the changes behind the scores. Figure 1(b) shows edits under matched conditions, while Appendices E and G.4 and Figure I.1 show different content and interventions. Whether a target style actually changes, the subject remains intact, or local structure is damaged requires examining the images. Quantitative results summarize aggregate trends; visualizations reveal specific changes. We interpret editing outcomes jointly through these sources of evidence, rather than through an isolated rise or fall in a metric.

I.7 Why does quantitative evaluation focus on a few concepts?

Our quantitative evaluation uses a fixed set of 10,000 open-content prompts covering people, objects, natural landscapes, and other content, as already discussed in Appendix B. A uniform comparison on this set requires considering whether the target attribute applies across generated content. Style and overall appearance directions facilitate comparisons on a fixed prompt set, whereas attributes such as dials and collars require the corresponding objects to be present. The same object-dependent editing task cannot simply be applied to every prompt.

However, evaluating SCFE on only a few concepts does not provide an accurate assessment. We therefore combine large-scale quantitative evaluation with object-dependent qualitative comparisons. The former compares overall performance under fixed conditions; the latter selects prompts containing the target objects.