CONTEXT-AWARE SALIENCY CONTROL FOR DIFFUSION-BASED OBJECT INSERTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern diffusion-based inpainting models and image-conditioned adapters provide reference-guided object insertion, but offer limited explicit control over the prominence of generated content. We introduce CASAdapter (Context-Aware Saliency Adapter), a module with 9.7M trainable parameters that modulates an object's CLIP image embedding according to the surrounding scene. Its orthogonal tangent-space update help to preserves the embedding norm after renormalization and bounds its angular displacement. This geometric constraint motivates limiting conditioning drift, without guaranteeing preservation of the generated object's identity. CASAdapter is trained through truncated DDIM sampling using a UniSal-based objective together with semantic and color-gradient regularization. Evaluation on a 300-image MS-COCO validation subset shows positive mean saliency gains under UniSal, DeepGaze IIE, SUM, and both TranSalNet variants at positive modulation settings, and negative mean gains at the tested negative setting. These observations support transfer beyond the training predictor. Larger orthogonal modulation produces higher mean saliency gains alongside lower CLIP similarity. Compared with the reported affine setting, the orthogonal variant offers a more favorable observed trade-off for four of the five predictors at a moderate setting; the strongest tested setting exceeds the affine setting in mean gain under all five predictors while retaining higher CLIP similarity.

1 INTRODUCTION

Visual attention is central to how images communicate meaning. Designers, photographers, and advertisers invest considerable effort in directing the viewer's gaze toward the elements that carry their core message. Consequently, a large body of work on computational saliency Droste et al. (2020); Hosseini et al. (2024); Lou et al. (2022); Xiong et al. (2024) has long sought to predict where humans look. However, the inverse problem, modifying an existing image so that a chosen object captures a specified visual attention has received comparatively little attention. To this day, saliency manipulation remains a largely manual post-production task, reliant on an editor's intuition regarding contrast, framing, and lighting. While a few recent generative approaches incorporate saliency as a signal Zhang et al. (2024); Jiang et al. (2021), they primarily use it as a layout prior during the initial image synthesis, rather than as a targeted editing objective for an already composed scene.

In this paper, we address this inverse problem at the object level as shown in Figure 1. Given an input image, a binary mask delimiting an object of interest, and a diffusion-based inpainting pipeline, we aim to regenerate the masked region such that the object naturally captures a larger (or smaller) visual attention, while encouraging reference consistency and limiting visible artifacts.

Modern inpainting models Rombach et al. (2022), combined with image-conditioned adapters such as an Image Prompt adapter (Ip-Adapter) Ye et al. (2023), enable photorealistic object insertion from a single reference image. Furthermore, spatial conditioning modules like ControlNet Zhang et al. (2023) and Text-to-image models (T2I-Adapter) Mou et al. (2024) enrich this generation space for extra guidance. Yet, while the text branch of these models can be prompted with saliency-evocative descriptors (*e.g.* *vibrant*, *eye-catching*, *high contrast*), the resulting effect is highly stochastic and insufficient to guarantee a measurable, controlled shift in attentional capture.

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.

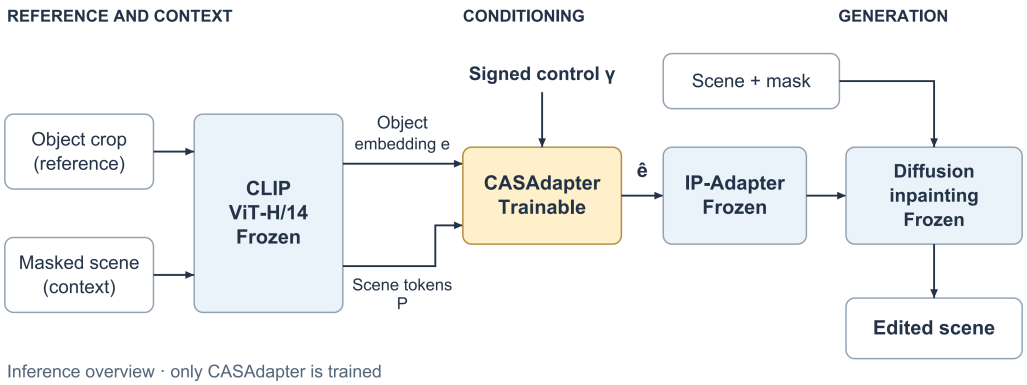

Figure 1: Overview of the CASAdapter pipeline integrated with an IP-Adapter-based diffusion inpainting framework. The proposed module modulates the CLIP image embedding conditioned on scene context to control object-level saliency during generation.

Beyond controllability itself, a practical challenge lies in the computational cost of adapting modern diffusion pipelines. Fully fine-tuning large generative backbones is resource-intensive and can destabilize pretrained generative priors. We therefore seek a lightweight adaptation strategy capable of operating on top of frozen diffusion architectures, while leaving the underlying generator fixed. CASAdapter is specifically designed around this principle: only a compact modulation module is trained, while the diffusion backbone, VAE, CLIP encoder, and IP-Adapter remain entirely frozen during optimization.

To address this limitation, we propose CASAdapter (Fig.1), a lightweight residual module inserted upstream of IP-Adapter. CASAdapter modulates the CLIP image-conditioning embedding according to the surrounding scene, with the aim of adjusting object-level saliency while encouraging semantic consistency with the reference.

CASAdapter combines object and scene features to predict a gated tangent-space update. Normalization restores the reference embedding norm, while the signed coefficient bounds angular displacement. These geometric properties constrain conditioning changes; image-level fidelity remains an empirical question.

The saliency objective encourages an increase in UniSal predictions within the target mask. It is jointly optimized with a CLIP cosine similarity loss and an HSV color-gradient penalty to discourage semantic and chromatic deviations. Saliency share and foreground–background contrast are reported as evaluation metrics.

In summary, our main contributions are three-fold:

- We formalise the problem of saliency-aware object insertion and highlight the limitations of current diffusion pipelines, which optimise for photorealism without explicitly modelling the attentional impact of the generated content.

- We introduce CASAdapter, a context-aware module with 9.7M trainable parameters operating upstream of a frozen IP-Adapter-based diffusion pipeline. Its orthogonal tangent-space update preserves embedding norm after rescaling and provides an explicit bound on angular displacement. This motivates a constrained conditioning change rather than a guarantee of semantic preservation.

- We evaluate saliency control on MS-COCO with UniSal and four additional predictor configurations. The observed mean changes transfer across these evaluators, with a measurable trade-off between saliency gains and CLIP similarity. A comparison with an affine modulation baseline examines this trade-off across conditioning variants.

2 Related Work

Visual saliency prediction. Computational saliency models estimate where observers are likely to look in an image. The DeepGaze family builds on features learned for visual recognition: DeepGaze I (Kümmerer et al., 2015) uses pretrained image features for fixation prediction, while DeepGaze IIE (**?**) combines predictions from different backbones and studies probabilistic calibration in and out of domain. Our evaluation uses DeepGaze IIE.

UniSal (Droste et al., 2020) jointly models image and video saliency through a shared encoder, recurrent, decoder architecture with domain-adaptive components. It supplies the differentiable training signal in our framework. TranSalNet (Lou et al., 2022) combines convolutional features with transformer modules to capture long-range context. Its two variants, TranSalNet_Dense and TranSalNet_Res, use DenseNet-161 and ResNet-50 backbones, respectively. Both belong to the same method and share the same publication; we evaluate them separately to assess sensitivity to the feature backbone. SUM (Hosseini et al., 2024) integrates Mamba-based context modeling with a U-Net architecture to unify saliency prediction across different image types.

These models provide different architectures for examining whether our edits improve saliency predictions beyond UniSal. Our analysis uses UniSal as the training-proxy evaluator and SUM, both TranSalNet variants, and DeepGaze IIE as additional evaluators. These additional predictors do not supply the training loss. Architectural diversity motivates this comparison, but does not imply independent errors or disjoint training data. Agreement would support transfer across the tested predictors, rather than establish human-attention control. CASAdapter addresses the complementary editing problem: modifying a selected object's predicted prominence through the conditioning space of a frozen generator.

Diffusion models and conditioning mechanisms. Latent diffusion models Rombach et al. (2022) have become the standard for high-quality image synthesis and inpainting. Beyond text conditioning, IP-Adapter Ye et al. (2023) enables reference-guided object insertion by injecting CLIP image embeddings through dedicated cross-attention layers. Furthermore, ControlNet Zhang et al. (2023) and T2I-Adapter Mou et al. (2024) introduce spatial conditioning (e.g., depth, edges) via trainable adapter networks attached to a frozen U-Net. A parallel direction explores parameter-efficient control over generative behaviour. Low-Rank Adaptation (LoRA) Hu et al. (2022) applies efficient weight updates, while Concept Sliders Gandikota et al. (2023) learn one-dimensional directions in the weight space corresponding to interpretable visual attributes (e.g., age, lighting). While these methods manipulate model weights to steer global image properties, CASAdapter operates directly within the CLIP conditioning space. This enables fine-grained, object-specific control relative to its surrounding context, without altering the global foundational weights.

Saliency in diffusion pipelines. The literature explicitly connecting visual saliency with generative pipelines remains limited. Among the earliest works, Jiang et al. (2021) introduced the task of saliency-guided image translation, where a GAN-based model, named SalG-GAN, performs image-to-image translation conditioned on a target saliency map. Their approach incorporates a saliency-guided attention mechanism to generate images consistent with the desired attention distribution.

More recently, GazeFusion Zhang et al. (2024) extended this idea to diffusion models by proposing a saliency-guided image generation framework. Their method conditions the diffusion process on a user-specified attention distribution to control the regions that are likely to attract human gaze. The authors validated their approach through eye-tracking user studies and model-based saliency analyses, demonstrating that the generated gaze distributions align with the intended attention targets.

These works use saliency as a guidance signal in generative pipelines. CASAdapter instead studies object-level prominence control through the conditioning space of a frozen reference-guided diffusion model, with semantic consistency encouraged by geometric constraints and regularization.

3 Method

CASAdapter learns a context-dependent transformation of the image embedding supplied to a frozen reference-guided diffusion model. The transformation provides a signed control over the conditioning displacement, while saliency and fidelity objectives determine the learned direction. We distin-

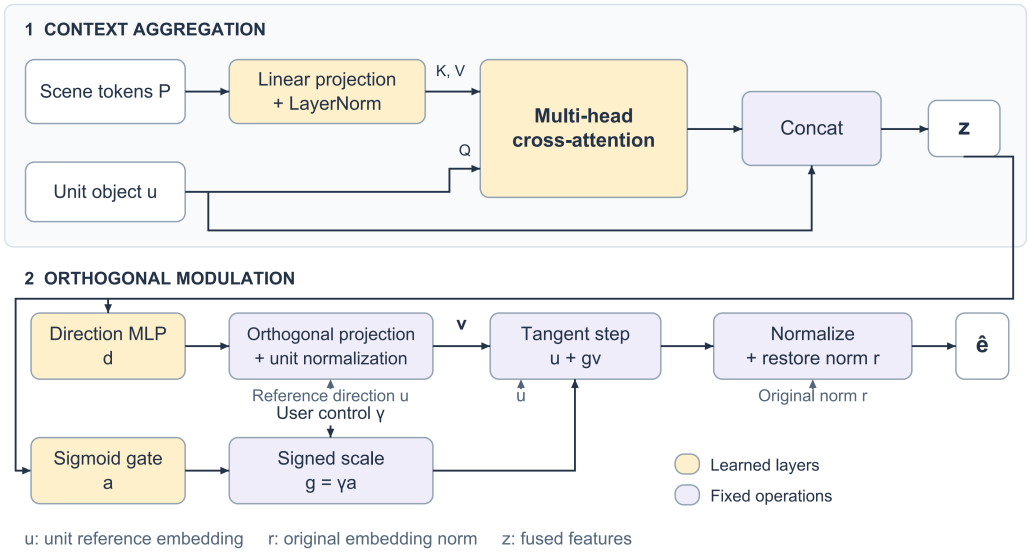

Figure 2: CASAdapter aggregates scene tokens using the object embedding as a query. The fused features predict an update direction and a scalar gate. Orthogonal projection, normalization, and rescaling constrain the embedding supplied to the frozen IP-Adapter.

guish the geometry of the conditioning update from its empirical effects on saliency and reference consistency.

3.1 Problem formulation

The inputs are a scene image, a binary mask defining the target region, and an object reference image. In our experiments, the reference is the bounding-box crop of the target object in the original scene. A frozen CLIP image encoder extracts a 1024-dimensional object embedding and 1280-dimensional scene patch features. Context is taken from the scene with the target region masked out, following the experimental protocol.

CASAdapter transforms the reference embedding before it is passed to a frozen Stable Diffusion inpainting model (Rombach et al., 2022) equipped with IP-Adapter (Ye et al., 2023). Only CASAdapter is trained. At inference, a signed coefficient γ adjusts the conditioning displacement to influence the target's predicted prominence while encouraging consistency with the reference. This coefficient controls modulation strength rather than specifying a desired saliency percentage. Setting it to zero recovers the original conditioning embedding, not the original scene: the frozen pipeline still performs inpainting.

3.2 Context-aware conditioning

Object prominence depends on its surroundings. CASAdapter therefore combines the reference embedding with scene context before predicting its update (Figure 2). The object embedding is normalized, while scene patch features are linearly projected to the same dimension and layer-normalized. Multi-head cross-attention uses the object embedding as a single query and the projected scene features as keys and values.

The resulting context feature is concatenated with the normalized object embedding. A small multilayer perceptron predicts an update direction from these fused features, and a sigmoid gate predicts its relative strength. Multiplying this gate by the user coefficient makes the update both context-dependent and externally adjustable. CASAdapter contains approximately 9.7M trainable parameters; the CLIP encoder, IP-Adapter, diffusion backbone, and VAE remain frozen.

3.3 ORTHOGONAL MODULATION

Tangent-space update. Let $\mathbf{e}$ denote the reference embedding, $r = \|\mathbf{e}\|_2$ its norm, and $\mathbf{u} = \mathbf{e}/r$ its unit direction. The network predicts a direction $\mathbf{d}$ and a sigmoid gate a between zero and one. The signed step is $g = \gamma a$, so $|g| \leq |\gamma|$. We remove the predicted direction's component parallel to $\mathbf{u}$:

$$\mathbf{d}_\perp = \mathbf{d} - \langle \mathbf{d}, \mathbf{u} \rangle \mathbf{u}, \qquad \mathbf{v} = \frac{\mathbf{d}_\perp}{\|\mathbf{d}_\perp\|_2}. \tag{1}$$

For a nonzero projection, $\mathbf{u}^\top \mathbf{v} = 0$ and $\|\mathbf{v}\|_2 = 1$. The modulated embedding is obtained by a tangent step followed by normalization and rescaling:

$$\hat{\mathbf{e}}_\gamma = r \frac{\mathbf{u} + g\mathbf{v}}{\|\mathbf{u} + g\mathbf{v}\|_2} = r \frac{\mathbf{u} + g\mathbf{v}}{\sqrt{1 + g^2}}. \tag{2}$$

The expression assumes a nonzero reference embedding and projected direction. A zero projected direction can be assigned an identity update; near-zero norms require numerical stabilization.

Geometric interpretation. Orthogonality gives $\|\mathbf{u} + g\mathbf{v}\|_2^2 = 1 + g^2$. Consequently, Equation 2 preserves the reference norm, $\|\hat{\mathbf{e}}_\gamma\|_2 = r$. If α denotes the unsigned angle between the original and modulated embeddings, then

$$\cos \alpha = \frac{1}{\sqrt{1 + g^2}}, \qquad \alpha = \arctan |g| \leq \arctan |\gamma|. \tag{3}$$

Thus, $|\gamma|$ sets an angular budget around the reference direction. At $\gamma = 0$, the original embedding is recovered exactly. Changing the sign of γ reverses the tangent step, although the generated saliency response need not be symmetric or monotonic.

Preserving magnitude while limiting directional displacement provides a geometric motivation for restricting conditioning drift. This property concerns the input to IP-Adapter: it does not bound the semantic difference between generated images, nor guarantee preservation of object identity after diffusion sampling. We therefore use orthogonal modulation as a conditioning regularizer, complemented by image-level fidelity objectives. Its effect on generated content is assessed empirically through the saliency–fidelity trade-off.

3.4 TRAINING OBJECTIVE

Differentiable generation. Training uses an unmodulated branch conditioned on $\mathbf{e}$ and a modulated branch conditioned on $\hat{\mathbf{e}}_\gamma$. Starting from a noisy latent at a timestep sampled in $[500, 700]$, each branch runs 30 DDIM denoising steps (Song et al., 2020). The unmodulated branch is evaluated without gradients. In the modulated branch, gradients are retained through the final two steps and the VAE decoder. Freezing these networks means that their weights are not updated; the retained computations still propagate gradients to CASAdapter.

The original scene is the reference for all three losses: saliency and color use the original image, while semantic consistency uses its target-object crop. The image generated without modulation provides a separate comparison and is not the target of the training losses.

Saliency and fidelity. The total objective is

$$\mathcal{L} = \lambda_{\text{sal}} \mathcal{L}_{\text{sal}} + \lambda_{\text{sem}} \mathcal{L}_{\text{sem}} + \lambda_{\text{col}} \mathcal{L}_{\text{col}}, \tag{4}$$

with weights of 0.6 for saliency, 0.4 for semantic consistency, and 0.4 for color regularization.

Saliency. Frozen UniSal (Droste et al., 2020) provides the differentiable saliency signal. The loss is the ratio between the mean predicted saliency inside the target mask in the original image and that in the modulated output. Minimizing it encourages a larger UniSal response in the target region. This objective optimizes predicted saliency rather than human gaze; negative control is evaluated at inference.

Semantic consistency. We penalize cosine distance between frozen CLIP embeddings of the original and generated object regions. This encourages semantic alignment without requiring pixelwise reconstruction or guaranteeing exact visual identity.

Color consistency. We compare spatial gradients in HSV space between the original image and the modulated output. Hue-gradient differences are weighted by local saturation, while saturation and value gradient differences receive absolute-difference penalties. This discourages abrupt chromatic changes while allowing some lighting variation.

Together, these objectives and the angular budget aim to balance predicted saliency gain with reference consistency.

Optimization. We train CASAdapter with AdamW, an initial learning rate of 10^{-4}, and cosine decay to 10^{-6}. A batch size of 2 with eight gradient accumulation steps gives an effective batch size of 16; the gradient norm is clipped at 1.0. UniSal and the semantic encoder remain frozen throughout optimization. SUM, TranSalNet, and DeepGaze IIE are used to evaluate the generated outputs and do not contribute gradients to this training objective.

4 Experimental setup

We evaluate whether the learned modulation changes object-level saliency beyond the UniSal predictor used during training, and quantify the accompanying change in reference similarity. All saliency gains and CLIP similarities are measured against the original image, including those of the unmodulated baselines.

4.1 Data and generation protocol

CASAdapter is trained on MS-COCO `train2017` and evaluated on a subset of 300 `val2017` images (Lin et al., 2014) that are not used for its training. Instance segmentation masks define the target regions, and the original object crops provide reference conditioning. The target-masked scene supplies contextual features. A local image crop around the target, padded by 10 pixels, is resized with its mask to 512×512 for inpainting. The generated patch is then resized back and reinserted into the scene. Saliency predictors evaluate the resulting full-scene images.

Every benchmark uses the same 300 source images and 1,200 generations, with four outputs per image. The same initial seed is used across configurations. Metrics are computed per original/generated pair, averaged over the four outputs, then averaged equally across images. Reported dispersions are population standard deviations of image-level means, not confidence intervals.

4.2 Baselines and modulation settings

The first baseline uses the frozen IP-Adapter pipeline with $\gamma = 0$, which leaves the reference embedding unchanged. The second also uses $\gamma = 0$, but adds a text prompt explicitly requesting increased visibility and contrast of the target object. This comparison tests whether a saliency-oriented prompt alone produces changes comparable to embedding modulation.

CASAdapter is evaluated at $\gamma \in \{-0.3, 0.3, 0.5, 0.8\}$. The negative setting tests suppression; the positive settings test increasing modulation strengths. Apart from the custom-prompt baseline, the comparison varies the embedding modulation under the shared generation procedure. We report an affine CASAdapter variant at modulation strength 0.15 using the same benchmark images and generation parameters. Its coefficients need not produce the same saliency response as orthogonal coefficients; we therefore compare achieved gains and CLIP similarity rather than equating the coefficient values. The affine and orthogonal variants were trained with exactly the same training parameters. This comparison evaluates their modulation formulations under matched training settings.

The unmodulated generated image is a baseline output, whereas the original scene is the reference for every reported gain. Thus, zero modulation need not yield zero saliency change: the diffusion pipeline still regenerates the target region.

4.3 Predictors and evaluation metrics

We evaluate the stored outputs using UniSal (Droste et al., 2020), SUM (Hosseini et al., 2024), TranSalNet_Dense and TranSalNet_Res (Lou et al., 2022), and DeepGaze IIE. UniSal supplies the

Table 1: Mean relative saliency-share gain ΔA (%) and CLIP similarity. Dense and Res denote TranSalNet variants. UniSal supplies the training objective. Every configuration uses 300 source images and 1,200 generations; a dash denotes an unavailable score. Higher CLIP indicates greater reference similarity, not guaranteed identity preservation.

Configuration	UniSal	DeepGaze IIE	SUM	Dense	Res	CLIP
Unmodulated	-1.10	1.45	—	1.90	0.46	0.838
Unmodulated + prompt	-0.70	2.20	0.57	1.77	1.01	0.836
Affine, 0.15	20.15	26.80	51.21	16.21	14.70	0.641
Orthogonal, -0.3	-7.59	-4.09	-8.26	-1.41	-3.06	0.822
Orthogonal, 0.3	11.52	14.43	18.72	9.43	8.13	0.821
Orthogonal, 0.5	21.93	28.30	39.28	17.32	16.30	0.768
Orthogonal, 0.8	36.07	45.59	73.97	31.93	28.65	0.677

training objective; the other predictors assess transfer beyond that objective. These results concern transfer across the evaluated predictors, not direct measurements of human gaze.

Saliency share (A) is the percentage of total predicted scene saliency inside the target mask. We report it for the original and generated images as A_{ref} and A_{gen}. **Saliency share gain** (ΔA) is their relative change, computed for each pair before averaging, with a denominator stabilization of 10^{-8} applied to the underlying zero-to-one masses. It is expressed in percent, not percentage points. Consequently, the mean relative gain cannot generally be recovered from the two mean shares.

Saliency contrast gain (ΔC) is the relative change in the ratio of mean saliency inside and outside the mask, with the same denominator stabilization. Share and contrast summarize related aspects of predicted object prominence; they are not independent validations. **CLIP cosine similarity** compares normalized embeddings of the original and generated object crops. It is computed once per output and does not depend on the saliency evaluator. Since CLIP also contributes to training, this metric is a measure of reference consistency rather than an independent test of identity preservation.

5 Results and discussion

Table 1 summarizes saliency-share gains and reference similarity. All gains are relative to the original image. The comparisons are descriptive: we do not report paired confidence intervals or significance tests. Figure 3 presents earlier qualitative examples; their original settings differ from the new benchmark.

5.1 Baselines and transfer across predictors

Without modulation, mean share gains range from -1.10% to $+1.90\%$ among available predictors. The saliency-oriented prompt yields similarly small changes, from -0.70% to $+2.20\%$, with CLIP similarity of 0.836 compared with 0.838 without the custom prompt. The prompt does not improve every evaluator's mean: TranSalNet_Dense decreases slightly. These findings concern this particular prompt and do not rule out other text-based strategies.

For each positive orthogonal setting, all five predictors report positive mean share and contrast gains. At $\gamma = 0.3$, share gains range from 8.13% to 18.72%, including 11.52% under UniSal. At $\gamma = 0.8$, they range from 28.65% to 73.97%. DeepGaze IIE, SUM, and both TranSalNet variants thus register changes beyond the training predictor. The affine variant also produces positive mean gains under all five models, so cross-predictor transfer is not unique to the orthogonal update. Agreement supports transfer across the tested predictors, without excluding shared biases or establishing an effect on human gaze.

5.2 Signed control and reference consistency

Mean share and contrast gains increase across the three positive orthogonal strengths. At $\gamma = -0.3$, all five predictors instead show negative mean gains: share changes range from -1.41% to -8.26%.

These observations support a signed average response at the evaluated settings, not symmetric effects or monotonic behavior for each image. Figure 3 illustrates the modulation mechanism on chair, cow, car, and bag examples, pairing generated images with saliency overlays. Their original coefficients (-0.2, 0.2, and 0.5) differ from the current benchmark settings, so they provide qualitative context rather than direct visual validation of the reported averages.

The positive response comes with lower CLIP similarity: 0.821 at $\gamma = 0.3$, 0.768 at 0.5, and 0.677 at 0.8. Moderate modulation therefore illustrates a less aggressive operating point, rather than an established optimum. This trade-off is consistent with the scope of the geometry: restricting conditioning displacement does not ensure preservation of generated appearance. The generated-image rows in Figure 3 allow appearance changes to be inspected alongside the corresponding saliency overlays; the latter visualize model predictions, not measured human gaze.

5.3 Comparison with affine modulation

The affine baseline at 0.15 reaches a mean CLIP similarity of 0.641. Orthogonal modulation at 0.5 achieves slightly higher mean share gains under four predictors: UniSal (21.93% versus 20.15%), DeepGaze IIE (28.30% versus 26.80%), and both TranSalNet variants. Its CLIP similarity is higher by 0.127, reaching 0.768. SUM is an exception: the orthogonal gain is 39.28%, below the affine gain of 51.21%. The achieved saliency levels are therefore close for four predictors, but not uniformly matched across evaluators.

At orthogonal strength 0.8, mean share and contrast gains exceed the reported affine baseline under every predictor, while CLIP similarity remains higher (0.677). These operating points support a more favorable observed saliency–similarity trade-off for the orthogonal formulation under identical training parameters. This comparison does not establish superiority across the full affine response curve or isolate the projection from every other difference between the modulation formulations.

5.4 Variability and limitations

Relative gains vary substantially across images. SUM gives $73.97 \pm 192.79\%$ for orthogonal modulation at 0.8 and $51.21 \pm 121.31\%$ for affine modulation at 0.15. Small original saliency masses can amplify relative changes. For example, the former setting increases SUM's mean absolute share from 24.871% to 29.725%, or 4.854 percentage points. Mean pairwise relative gains and differences of mean shares therefore answer different questions. Standard deviations describe dispersion, not uncertainty on the mean; medians, success proportions, and paired confidence intervals cannot be recovered from these aggregates.

Only one affine strength is reported. Although training parameters are identical, this comparison concerns the complete modulation formulations rather than the projection alone. CLIP similarity and cross-predictor agreement remain proxies for fidelity and attention, not guarantees of identity preservation or human gaze control.

6 Conclusion

The results support the general concept of controlling predicted object prominence through a compact, context-aware conditioning module while keeping the diffusion backbone frozen. CASAdapter produces signed mean saliency changes that transfer beyond UniSal, with a measurable trade-off between saliency gain and CLIP similarity. Compared with the reported affine baseline, the orthogonal variant offers a more favorable observed trade-off at several operating points. Its geometric constraints restrict embedding displacement, without guaranteeing image-level identity preservation.

The next step is to apply this methodology to more recent diffusion backbones and assess whether stronger generative models improve the balance between saliency control and reference consistency. Extending it to video would additionally require temporally coherent modulation, consistent object appearance, and evaluation of attention across frames. These directions would test how broadly the concept transfers beyond the current image-inpainting setting.

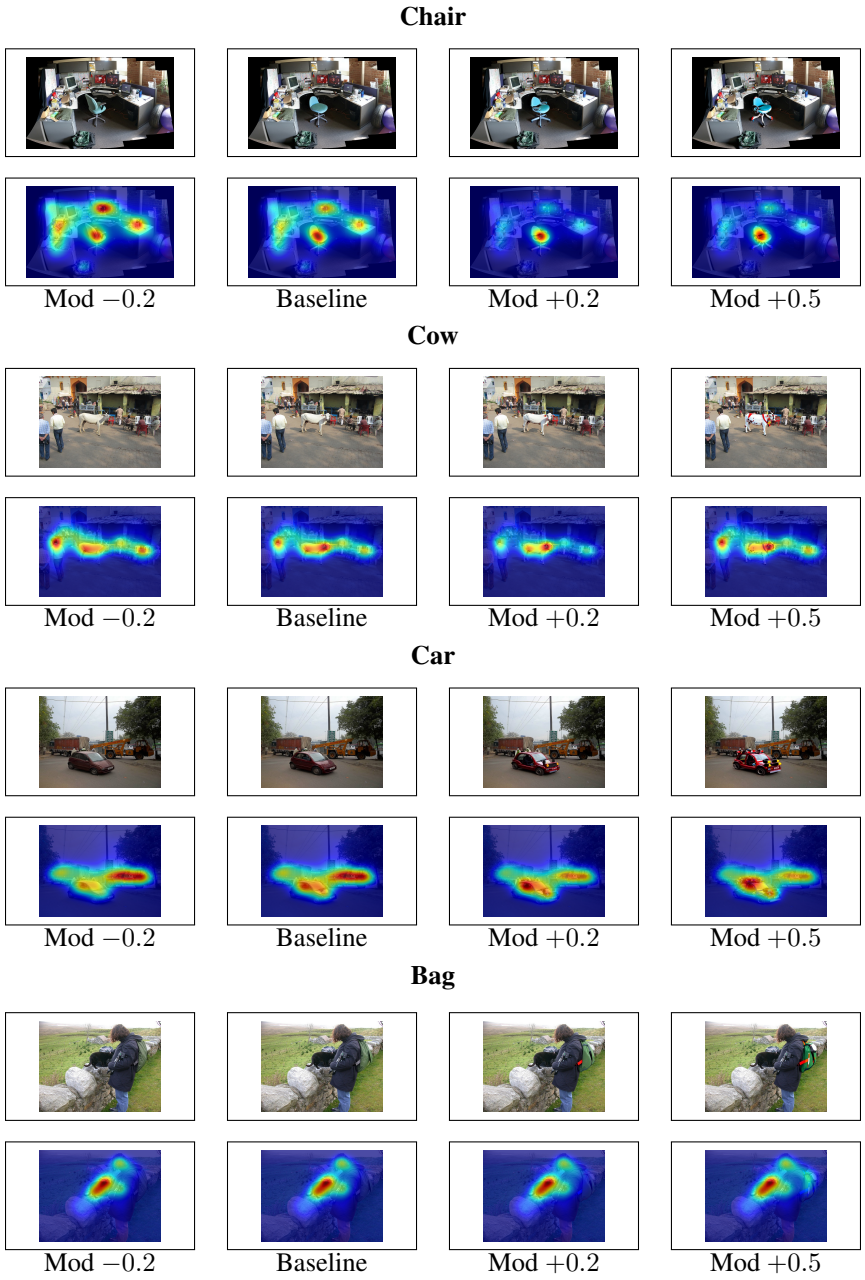

Figure 3: Qualitative orthogonal-modulation examples from the earlier experiment: chair, cow, car, and bag scenes. For each scene, generated images appear above their saliency overlays at the original coefficients. These illustrative settings differ from the current benchmark and do not depict the affine comparison.

AI USE STATEMENT

Artificial intelligence was used as an assistive tool at several stages of this work. It contributed to the drafting and reformulation of selected paragraphs, as well as to adapting the manuscript to the ICLR template. During the research phase, AI tools were used in addition to standard bibliographic methods to identify prior work, explore the state of the art, and map existing approaches and solutions. AI also assisted in writing some of the code, particularly scripts used to build the evaluation benchmark for testing multiple saliency-prediction models, to automate the generation of result images, and to develop the Web-based demonstration application. All experimental results, analyses, and AI-assisted content were reviewed and validated by the authors.

REFERENCES

Richard Droste, Jianbo Jiao, and J. Alison Noble. Unified image and video saliency modeling. In *ECCV*, 2020.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Concept sliders: LoRA adaptors for precise control in diffusion models. *arXiv preprint arXiv:2311.12092*, 2023.

Alireza Hosseini et al. SUM: Saliency unification through mamba for visual attention modeling. *arXiv preprint arXiv:2406.17815*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *ICLR*, 2022.

Lai Jiang, Mai Xu, Xiaofei Wang, and Leonid Sigal. Saliency-guided image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16509–16518, 2021.

Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. In *ICLR Workshop*, 2015.

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

Jianxun Lou, Hanhe Zhang, and Rabab Hamzah. Transalnet: Towards perceptually relevant visual saliency prediction. In *Neurocomputing*, volume 494, pp. 455–467, 2022.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *AAAI*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Junwen Xiong, Peng Wang, et al. DiffSal: Joint audio and video learning for diffusion saliency prediction. In *CVPR*, 2024.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

Yunxiang Zhang, Nan Wu, Connor Z Lin, Gordon Wetzstein, and Qi Sun. Gazefusion: Saliency-guided image generation. *ACM Transactions on Applied Perception*, 21(4):1–19, 2024.