AMPLIFY WHAT YOU GAZE AT: TARGET SALIENCY BOOSTING IN TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image generation has advanced in controlling what, where, and how objects appear, yet how visual attention is distributed among objects remains largely unexplored. In this paper, we introduce *Target Saliency Boosting*, a new task aimed at boosting the visual saliency of a specific object during text-to-image generation without requiring any visual priors. Our key insight is that visual saliency is inherently relative: boosting the saliency of a target object also depends on the global saliency distribution across all objects in the scene. Based on this insight, we propose GazeME, a lightweight framework that uses saliency-marked prompts, inserting learnable marker tokens around object descriptions to indicate which objects to visually emphasize or suppress. To learn these markers, we construct a saliency-semantics dataset that associates objects in image–prompt pairs with object-level saliency scores, and propose Saliency Prior Marker Activation (SPMA), a saliency-aware stochastic marker activation strategy that exploits relative saliency relationships for robust training. During inference, GazeME automatically inserts appropriate markers into the prompt, thereby directly enhancing the visual saliency of the target object. Extensive experiments demonstrate that GazeME effectively boosts target saliency while preserving both semantic alignment and image quality.

1 INTRODUCTION

Text-to-image generation models have achieved remarkable success in synthesizing high-fidelity images from text prompts, empowering creative workflows across design, advertising, and the visual arts. Yet semantic fidelity marks only the beginning of effective visual communication. For designers, the decisive question is not simply what appears in the frame, but what seizes the viewer's attention first (Sutcliffe & Namoune, 2008). An advertisement must make its product stand out, and a story illustration must direct focus to its protagonist. This shifts the challenge beyond semantics to the dynamics of visual attention: semantics dictates what to generate, whereas attention orchestrates where the eye is drawn, which is important for memory retention (Santangelo, 2015), cognitive comprehension (Jamet et al., 2008), and emotional resonance (Subramanian et al., 2014).

Visual saliency provides a standard approach for modeling human visual attention, using a saliency map to quantify the spatial distribution of human attention over an image (Kumain et al., 2025). While prior work has largely focused on predicting fixations from given images (Zhao et al., 2026; Kümmerer et al., 2016; Kroner et al., 2020), the inverse problem, generating images in which specified targets become visually salient, remains largely unexplored. A recent method, GazeFusion (Zhang et al., 2024), conditions text-to-image diffusion models on explicit saliency map inputs, aiming to generate images whose saliency distribution is consistent with the provided maps. However, requiring users to manually specify meaningful saliency maps before image generation places a burden on them. This calls for more intuitive interaction paradigms that allow users to achieve saliency-driven generation with minimal explicit input.

To address this, we introduce a task named *Target Saliency Boosting* in text-to-image generation, which offers an intuitive way to interact and control visual attention without explicit visual priors. Given a text prompt in which a target object is explicitly described, the objective is to generate an image in which the target-object region has high visual saliency. However, achieving this goal presents several challenges. (1) Visual attention is complex, entangled across low-level contrast,

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.

global composition, and high-level semantics. Thus, traditional control methods are often biased, lacking precision for effective modulation. (2) Saliency is relative and context-dependent. Boosting a target may suppress other elements and distort composition, breaking scene coherence. (3) A generalizable dataset that provides text prompts and corresponding object-level saliency scores is currently unavailable, which precludes effective supervision for our task.

In this work, we propose GazeME, a lightweight framework for this task. Our key insight is that visual saliency is inherently relative: boosting the prominence of a target object depends on the attention distribution across non-target entities. Based on this insight, we introduce the concept of saliency-marked prompts, where learnable marker tokens are inserted around object descriptions to explicitly indicate which objects should be visually emphasized or suppressed. To learn these marker embeddings, we construct an automatic saliency-semantics dataset that maps individual objects within image–prompt pairs to their relative visual saliency scores. We further propose saliency prior marker activation (SPMA), a saliency-aware stochastic marker activation strategy that exploits the relative saliency relationships among objects to construct diverse training samples. Only the marker embeddings are optimized, and this approach is lightweight and preserves the generalization of the pretrained model. During inference, given a source prompt and a target object, GazeME automatically inserts the appropriate markers into the prompt, enabling direct manipulation of the relative visual prominence of objects without modifying their underlying semantic content.

Additionally, we conduct extensive experiments to evaluate the effectiveness of GazeME. Quantitative and qualitative comparisons, shortcut analyses, and ablation studies demonstrate that our method can effectively enhance target saliency while avoiding trivial solutions such as blurring the background. Subjective evaluations from user studies further validate the empirical advantages of GazeME, demonstrating alignment with human perceptual preferences.

In summary, our contributions are threefold:

- We introduce *Target Saliency Boosting*, a novel task formulation that enables users to control visual saliency in text-to-image generation through natural language, without requiring any explicit saliency maps. This defines a new interaction paradigm where the model directly translates linguistic intent into visual saliency.
- We propose GazeME, a lightweight framework that introduces learnable marker tokens to encode the intent of enhancing or suppressing visual saliency. The framework integrates saliency-aware data construction, embedding learning, and marker-guided generation, with saliency prior marker activation (SPMA) enabling robust training and the saliency-guided generation strategy reinforcing the desired visual saliency during inference.
- We construct a language-grounded multi-object saliency dataset, where objects described in a prompt are associated with their visual saliency rankings and scores. We also develop an automatic annotation pipeline that derives object-level saliency supervision from image–prompt pairs, enabling scalable learning of saliency-aware marker embeddings.

2 Related Work

In this section, we review related works in visual saliency and controllable text-to-image generation.

2.1 Visual Saliency

Visual saliency, originally conceptualized in cognitive psychology, refers to the visual property that makes certain regions stand out from their surroundings and naturally draw human attention (Itti, 2007). We review it from three aspects: prediction methods, datasets, and applications.

Saliency Prediction. Saliency prediction aims to model where humans tend to look in a scene. This target is usually represented as a saliency map (e.g., a heatmap where pixel values encode the likelihood of attracting fixations), with brighter regions indicating higher visual priority (Koch & Ullman, 1985). To this end, early computational models rely on hand-crafted features, such as multi-scale center-surround contrast (Itti et al., 1998) and spectral residuals (Hou & Zhang, 2007). Deep learning methods then adopt convolutional neural networks (Kümmerer et al., 2016) and Transformers (Lou et al., 2022) to learn hierarchical representations, significantly improving prediction

accuracy. Recently, unified foundation models like Attend-to-Anything Model (AAM) (Zhao et al., 2026) consolidate multiple saliency tasks into a single framework. These advances provide reliable tools for estimating visual attention and lay the foundation for downstream applications.

Saliency Datasets. Existing saliency datasets are typically collected by presenting visual stimuli to multiple observers under free-viewing conditions and recording their gaze trajectories with eye trackers. Representative datasets include MIT1003 (Judd et al., 2009), U-EYE (Jiang et al., 2023), CAT2000 (Borji & Itti, 2015), SalECI (Jiang et al., 2022), SALICON (Jiang et al., 2015), and OSIE (Xu et al., 2014), covering natural images, web pages, and e-commerce content. However, existing datasets lack prompt-level grounding and sufficient scene diversity; although some provide annotations for salient regions or the most salient objects, they still offer insufficient supervision for low-saliency objects as negative samples.

Saliency-Guided Applications. Saliency information has been widely exploited in downstream vision tasks and practical applications. In salient object detection and salient object segmentation, saliency cues help identify and localize visually prominent objects (Borji et al., 2019). Saliency has also been used for image cropping (Xu et al., 2021), and image retrieval (Wei et al., 2019), as well as region-of-interest (ROI) coding for perceptually efficient video compression (Zhu & Xu, 2018). Recent methods such as GazeFusion further incorporate saliency information to guide image generation (Zhang et al., 2024). Despite these advances, creating images with user-specified saliency remains challenging, both in terms of interaction difficulty and the ability of generative models to understand and faithfully translate saliency intent.

2.2 Controllable Text-to-Image Generation

Controllable text-to-image generation extends standard text-to-image models with additional mechanisms that allow users to specify and refine desired properties of the generated images. We review it from two aspects: control modalities and control mechanisms.

Control Modalities. Controllable text-to-image generation extends natural language prompts with additional conditions to constrain the generated content. Common modalities include semantic segmentation masks, edge maps, depth maps, human poses, color palettes, and layout sketches, providing explicit spatial or structural guidance (Mou et al., 2024; Zhang et al., 2023; Huang et al., 2023). These conditions are typically encoded by dedicated modules and injected into pre-trained diffusion models. For example, AnyControl (Sun et al., 2024) supports multiple spatial conditions within a unified framework; GLIGEN (Li et al., 2023) grounds textual descriptions to specified spatial regions through bounding boxes; and SpaText (Avrahami et al., 2023) constructs spatially aligned semantic embeddings to enforce object-level layout constraints. Such modalities enable users to explicitly control properties, but using them generally requires semantic or spatial intent.

Control Mechanisms. Controllable generation can be achieved through different intervention mechanisms, including condition injection, embedding adaptation, attention manipulation, feature modulation, and denoising-process control. Condition-based methods such as ControlNet (Zhang et al., 2023) and T2I-Adapter (Mou et al., 2024) introduce auxiliary modules for external signals, while Textual Inversion (Gal et al., 2023) learns token embeddings to encode new concepts and LoRA (Hu et al., 2022) enables parameter-efficient adaptation. Attention- and feature-based methods directly modify cross-attention maps or intermediate representations for semantic and appearance control (Hertz et al., 2023). Alternatively, recent methods intervene in the denoising process itself, such as NoiseCtrl (Dai et al., 2025) and CogBlender (Dang et al., 2026), which manipulate sampling or velocity trajectories during generation. Despite these diverse mechanisms, existing methods primarily control *what* appears, *where* it appears, or *how* it looks, while leaving *how visual attention is distributed among objects* largely unexplored.

Overall, existing saliency prediction models and controllable text-to-image methods are insufficient to support our task. Saliency models mainly predict attention from existing images without semantic grounding, while controllable text-to-image methods focus on semantic, spatial, or appearance control, with limited support for saliency-aware generation. To bridge this gap, we introduce GazeME, enabling users to boost the visual saliency of a target object through natural language prompting.

Original Prompt	a photo of a dog and a cat.	a picture of a woman, a man and a child.
Marked Prompt	a photo of <focus> a dog </focus> and <unfocus> a cat </unfocus>.	a picture of <focus> a woman </focus>, a man and <unfocus> a child </unfocus>.
Marker Meaning	increase saliency of the dog, and suppress saliency of the cat	increase saliency of the woman, suppress saliency of the child, leave the man unconstrained

Figure 1: Examples of marked prompts, where `<focus></focus>` wraps the target object whose saliency should be boosted, while `<unfocus></unfocus>` wraps the object whose saliency should be suppressed.

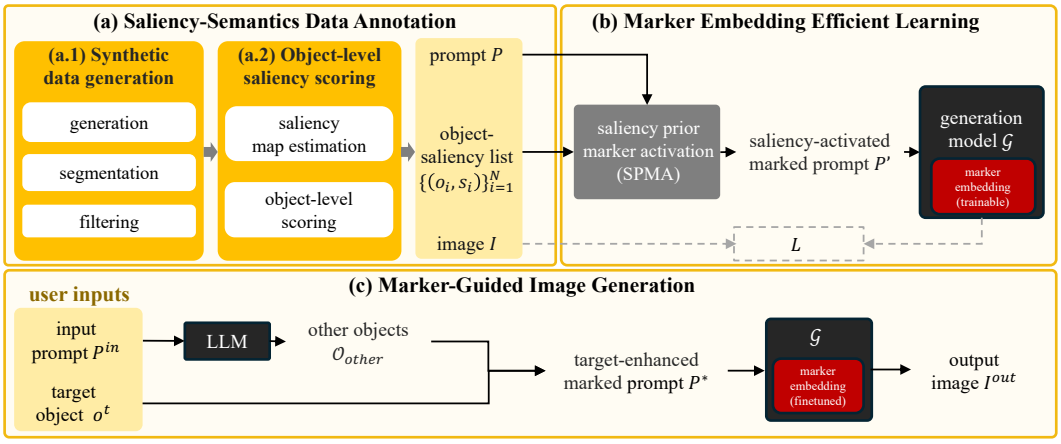

Figure 2: Overview of GazeME, which consists of three stages: (a) saliency-semantics data annotation, (b) marker embedding efficient learning, and (c) marker-guided image generation.

3 METHODOLOGY

In this section, we present GazeME for *Target Saliency Boosting*, centered on the concept of **saliency-marked prompts**. As shown in Figure 1, we introduce two pairs of bidirectional marker tokens, `<focus>/</focus>` and `<unfocus>/</unfocus>`, which are inserted around object descriptions to indicate whether an object should be visually emphasized or suppressed. GazeME learns saliency knowledge in these marker embeddings and uses them to guide image generation. As illustrated in Figure 2, our framework consists of three stages: (a) **saliency-semantics data annotation**, which automatically infers object-level relative saliency from image–prompt pairs; (b) **marker embedding efficient learning**, which optimizes only the marker embeddings to encode saliency knowledge; and (c) **marker-guided image generation**, which inserts the learned markers into a user prompt P^{in} according to a target object o^t and generates the output image I^{out}. We detail each stage below.

3.1 SALIENCY-SEMANTICS DATA ANNOTATION

We construct a dataset of 1,637 image–prompt pairs annotated with object-level saliency scores to train saliency-aware marker embeddings. Each sample is a tuple $\{I, P, \{(o_i, s_i)\}_{i=1}^N\}$, where I is an image, P is the corresponding prompt, and s_i is the saliency score of the object o_i mentioned in P, quantifying the average saliency within the object region relative to the global distribution. As shown in Figure 2, the data construction consists of two stages: (a.1) **synthetic data generation**, which produces diverse image–prompt pairs with object masks; and (a.2) **object-level saliency scoring**, which assigns a saliency score to each annotated object.

Synthetic data generation. We obtain the data through three steps: data generation, segmentation, and filtering. (1) *Generation.* We first prompt a Large Language Model (LLM, e.g., DeepSeek-V4-Flash in our implementation) to generate a diverse set of prompts, with each prompt P accompanied by a list of noun phrases $\{o_1, o_2, \ldots, o_N\}$ denoting the objects in the scene. We then synthesize the corresponding image I using a text-to-image generation model (e.g., FLUX.1-dev in our imple-

mentation). (2) *Segmentation.* Given I, we obtain a binary object mask M_i for each object o_i using Segment Anything Model 3 (SAM3) (Carion et al., 2025), a state-of-the-art segmentation model capable of segmenting objects specified by natural-language concepts. (3) *Filtering.* To ensure reliable object-level annotations, we remove samples with repeated object categories in the prompt, which may cause ambiguity in automatic object segmentation and matching. We also discard samples in which the queried object cannot be detected by SAM3.

Object-level saliency scoring. We compute object-level saliency scores in two steps. (1) *Saliency map estimation.* Given the synthesized image I, we compute a pixel-level saliency map S using the Attend-to-Anything Model (AAM) (Zhao et al., 2026), a state-of-the-art foundation model for visual saliency prediction. (2) *Object-level scoring.* Given the saliency map S and the object mask M_i, we compute an object-level saliency score for each object by adapting the formulation of the normalized scanpath saliency (NSS) (Bylinskii et al., 2019), defined as

$$s_i = \frac{1}{\sigma_S}\left(\frac{\sum_p S(p)M_i(p)}{|M_i|} - \mu_S\right), \tag{1}$$

where $S(p)$ denotes the saliency value at pixel p, and $M_i(p) \in \{0,1\}$ indicates whether pixel p belongs to the object mask M_i, μ_S and σ_S denote the mean and the population standard deviation of S over the whole image, respectively, and $|M_i|$ denotes the area of mask M_i in pixels. The higher the value of s_i, the more salient the object region is; conversely, a lower value indicates lower saliency.

3.2 Marker Embedding Efficient Learning

With the constructed dataset, we train the saliency-aware marker embeddings to instill saliency awareness into the text-to-image model. To enhance robustness, generalization, and discriminative power along both positive and negative saliency directions, we propose **saliency prior marker activation** (SPMA), a lightweight, saliency score-conditioned mechanism that dynamically inserts markers into prompts during training. Moreover, instead of fine-tuning the full model, we freeze all backbone parameters and optimize only the newly introduced marker token embeddings. We next describe SPMA and the training procedure for learning the marker embeddings.

Saliency prior marker activation (SPMA). SPMA takes a prompt P as input and dynamically outputs a saliency-activated marked prompt P' based on the saliency scores $\{s_i\}$ among objects $\{o_i\}$, forming diverse marker configurations that serve as an implicit ensemble over different saliency directions and object combinations. It encourages the marker embeddings to capture saliency relationships rather than memorize a fixed pattern. Specifically, SPMA consists of three steps: (1) *saliency-based grouping*, which divides objects into high- and low-saliency groups; (2) *stochastic marker activation*, which probabilistically selects objects to apply markers; and (3) *saliency-consistent truncation*, which applies a truncation operation to correct the activation values while preserving the relative saliency ordering among objects.

(1) *Saliency-based grouping.* Given the saliency score s_i of each object o_i, we first divide the objects into high- and low-saliency groups using a threshold γ:

$$y_i = \begin{cases} 1, & s_i \geq \gamma, \\ 0, & s_i < \gamma, \end{cases} \tag{2}$$

where $y_i = 1$ and $y_i = 0$ indicate high- and low-saliency objects, respectively. We set γ as the per-sample mean saliency score to separate objects with positive and negative saliency score values.

(2) *Stochastic marker activation.* We introduce a binary activation variable $z_i \in \{0,1\}$ for each object o_i to indicate whether a marker is inserted around its description in the prompt, where $z_i = 1$ means that the object is selected and wrapped by its corresponding marker, and $z_i = 0$ means that no marker is inserted around it. The marker type is determined by the grouping label y_i. Objects with $y_i = 1$ are wrapped by `<focus>`/`</focus>`, corresponding to high saliency, while objects with $y_i = 0$ are wrapped by `<unfocus>`/`</unfocus>`, corresponding to low saliency.

For each object, we sample z_i from a Bernoulli distribution with saliency-dependent probability p_i. Specifically, we consider the following four cases. For high-saliency objects ($y_i = 1$): the most salient object, i.e., the one with maximum s_i, is always activated with $p_i = 1$. For each remaining

high-saliency object, its activation probability is defined by:

$$p_i = \frac{s_i - \gamma}{s_i^{\text{up}} - \gamma}, \tag{3}$$

where $s_i^{\text{up}} = \min_{s_j > s_i} s_j$ denotes the nearest higher saliency score. For low-saliency objects ($y_i = 0$): the least salient object, i.e., the one with minimum s_i, is always activated with $p_i = 1$. For each remaining low-saliency object, its activation probability is defined by:

$$p_i = \frac{\gamma - s_i}{\gamma - s_i^{\text{down}}}, \tag{4}$$

where $s_i^{\text{down}} = \max_{s_j < s_i} s_j$ denotes the nearest lower saliency score.

(3) *Saliency-consistent truncation.* Finally, we enforce saliency ordering consistency by truncating the sampled activation pattern $\{z_i\}$. Specifically, for the high-saliency group, an object can remain activated only if all objects with higher saliency scores are also activated; otherwise, it is deactivated. Conversely, for the low-saliency group, an object can remain activated only if all objects with lower saliency scores are also activated; otherwise, it is deactivated.

This design has three key effects. Objects with a saliency score closer to γ are more likely to be deactivated, while objects farther from γ are more likely to be activated. Moreover, objects with similar saliency scores tend to be activated together. Additionally, the truncation rule preserves their relative saliency ordering, preventing contradictory activation patterns.

Training procedure. We adopt a lightweight training strategy. With all other model parameters frozen, we optimize θ (the embeddings of four marker tokens: `<focus>`, `</focus>`, `<unfocus>`, and `</unfocus>`) using the flow matching objective:

$$L = \mathbb{E}_{t,\epsilon}\left[\left\|v_\theta\left(x_t, t, P'\right) - u\right\|_2^2\right] \tag{5}$$

where x_t denotes the latent representation of the original image I after adding Gaussian noise (parameterized by ϵ) at timestep t, u denotes the target velocity, and $v_\theta(x_t, t, P')$ denotes the velocity predicted by the generative model conditioned on the saliency-activated marked prompt P'.

3.3 Marker-Guided Image Generation

Based on the learned marker embeddings, we employ a saliency-enhanced generation strategy to explicitly enforce object-level saliency guidance during image generation. Given an input prompt P^{in} and a target object o^t, we first employ an LLM to identify the other objects $\mathcal{O}_{other}$ mentioned in P^{in}. We then construct a saliency-enhanced marked prompt P^* by assigning the `<focus>` marker to the target object and the `<unfocus>` marker to the remaining objects. The resulting prompt explicitly represents o^t as the focus object and the objects in $\mathcal{O}_{other}$ as unfocus objects. The constructed prompt P^* is then fed into the text-to-image generative model with the learned marker embeddings to synthesize the final image I^{out}, thereby enhancing the visual saliency of the target object while suppressing the saliency of competing objects.

Implementation details. We adopted FLUX.1-dev as the base text-to-image model. The model parameters were optimized using the AdamW optimizer with a learning rate of 3×10^{-4}.Training was performed for 2,000 steps with a batch size of 2 on a single NVIDIA H200 GPU, taking approximately two hours. At inference, images were generated at 512×512 resolution with 30 sampling steps. We set the random seed to 42 for all experiments.

4 Experiments

In this section, we first describe the experimental setup, then present the main comparison against baselines with shortcut analysis, followed by ablation studies. Additionally, we conduct a user study to validate the effectiveness of GazeME.

4.1 Experimental Setup

In this section, we briefly introduce the experimental setup, including testset construction, baseline and metric selection. For detailed implementation, please refer to Appendix B.

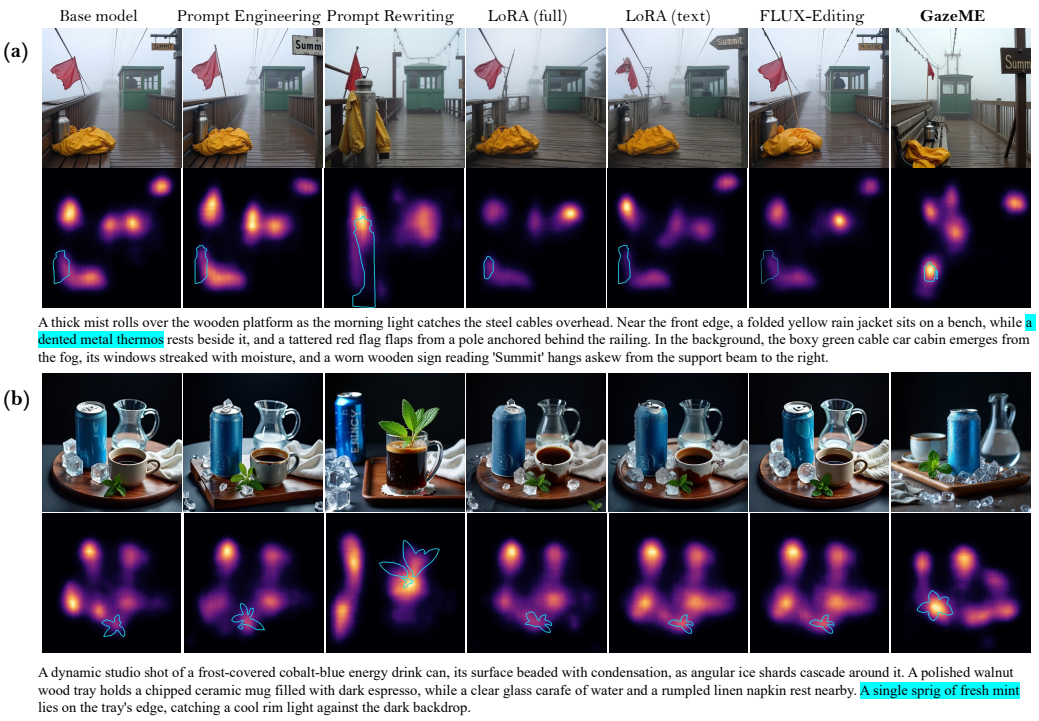

Figure 3: Qualitative results. We show the generated images and the corresponding saliency maps. The target object regions are outlined in cyan on the saliency maps.

Testset. We construct a target-saliency testset containing 432 challenging target saliency boosting tasks and each takes a prompt–target object pair as input. To build each pair, we generate diverse prompts, render images with the frozen base model, and select as the target an object that the untrained base model places in the bottom half of its own saliency score distribution (i.e., an object that is not salient by default). We will release the testset to ensure reproducibility.

Baselines. We compare GazeME with five representative methods: (1) *Prompt Engineering*, which directly injects the target-saliency intent into the text-to-image prompt; (2) *Prompt Rewriting*, which queries an LLM to rewrite the input prompt into a more descriptive caption that emphasizes the target object; (3) *LoRA (full)* (Hu et al., 2022), which adopts low-rank adaptation to finetune the base model for target saliency enhancement, adapting all attention and feed-forward layers; (4) *LoRA (text)*, a capacity-localized LoRA variant that restricts adaptation to the text-side attention layers; (5) *FLUX-Editing*, which first generates an image and then edits it with an instruction to improve target saliency.

Metrics. We evaluate all methods along three dimensions. (1) *Saliency alignment.* We report Target-NSS, the saliency score of the target object, computed by Equation 1. Higher values indicate that the target object is more visually salient. (2) *Semantic alignment.* We report CLIPScore (Hessel et al., 2021), the semantic similarity between the generated image and the input prompt. Higher values indicate better prompt fidelity. (3) *Image quality.* We report CLIPIQA (Wang et al., 2023), a no-reference perceptual quality metric. Higher values indicate better perceived image quality.

4.2 Baseline Comparison

We compare GazeME with the baselines from both quantitative and qualitative aspects. In addition, we perform a shortcut analysis with baselines on low-level factors that may act as saliency shortcuts.

Quantitative Comparison. Table 1 compares all methods on the testset. In terms of saliency alignment (Target-NSS), GazeME achieves the highest score and clearly outperforms prompt-level, training-based, and editing-based baselines. For semantic alignment (CLIPScore), GazeME remains competitive with the baselines and stays within a similar range, lying between Prompt Rewriting and

Table 1: Quantitative comparison.

Method	Target-NSS↑	CLIPScore↑	CLIPIQA↑
Base Model	0.1791	0.2890	0.9463
Prompt Engineering	0.2340	0.2846	0.9523
Prompt Rewriting	0.2378	0.2774	0.9485
LoRA (full)	0.2521	0.2872	0.9429
LoRA (text)	0.2313	0.2896	0.9458
FLUX-Editing	0.2554	0.2899	0.9426
GazeME	0.2761	0.2820	0.9516

Table 2: Results of the ablation study.

Method	Target-NSS↑	CLIPScore↑	CLIPIQA↑
GazeME($\gamma{=}\mu_{\text{sample}}$)	0.2761	0.2820	0.9516
$\gamma{=}0$	0.2724	0.2819	0.9508
$\gamma{=}m_{\text{dataset}}$	0.2744	0.2821	0.9510
$\gamma{=}\mu_{\text{dataset}}$	0.2720	0.2820	0.9515
$\gamma{=}m_{\text{sample}}$	0.2757	0.2814	0.9512
w/o SPMA	0.2276	0.2864	0.9508
w/o training	0.2192	0.2870	0.9503
w/o <unfocus> in P^*	0.2521	0.2864	0.9501

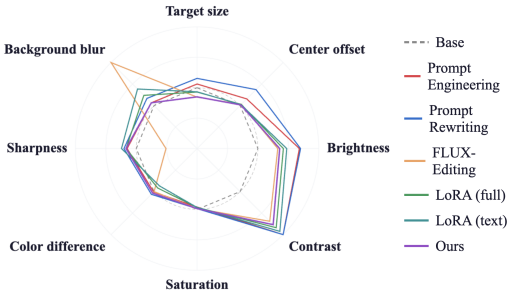

Figure 4: Shortcut analysis.

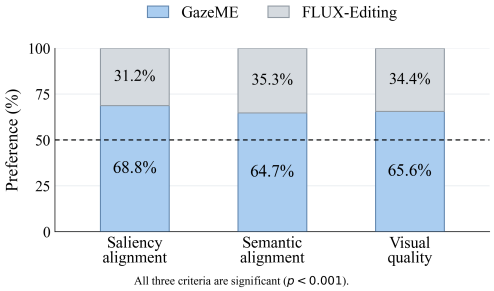

Figure 5: User study results.

Base Model, suggesting no substantial degradation in prompt fidelity. For visual quality (CLIPIQA), GazeME reaches the top level, matching the best baseline. Overall, these results demonstrate that GazeME is effective at enhancing target saliency while preserving semantic and perceptual quality, rather than merely optimizing a single metric.

Qualitative Comparison. Figure 3 qualitatively compares GazeME with the baselines. GazeME consistently makes the target object the primary visual focus, whereas other methods, although capable of partially enhancing the target's saliency, often also amplify other objects or introduce unnatural artifacts, and thus fail to make the target stand out naturally. These examples are consistent with the quantitative results and further demonstrate that GazeME enhances target saliency in a natural and controllable manner.

Shortcut Analysis. Saliency is not a direct function of low-level pixel statistics, yet low-level factors can mimic it, allowing methods to appear to improve saliency by biasing these factors rather than redistributing attention. To examine this, we analyze eight low-level visual factors, as shown in Figure 4. The radar chart (log scale) shows the ratio of each method's factor value to the base model's value (1.0) on each axis, except for center offset, where we plot the inverse ratio. The axes cover target size, center offset, brightness, contrast, saturation, color difference, sharpness, and background blur. FLUX-Editing mainly blurs the background, while prompt-level methods tend to enlarge the target, shift it toward the center, and increase brightness and contrast. GazeME shows deviations in brightness, and contrast, far smaller than other baselines and not concentrated on a single axis. Overall, GazeME keeps the factor distribution closest to the base model, confirming that it improves saliency mainly through saliency redistribution rather than low-level factor shortcuts. For detailed implementation of the factors, please refer to Appendix B.3.

4.3 Ablation Study

We conduct ablation studies from two aspects of GazeME: the marker mechanism and the parameter selection.

Marker mechanism. We ablate three design choices of our marker mechanism: SPMA supervision, marker training, and the inclusion of the <unfocus> marker in P^* (Table 2). Removing SPMA (*w/o SPMA*, i.e., only the marker of the highest-saliency object is activated during training) drops Target-NSS from 0.2761 to 0.2276, and *w/o training*, which keeps the marker embeddings fixed

without optimization, drops it to 0.2192, showing that both are essential for steering attention toward the target. Removing <unfocus> in P^* reduces Target-NSS to 0.2521, confirming that explicitly suppressing non-target objects further improves saliency control.

Parameter selection. We ablate the threshold γ that splits objects into high- and low-saliency groups for SPMA. We compare a fixed threshold (0), dataset-level statistics (m_{dataset} for the saliency score median and μ_{dataset} for the saliency score mean), and per-sample statistics (m_{sample} and μ_{sample}). As shown in Table 2, Target-NSS varies only slightly across all variants, while CLIPScore and CLIPIQA remain nearly constant. This indicates that GazeME is robust to the exact choice of γ. We adopt the per-sample mean setting (μ_{sample}), which gives the best Target-NSS and guarantees that every sample contains both high- and low-saliency groups for bidirectional supervision.

4.4 User Study

We conducted a user study with 26 participants to compare GazeME with the strongest baseline, *FLUX-Editing*. We randomly selected 17 prompt–target object pairs as evaluation tasks, and for each task we generated an image pair consisting of one image from GazeME and one from the baseline. Each participant evaluated these 17 image pairs. The order of the two images within each pair was randomized. For each pair, participants answered three two-alternative questions to evaluate (1) *saliency alignment*, (2) *semantic alignment*, and (3) *visual quality*. Detailed descriptions of the questions and the procedure can be found in Appendix B.4.

As shown in Figure 5, GazeME received 68.8% of the votes for saliency alignment, 64.7% for semantic alignment, and 65.6% for visual quality. We tested these participant-level majorities using two-sided exact binomial tests against a 50% chance level. All three results remain significant ($p < 10^{-3}$), indicating a consistent preference for GazeME across the three criteria. The user study confirms the perceptual advantage of GazeME against the strongest baseline.

5 Discussion

Our results suggest that target saliency is best treated as a relative property. The bidirectional markers provide a lightweight interface for controlling this relativity. SPMA encourages the marker embeddings to learn a robust saliency knowledge. Our work complements existing control over *what* objects appear or *where* they appear by targeting *how attention is distributed* among them.

Limitations and future work. Our approach has several limitations. First, we rely on automatic predictors for saliency estimation and object segmentation, which may introduce noise and bias into training. A promising solution is iterative refinement with human feedback, though this requires a careful annotation pipeline and a robust learning scheme for noisy or conflicting supervision. Second, while our markers encode relative saliency, we do not explicitly model spatial interactions or occlusion, which may limit performance in cluttered scenes. Extending the marker design to structured or pairwise relations is non-trivial due to their combinatorial nature. Finally, our framework targets static images; extending it to video would require temporally coherent markers to maintain saliency intent across frames, which is beyond our current scope.

6 Conclusion

In this paper, we introduce *Target Saliency Boosting*, a new task for controlling visual attention in text-to-image generation without explicit visual priors. We propose GazeME, which encodes saliency knowledge into bidirectional marker tokens and applies them to guide image generation. Extensive experiments demonstrate that GazeME effectively boosts target saliency while preserving semantic alignment and image quality. We will release the constructed dataset and the source code to facilitate future research. We hope our work can inspire further exploration of visual attention-aware controllable generation and broader applications in areas such as computational advertising, visual design, and educational content production.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic datasets, cleaning and reformatting datasets, supporting qualitative/thematic data analysis, assisting with translation, designing or providing feedback on research methodology or experiments, interpreting results and implementing methods. We have not used generative AI tools for the other required-disclosure tasks, including developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses; these tasks are not applicable to this work. Additionally, we used generative AI tools for formulating questions for surveys or interviews, creating or modifying scientific figures or images, creating or editing software code, brainstorming, identifying relevant literature, proposing a title or keywords for the research paper, and editing the research paper to improve readability. We have reviewed all AI-assisted work. Specifically, synthetic datasets were checked for validity and consistency with the intended data-generating process; AI-assisted qualitative/thematic analysis was reviewed by the authors against the underlying data; translations were checked by an author fluent in the relevant languages; AI-assisted methodology or experiment design and feedback were reviewed and approved by the authors; AI-assisted method implementation was tested and verified by the authors; survey/interview questions were reviewed and revised by the authors; scientific figures/images were checked for accuracy and approved by the authors; AI-generated or AI-edited code was tested for correctness by the authors; AI-drafted text was fact-checked and edited by the authors; literature identified with AI assistance was verified against the original sources; and title, keywords, and readability edits were reviewed and approved by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work involves a user study with 26 voluntary human participants. The user study was conducted via an anonymous online questionnaire, and participants could withdraw from this study at any time. No personally identifiable information was gathered beyond coarse demographic statistics (age range and gender). The study did not involve invasive devices, sensitive personal data, or more than minimal risk. All images shown to participants were generated by the generative models and were reviewed to ensure that they did not contain harmful, offensive, or NSFW content.

REPRODUCIBILITY STATEMENT

We have made extensive efforts to ensure the reproducibility of our work. The source code and the constructed dataset will be publicly released upon publication. Detailed dataset statistics and annotation examples are provided in Appendix A. The experiment details, including baseline implementations, metric computations, shortcut factor definitions, and user study procedure, are fully described in Appendix B. Additional results are reported in Appendix C.

REFERENCES

Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18370–18380, 2023.

Ali Borji and Laurent Itti. CAT2000: A large scale fixation dataset for boosting saliency research. *CoRR*, abs/1505.03581, 2015.

Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Computational Visual Media*, 5(2):117–150, 2019.

Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):740–757, 2019.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma,

Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. *CoRR*, abs/2511.16719, 2025.

Longquan Dai, He Wang, and Jinhui Tang. Noisectrl: A sampling-algorithm-agnostic conditional generation method for diffusion models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18093–18102, 2025.

Shengqi Dang, Jiaying Lei, Yi He, Ziqing Qian, and Nan Cao. Cogblender: Towards continuous cognitive intervention in text-to-image generation. *CoRR*, abs/2603.09286, 2026.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021.

Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

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. In *International Conference on Learning Representations*, 2022.

Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13753–13773, 2023.

L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998. doi: 10.1109/34.730558.

Laurent Itti. Visual salience. *Scholarpedia*, 2(9):3327, 2007.

Eric Jamet, Monica Gavota, and Christophe Quaireau. Attention guiding in multimedia learning. *Learning and Instruction*, 18(2):135–145, 2008.

Lai Jiang, Yifei Li, Shengxi Li, Mai Xu, Se Lei, Yichen Guo, and Bo Huang. Does text attract attention on e-commerce images: A novel saliency prediction dataset and method. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2078–2087, 2022.

Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1072–1080, 2015.

Yue Jiang, Luis A. Leiva, Hamed Rezazadegan Tavakoli, Paul R. B. Houssel, Julia Kylmälä, and Antti Oulasvirta. Ueyes: Understanding visual saliency across user interface types. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2023.

Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *Proceedings of International Conference on Computer Vision*, pp. 2106–2113, 2009.

Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, 4(4):219–227, 1985.

Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. Contextual encoder–decoder network for visual saliency prediction. *Neural Networks*, 129:261–270, 2020.

Sandeep Chand Kumain, Maheep Singh, and Lalit Kumar Awasthi. Revisited visual saliency detection with deep learning: A review of recent advancements. *ACM Computing Surveys*, 58(6), December 2025.

Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. Deepgaze II: reading fixations from deep features trained on object recognition. *CoRR*, abs/1610.01563, 2016.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: open-set grounded text-to-image generation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22511–22521, 2023.

Jianxun Lou, Hanhe Lin, David Marshall, Dietmar Saupe, and Hantao Liu. Transalnet: Towards perceptually relevant visual saliency prediction. *Neurocomputing*, 494:455–467, 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. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4296–4304, 2024.

Valerio Santangelo. Forced to remember: When memory is biased by salient information. *Behavioural Brain Research*, 283:1–10, 2015.

Ramanathan Subramanian, Divya Shankar, Nicu Sebe, and David Melcher. Emotion modulates eye movement patterns and subsequent memory for the gist and details of movie scenes. *Journal of Vision*, 14(3):31–31, 03 2014.

Yanan Sun, Yanchen Liu, Yinhao Tang, Wenjie Pei, and Kai Chen. Anycontrol: Create your artwork with versatile control on text-to-image generation. In *Proceedings of European Conference on Computer Vision*, volume 15069, pp. 92–109, 2024.

Alistair Sutcliffe and Abdallah Namoune. Getting the message across: visual attention, aesthetic design and what users remember. In *Proceedings of ACM Conference on Designing Interactive Systems*, DIS '08, pp. 11–20, 2008.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023.

Shikui Wei, Lixin Liao, Jia Li, Qinjie Zheng, Fei Yang, and Yao Zhao. Saliency inside: Learning attentive cnns for content-based image retrieval. *IEEE Transactions on Image Processing*, 28(9): 4580–4593, 2019.

Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of Vision*, 14(1):28–28, 01 2014.

Yifei Xu, Wujiang Xu, Mian Wang, Li Li, Genan Sang, Pingping Wei, and Li Zhu. Saliency aware image cropping with latent region pair. *Expert Systems with Applications*, 171:114596, 2021.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pp. 3813–3824, 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), November 2024.

Wenzhuo Zhao, Ronghao Xian, Keren Fu, and Qijun Zhao. Attend to anything: Foundation model for unified human attention modeling. *CoRR*, abs/2606.03540, 2026.

Shiping Zhu and Ziyao Xu. Spatiotemporal visual saliency guided perceptual high efficiency video coding with neural network. *Neurocomputing*, 275:511–522, 2018.

A Dataset Details

The annotated dataset consists of 1,637 image–prompt pairs. Each prompt depicts a multi-object scene and contains a list of objects. Table 3 summarizes the statistics of the annotated dataset. The object-level saliency scores in the dataset range from -0.85 to 6.19 (mean 1.05, median 0.93; 25th–75th percentiles: 0.40–1.57). Finally, we illustrate a representative data sample in Figure 6.

Table 3: Statistics of the annotated dataset. Here, p25–p75 denotes the 25th and 75th percentiles.

Statistic	Value	Statistic	Value
Number of image–prompt pairs	1,637	Average objects per sample	4.26
Average prompt length	49.0 words	Object count (range; median)	2–8; 4
Prompt length (p25–p75)	37–59	Average focus phrase length	2.98 words

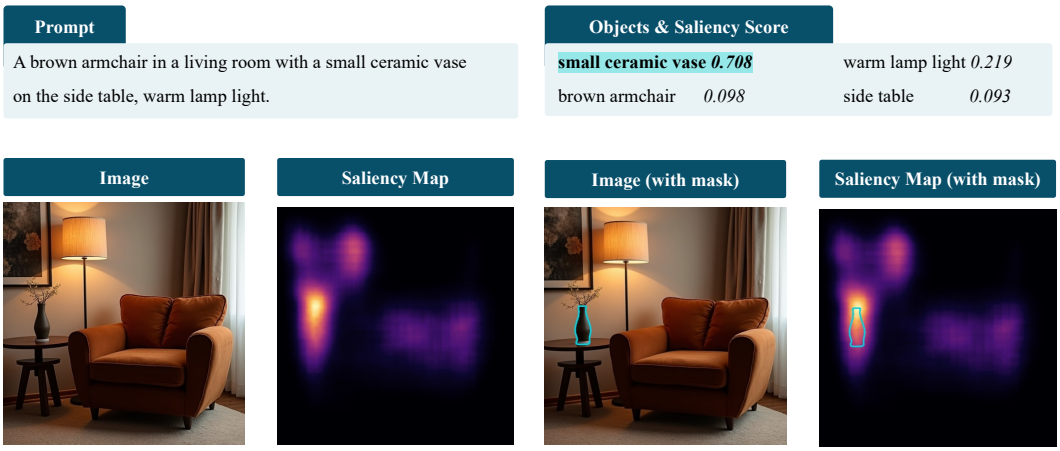

Figure 6: An example from the annotated image–prompt dataset.

B Experimental Details

We provide additional experimental information, including baseline implementations, evaluation metric computation, shortcut factor definitions, and details of the user study.

B.1 Baseline Implementation

We compare GazeME against five representative baselines, covering three categories: *prompt-level* methods (Prompt Engineering and Prompt Rewriting), *training-based* methods (LoRA (full) and LoRA (text)), and an *editing-based* method (FLUX-Editing). For each baseline, we describe its core mechanism, key hyperparameters, and how it is adapted to Target Saliency Boosting. All methods use the same base model and testset unless otherwise specified.

Prompt Engineering. This baseline operates purely at the prompt level. We simply append a fixed natural-language suffix, "focused on [o]", to the original prompt, where o denotes the target object. The base model then directly generates an image from the modified prompt.

Prompt Rewriting. This baseline uses a large language model (LLM, e.g., Deepseek-V4-Flash in our implementation) to rewrite the original prompt so that the target object is explicitly emphasized. We query the LLM with the following system and user prompts:

> **System prompt:** You are a professional art director at rewriting text-to-image generation prompts. Rewrite the prompt so that a viewer's attention is unmistakably drawn to the FOCUS OBJECT. Feel free to use professional visual-design vocabulary (e.g., focal emphasis,

> visual hierarchy, color contrast, leading lines, negative space, tonal separation, compositional weight) to emphasize the focus object.
>
> Hard rules:
> 1. Do NOT add, remove, or change any object in the scene — same objects, same scene.
> 2. The focus phrase given to you must appear VERBATIM in your rewrite.
> 3. Keep it natural professional prose, 1–3 sentences. No lists, no markers, no tags.
> 4. Output JSON only with a single field: `rewrite`.
>
> **User prompt:** ORIGINAL PROMPT: [original prompt]. FOCUS OBJECT (must appear verbatim): [target object]. OTHER OBJECTS (must all remain): [other objects].

The base model then generates an image from the rewritten prompt.

LoRA (full). Building on Prompt Engineering, we train LoRA adapters on all attention and feed forward layers of the base model. We use rank 16, $\alpha = 16$, and dropout rate 0.0 for LoRA layers. Optimization uses AdamW with a learning rate of 10^{-4}, a cosine learning schedule with 200 warmup steps, and an effective batch size of 16, for $4{,}000$ steps.

LoRA (text). This variant therefore restricts LoRA to the text-side attention projection layers. Rank, α, optimizer, and training steps are identical to LoRA (full).

FLUX-Editing. This baseline is a two-stage pipeline. The base model first generates an initial image from the original prompt. Given a target object o, the image is then edited by the same model (FLUX.1-dev) using the instruction below " Refine this image so that [o] becomes the unmistakable visual focal point. Keep the scene, objects, and composition exactly the same. " Here, we set the editing strength to 0.6 over 30 total sampling steps.

B.2 Metric Implementation

We describe the computation details of the three evaluation metrics. All metrics are computed per sample and then averaged over the test set, and higher values indicate better performance.

Target-NSS. We measure the saliency of the target object by adapting the normalized scanpath saliency (NSS) formulation (Bylinskii et al., 2019). Specifically, we define Target-NSS as:

$$\text{Target-NSS} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sigma_{S_i}}\left(\frac{\sum_p S_i(p)M_i^t(p)}{|M_i^t|} - \mu_{S_i}\right), \tag{6}$$

where S_i is the predicted saliency map for the i-th test image I_i, M_i^t is the mask of the target object o_i^t, $S_i(p)$ denotes the saliency value at pixel p, $M_i^t(p) \in \{0,1\}$ indicates whether pixel p belongs to the target mask, μ_{S_i} and σ_{S_i} are the mean and the population standard deviation of the entire saliency map S_i, $|M_i^t|$ is the number of pixels in the target mask, and N is the number of test samples. Higher Target-NSS indicates that the target object is more visually salient.

CLIPScore. We evaluate semantic alignment between the generated image and the original prompt using a frozen CLIP model. For each sample, we compute the cosine similarity between the CLIP image embedding and the CLIP text embedding:

$$\text{CLIPScore} = \frac{1}{N}\sum_{i=1}^{N}\cos(f_I(I_i), f_T(P_i)), \tag{7}$$

where f_I and f_T denote the frozen CLIP image and text encoders, respectively, and P_i is the original prompt used to generate I_i. The same original prompt is used for all methods.

CLIPIQA. We adopt the official CLIPIQA implementation (Wang et al., 2023) to compute a no-reference perceptual quality score for each generated image. The final score is obtained by averaging the per-image quality scores:

$$\text{CLIPIQA} = \frac{1}{N}\sum_{i=1}^{N} q(I_i), \tag{8}$$

where $q(I_i)$ denotes the CLIPIQA score of image I_i.

B.3 SHORTCUT ANALYSIS DETAILS

We analyze eight low-level factors that could serve as appearance shortcuts. Let T denote the target region in the image and B its complement (the background region in the image). We detail the implementation of each factor below.

Target size. Target size measures the area of the target region as a percentage of the whole image:

$$\text{Target size} = \frac{|T|}{|I|} \times 100\%,$$

where $|T|$ and $|I|$ are the numbers of pixels in the target region and the whole image, respectively. A higher value means the target simply became larger, which can trivially increase saliency.

Center offset. Center offset measures the Euclidean distance between the centroid of the target region and the image center:

$$\text{Center offset} = \sqrt{(\bar{x}_T - c_x)^2 + (\bar{y}_T - c_y)^2},$$

where $(\bar{x}_T, \bar{y}_T)$ is the centroid of T in normalized image coordinates, and (c_x, c_y) denotes the image center. A smaller value means the target was dragged toward the center.

Brightness. Brightness measures the absolute brightening of the target relative to the whole image:

$$\text{Brightness} = \bar{g}_T - \bar{g}_I,$$

where $\bar{g}_T$ and $\bar{g}_I$ are the mean grayscale levels inside the target region and the entire image, respectively. A larger value indicates that the target was simply brightened.

Contrast. We use Michelson contrast to measure luminance contrast between the target and the background:

$$\text{Contrast} = \frac{\bar{L}_T - \bar{L}_B}{\bar{L}_T + \bar{L}_B},$$

where $\bar{L}_T$ and $\bar{L}_B$ are the mean grayscale levels inside the target region T and the background region B, respectively. A larger absolute value means the target was made more contrastive against the background, which can act as a shortcut.

Saturation. Saturation measures how much more saturated the target is relative to the background:

$$\text{Saturation} = \frac{\overline{\text{HSV-sat}}_T}{\overline{\text{HSV-sat}}_B},$$

where $\overline{\text{HSV-sat}}_T$ and $\overline{\text{HSV-sat}}_B$ are the mean HSV saturation values inside the target region T and the background region B, respectively. A larger value means the target was made more colorful than its surroundings.

Color difference. Color difference measures the Lab color separation between the mean target color and the mean background color:

$$\text{Color difference} = \left\| \mu_T^{\text{Lab}} - \mu_B^{\text{Lab}} \right\|_2,$$

where μ_T^{Lab} and μ_B^{Lab} are the mean target and background colors in CIE Lab, a perceptually uniform color space, and $\|\cdot\|_2$ denotes the Euclidean norm. A larger value means the target was hue-shifted or color-separated from the background.

Sharpness. Sharpness measures the relative crispness of the target compared with the background:

$$\text{Sharpness} = \frac{\text{Var}(\nabla^2 T)}{\text{Var}(\nabla^2 B)},$$

where $\text{Var}(\nabla^2 T)$ and $\text{Var}(\nabla^2 B)$ are the variances of the 4-neighbour Laplacian response inside the target region T and the background region B, respectively. The 4-neighbour Laplacian operator is applied to the grayscale image and measures local intensity variation by comparing each pixel with its four immediate neighbours. A larger value means the target was sharpened.

Background blur. Background blur measures the lack of background detail:

$$\text{Background blur} = \frac{10^4}{\mathrm{Var}(\nabla^2 B)},$$

where $\mathrm{Var}(\nabla^2 B)$ is the variance of the 4-neighbour Laplacian response inside the background region B. The scale factor is set to 10^4 in our implementation. A larger value means the background was blurred or exhibiting halation.

B.4 User Study Details

We detail the user study procedure and analysis below.

Participants and procedure. The study included 26 participants: 11 male and 15 female. Participants were 19–26 years old (median age: 23 years). All participants completed the questionnaire, resulting in 26 responses for each of the 17 image pairs. Each participant evaluated 17 image pairs. Within each pair, both images were generated using the same input prompt and target object, with one image generated by GazeME and the other by *FLUX-Editing*. The images were labeled A and B, and the assignment of methods to labels A and B was randomized across pairs. For each pair, participants answered the following three questions, selecting either image A or image B for each question: (1) *Saliency alignment:* "In the two images above, which image makes [target object] more visually salient than the other objects in the image?" (2) *Semantic alignment:* "Which of the two images better matches the following description: [input prompt for this image]?" (3) *Visual quality:* "Which of the two images has higher image quality?" The questionnaire used a two-alternative forced-choice format and did not provide a tie option.

Results and analysis. GazeME received 68.8% of choices for saliency alignment, 64.7% for semantic alignment, and 65.6% for visual quality. These percentages are reported descriptively. For participant-level statistical testing, we first summarized each participant's 17 responses for each criterion into a single binary preference label: a participant was labeled as preferring GazeME if they selected it in at least 9 of the 17 image pairs, and otherwise as preferring *FLUX-Editing*. This yielded 26 binary participant-level observations for each criterion. We then applied two-sided exact binomial tests against a 50% chance level. All three tests were significant at $p < 10^{-3}$.

To account for dependence across both participants and image pairs, we additionally fitted a crossed random-intercept Bayesian logistic regression model with participant- and stimulus-level random intercepts. The odds ratios for selecting GazeME over *FLUX-Editing* were 3.15 (approximate 95% credible interval: [2.45, 4.06]) for saliency alignment, 2.06 ([1.66, 2.55]) for semantic alignment, and 2.07 ([1.66, 2.57]) for visual quality. All three credible intervals exclude an odds ratio of 1.

C Additional Results

This section provides additional experimental results. We first present additional qualitative results, followed by a backbone ablation on Stable Diffusion 2.1, a cross-predictor evaluation, and an additional analysis of marker effectiveness. These results further validate the effectiveness and generalization of GazeME.

C.1 Additional Qualitative Results

We provide additional qualitative comparison results in Figure 7. These results cover a broader range of prompts and further demonstrate the effectiveness of GazeME in target saliency boosting. These examples complement the results shown in the main paper.

C.2 Backbone Ablation

To examine whether the marker mechanism extends beyond FLUX.1-dev, we apply the GazeME framework to Stable Diffusion 2.1 (SD2.1). We retrain the four marker embeddings in the SD2.1 text encoder while keeping all other model parameters frozen. As shown in Table 4, GazeME achieves the highest Target-NSS among all methods, slightly above SD2.1-Editing and LoRA (text), and clearly above prompt-level baselines and the base model. This indicates that the marker mechanism

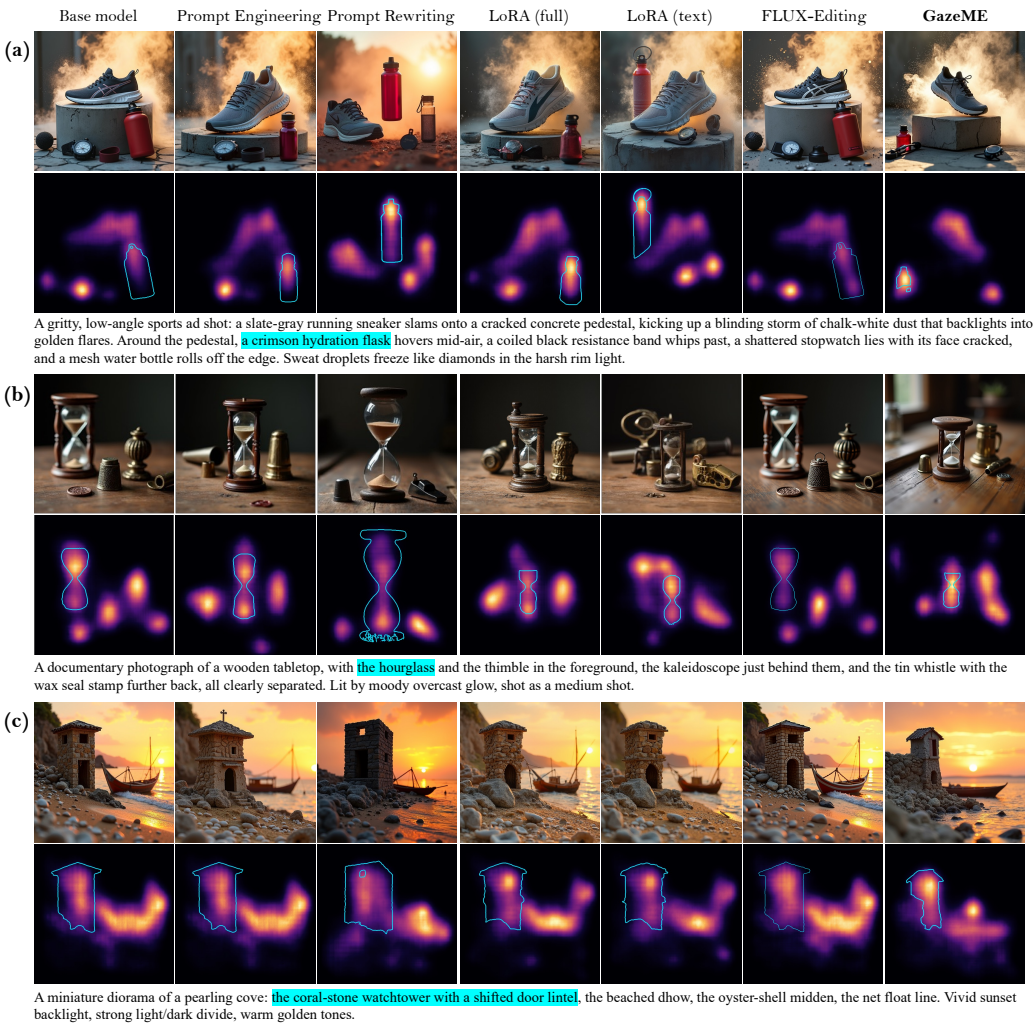

Figure 7: Additional qualitative comparisons with baselines.

Table 4: Results on another text-to-image generative model (Stable Diffusion 2.1).

Method	Target-NSS ↑	CLIPScore ↑	CLIPIQA ↑
Base Model	0.2060	0.2763	0.9231
Prompt Engineering	0.2808	0.2728	0.9261
Prompt Rewriting	0.2957	0.2722	0.9190
SD2.1-Editing	0.3375	0.2474	0.9330
LoRA (full)	0.2844	0.2805	0.9303
LoRA (text)	0.3213	0.2795	0.9300
GazeME	0.3405	0.2688	0.9259

Table 5: Effect of saliency predictor used for Target-NSS computation.

Predictor	AAM ↑	TranSalNet ↑
Base Model	0.1791	0.4791
Prompt Engineering	0.2340	0.5325
Prompt Rewriting	0.2378	0.5516
LoRA (full)	0.2521	0.5207
LoRA (text)	0.2313	0.5337
FLUX-Editing	0.2554	0.5571
GazeME	0.2761	0.5638

is not tied to a specific backbone and can effectively boost target saliency on a different text-to-image model. On CLIPScore, GazeME is slightly lower than LoRA (full) and LoRA (text) but remains comparable to the base model, suggesting that the saliency gain does not come at a substantial cost to prompt–image alignment. On CLIPIQA, GazeME is on par with the base model and Prompt Engineering, while SD2.1-Editing achieves the highest score. Overall, these results confirm that the marker mechanism generalizes across backbones and consistently improves target saliency while maintaining comparable semantic alignment and perceptual quality.

C.3 Cross Predictor Evaluation

We additionally evaluate the generated FLUX.1-dev images using TranSalNet (Lou et al., 2022) as an alternative saliency predictor. As shown in Table 5, GazeME achieves the highest Target-NSS under both AAM and TranSalNet, consistently outperforming all baselines. This ranking is stable across the two predictors: methods that perform well under AAM also tend to perform well under TranSalNet, and vice versa. For example, FLUX-Editing and Prompt Rewriting remain among the top baselines under both predictors, while the base model stays at the bottom. This consistency indicates that the improvement of GazeME is not an artifact of a specific saliency predictor, but reflects a genuine shift of visual attention toward the target object.

C.4 Additional Analysis of Marker Effectiveness

Table 6: Effectiveness of different marker assignments.

	Marker assignment			Contrast change
Configuration	Target	Most salient object	Other objects	Δgap $\uparrow$
Focus-only	`<focus>`	—	—	$+0.0465$
Boost–suppress	`<focus>`	`<unfocus>`	—	$+0.0577$
GazeME	`<focus>`	`<unfocus>`	all `<unfocus>`	$+0.0713$

Table 6 evaluates how different marker assignments affect the saliency distribution during inference. We compare three configurations. (1) *Focus-only* applies `<focus>` only to the target object. (2) *Boost–suppress* further applies `<unfocus>` to the most salient non-target object in the image generated directly from the base model (i.e., the object that competes most strongly with the target for attention). (3) GazeME applies `<unfocus>` to all non-target objects. To quantify the effect, we measure the change in the within-image contrast between the target and the most salient object:

$$\Delta\mathrm{gap} = (S_{\mathrm{target}} - S_{\mathrm{most}}) - (S^{\mathrm{base}}_{\mathrm{target}} - S^{\mathrm{base}}_{\mathrm{most}}),$$

where S_{target} and S_{most} are the saliency scores of the target object and the most salient object in the generated image, respectively, and $S^{\mathrm{base}}_{\mathrm{target}}$ and $S^{\mathrm{base}}_{\mathrm{most}}$ are the corresponding saliency scores in the image generated directly from the base model. A positive $\Delta\mathrm{gap}$ indicates that the target gains saliency relative to its strongest competitor. The results show that $\Delta\mathrm{gap}$ increases monotonically from Focus-only to Boost–suppress to GazeME, confirming that suppressing all non-target objects is more effective than suppressing only the most salient one.