PALO-STEER: NUISANCE-REDUCED SIGNAL CONSTRUCTION FOR TEXT-TO-IMAGE SAFETY STEERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image (T2I) diffusion models offer powerful generation capabilities but can also produce harmful content. To suppress harmful content, *activation-steering* methods construct signals from harmful-safe activation contrasts and apply them to intermediate activations during denoising. The effectiveness of these methods depends on how well the constructed signals capture harmful-relevant activation changes while minimizing nuisance. However, we identify two sources of nuisance in existing methods that can weaken safety steering: *Trajectory-induced confounding* mixes harmful-concept effects with differences in intermediate image states; *Aggregation-induced dilution* weakens the harmful-relevant contrast by pooling irrelevant image tokens. To address these issues, we propose *PaLo-Steer*, which constructs cleaner steering signals through state-controlled activation comparison and concept-aware spatial aggregation. (i) **Pa**rallel contrastive activation collection isolates the harmful-concept effect by collecting harmful- and safe-conditioned activations at shared intermediate states along a single harmful trajectory. (ii) Spatially **Lo**calized activation selection uses model-intrinsic cross-attention maps to locate image tokens relevant to harmful content. It then aggregates activation contrasts only over these tokens to reduce concept dilution. PaLo-Steer supports both vector- and transport-map-based steering. Across multiple safety benchmarks, it achieves state-of-the-art harmful-content suppression among parameter-free methods, reducing harmful rates by up to 12.0% over existing activation-steering methods while preserving benign generation quality.

Warning: This paper contains potentially offensive content.

1 INTRODUCTION

Recent advances in text-to-image (T2I) diffusion models (Rombach et al., 2022; Esser et al., 2024; Gao et al., 2025) have enabled highly realistic and semantically faithful image generation, while also raising significant safety concerns, as these models can be exploited to produce harmful or unsafe content (Li et al., 2025; Zhang et al., 2026). To mitigate harmful generation, various safeguards have been proposed, including prompt filtering (Rando et al., 2022), inference-time guidance (Schramowski et al., 2023; Gaintseva et al., 2026a), alignment-based methods (Liu et al., 2025), and model editing techniques (Gandikota et al., 2023; 2024; Gong et al., 2024). Among these approaches, inference-time guidance modifies the denoising process away from unsafe concepts through generation-time interventions, offering flexible safety control without additional training. While Safe Latent Diffusion (Schramowski et al., 2023) intervenes directly in the latent space, activation steering (Rodriguez et al., 2024; Gaintseva et al., 2026a;b) instead operates on intermediate hidden activations[1], enabling more fine-grained control over safety-relevant concepts without modifying model parameters.

Existing activation steering methods construct activation contrasts from harmful-safe prompt pairs[2]. For each pair, they collect image-token activations from independent denoising trajectories and aggregate each prompt's activations across all spatial positions. These aggregated activations are used to construct *steering signals*: harmful-safe difference vectors (Gaintseva et al., 2026a) or transport maps from harmful to safe activation distributions (Rodriguez et al., 2024) (Fig. 1(a)). Dur-

[1] Activations here refer to outputs of intermediate layers within the denoising network.

[2] Paired prompts share a benign context, concatenated with harmful and safe expressions, respectively.

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.

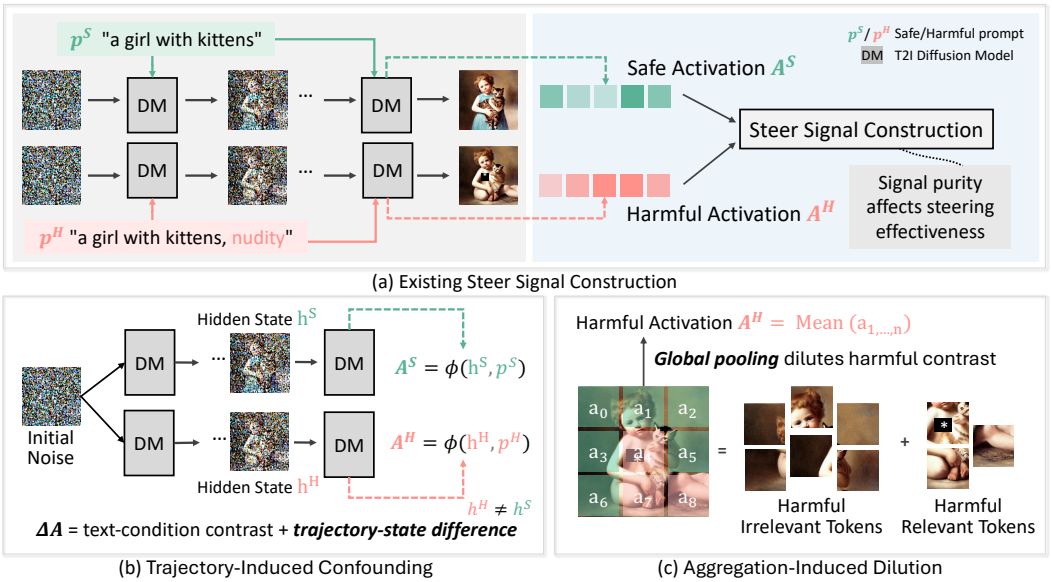


Figure 1: Trajectory confounding and concept dilution in existing steering-signal construction.

ing denoising, these signals are applied by subtracting the vectors from intermediate activations or mapping activations toward safe counterparts to suppress harmful-content generation. Within this pipeline, steering signals directly determine how activations are modified to guide generation, making their quality crucial to effective harmful-content suppression. Concretely, an effective signal should capture activation changes associated with the harmful concept while excluding harmful-irrelevant variation (which we term *nuisance*). Our perturbation experiments[3] further validate that including nuisance during signal construction can compromise signal quality and weaken harmful-content suppression (Sec. 3.2). This motivates a central question: *How can we construct steering signals that capture the harmful-concept effect while excluding nuisance?*

Through a detailed analysis of signal construction, we identify two sources of nuisance in existing methods. These arise during activation collection and aggregation despite harmful-safe prompt pairing: (i) **Trajectory-induced confounding** (Fig. 1(b)): Existing methods collect harmful and safe activations from independent denoising trajectories. Since activations depend on both the intermediate hidden state and the text condition, the resulting contrast mixes harmful-concept effects with state differences. These state differences can include harmful-irrelevant changes, such as the cat's fur color in Fig. 1(b), thereby contaminating the steering signal (Sec. 3.3). (ii) **Aggregation-induced dilution** (Fig. 1(c)): Existing methods pool all spatial image tokens, although harmful-concept evidence is localized. Including harmful-irrelevant activations dilutes the harmful-relevant contrast, as demonstrated by its progressive weakening in our controlled background-token inclusion analysis (Sec. 3.4). Together, these findings reveal why existing signal-construction pipelines are insufficient to isolate harmful-concept information from nuisance.

To address these two sources of nuisance, we propose *PaLo-Steer*, which contrasts harmful and safe activations at a shared state and aggregates the differences over harmful-relevant regions, yielding cleaner steering signals. It comprises two components: (i) **Parallel contrastive activation collection:** We compare harmful and safe text conditions at shared intermediate states to isolate the harmful-conditioning effect. Along a single harmful-conditioned trajectory, we reuse the state at each collection site to obtain a safe-conditioned reference alongside the harmful activation. Their contrast excludes differences between independently evolved states, addressing trajectory confounding in signal construction. (ii) **Spatially localized activation selection:** We use model-intrinsic information to locate regions associated with harmful content. Specifically, each image token is scored by its cross-attention to harmful text tokens, and the highest-scoring positions are selected as the spatial support for aggregating both harmful and safe-conditioned activations. This filters out irrelevant activations and reduces dilution of the harmful-concept contrast. Together, these components support more targeted interventions to suppress harmful content.

[3]We inject Gaussian noise into harmful and safe activations during signal construction to simulate nuisance. As noise increases, harmful-safe activation distributions become less separable and the harmful rate rises.

By constructing more effective steering signals, PaLo-Steer enhances existing activation-steering methods, including vector-based and transport-map-based. This makes it a training-free, plug-and-play approach that requires neither diffusion-model parameter updates nor external localization models. Across multiple safety benchmarks, PaLo-Steer improves both steering formulations, reducing harmful rates by up to 12.0% over the corresponding methods while preserving benign generation quality. Overall, our main contributions are as follows:

- We show that irrelevant activation variation weakens safety steering, and identify trajectory confounding and concept dilution as two sources of nuisance in existing signal-construction pipelines.
- We propose **PaLo-Steer**, which combines parallel contrastive activation collection and spatially localized activation selection to reduce these identified nuisances and construct more effective steering signals for harmful-content suppression.
- Across multiple safety benchmarks, PaLo-Steer improves both vector- and transport-map-based steering, reducing harmful rates by up to 12.0% while preserving benign generation quality.

2 Related Work

2.1 Text-to-Image Safeguards

Text-to-image (T2I) safeguards suppress harmful generation through prompt filtering (Rando et al., 2022), inference-time safety guidance (Schramowski et al., 2023; Yoon et al., 2025; Gaintseva et al., 2026a), safety alignment (Liu et al., 2025), and concept erasure or model editing (Gandikota et al., 2023; 2024; Gong et al., 2024). These approaches intervene at different stages: filtering screens inputs, alignment and editing modify the model, and inference-time guidance redirects generation during denoising. Among inference-time safeguards, Safe Latent Diffusion (SLD) (Schramowski et al., 2023) suppresses unsafe concepts by incorporating safety guidance into the predicted noise during denoising. Activation steering instead modifies internal activations using signals constructed from harmful and safe representations. Our work improves the construction of these signals to strengthen activation-level safety control.

2.2 Activation Steering for Generative Models

Activation steering controls generative models by applying concept-specific signals to internal activations during inference. First studied in large language models (LLMs) (Turner et al., 2023; Li et al., 2023; Wang et al., 2024; Lee et al., 2025), it provides a lightweight, parameter-free alternative to fine-tuning and model editing. Prior work on LLM steering vectors shows that greater separability between contrastive activations along the steering direction predicts stronger and more reliable steering, making separability a useful diagnostic of signal quality (Braun et al., 2025). Recent methods extend activation steering to text-to-image (T2I) diffusion models for visual control and harmful-content suppression (Rodriguez et al., 2024; Gaintseva et al., 2026a). Unlike LLMs, where the last-token activation can provide a compact sequence representation, T2I models produce spatial visual tokens corresponding to different image regions and interacting with text through cross-attention (Rombach et al., 2022; Esser et al., 2024). T2I signal construction must therefore account for both the denoising states from which activations are collected and the spatial tokens over which they are aggregated. We investigate how these two stages introduce nuisance, using activation separability and downstream harmful rates to assess its effect on safety steering.

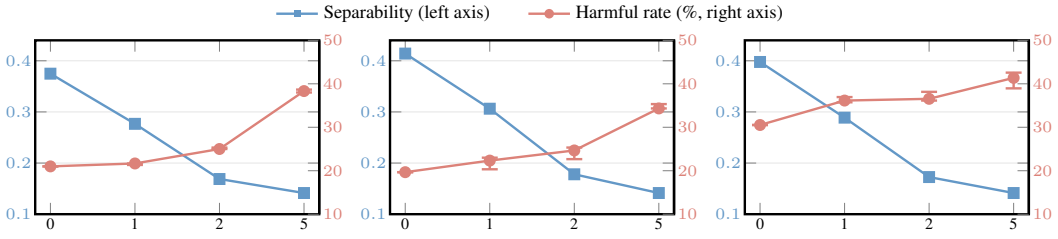
3 Nuisance in T2I Steering-Signal Construction

We examine how harmful-irrelevant variation enters steering-signal construction and affects harmful content suppression. We first introduce two typical activation steering methods (Sec. 3.1), then examine their sensitivity to irrelevant variation during signal construction (Sec. 3.2) and identify trajectory confounding and concept dilution (Secs. 3.3 and 3.4).

3.1 Preliminaries

To collect activations for steering-signal construction, existing methods first construct harmful-safe prompt pairs. Each pair consists of harmful prompt $p_i^H = b_i \oplus c_H$ and safe prompt $p_i^S = b_i \oplus c_S$, where b_i is a shared benign context[4], c_H is the harmful expression, c_S is a safe expression or the

[4] b_i is formed by combining person and scene phrases from predefined vocabularies (e.g., *a boy on the beach*).



(a) Sexual (b) Violence (c) Illegal Activities

Figure 2: Effect of activation noise on harmful-safe separability and harmful rate on SD-v1.5 over three runs.

empty string, and $\oplus$ denotes concatenation. Following CASteer (Gaintseva et al., 2026a), we collect image-token activations at the output of cross-attention, including the residual connection. At layer l and timestep t, the pair yields harmful activation $\mathbf{a}^H_{i,l,t}$ and safe activation $\mathbf{a}^S_{i,l,t} \in \mathbb{R}^{N\times d}$, with N tokens and d channels; $\mathbf{a}^r_{i,l,t}[n] \in \mathbb{R}^d$ denotes the activation of token n under condition $r \in \{H, S\}$. Existing methods globally pool these activations:

$$\mathbf{A}^H_{i,l,t} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{a}^H_{i,l,t}[n], \qquad \mathbf{A}^S_{i,l,t} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{a}^S_{i,l,t}[n]. \tag{1}$$

Given M prompt pairs, vector-based methods such as CASteer construct a mean-difference direction $\mathbf{v}_{l,t}$, whereas transport-based methods such as AcT (Rodriguez et al., 2024) learn a sample-wise map $\mathcal{T}_{l,t}$:

$$\mathbf{v}_{l,t} = f_{\text{norm}}\left(\frac{1}{M}\sum_{i=1}^{M}\left(\mathbf{A}^H_{i,l,t} - \mathbf{A}^S_{i,l,t}\right)\right), \qquad \mathcal{T}_{l,t} : \mathbf{A}^H_{i,l,t} \mapsto \mathbf{A}^S_{i,l,t}, \tag{2}$$

where we apply L2 normalization via f_{norm}. The steering vector $\mathbf{v}_{l,t}$ points from the safe to the harmful activation mean, while $\mathcal{T}_{l,t}$ maps each harmful activation to its paired safe activation.

During inference, the corresponding signal is applied to each token activation by vector subtraction, yielding $\tilde{\mathbf{a}}^{\text{vec}}_{l,t}[n]$, or transport, yielding $\tilde{\mathbf{a}}^{\text{map}}_{l,t}[n]$:

$$\tilde{\mathbf{a}}^{\text{vec}}_{l,t}[n] = \mathbf{a}_{l,t}[n] - \lambda_{l,t,n}\mathbf{v}_{l,t}, \qquad \tilde{\mathbf{a}}^{\text{map}}_{l,t}[n] = \mathcal{T}_{l,t}(\mathbf{a}_{l,t}[n]), \tag{3}$$

where $\lambda_{l,t,n}$[5] scales the intervention by the token's alignment with the harmful direction, following CASteer. These interventions shift intermediate activations away from the harmful direction or toward safe representations, thereby guiding subsequent denoising to suppress harmful content.

3.2 Effect of Irrelevant Variation on Steering

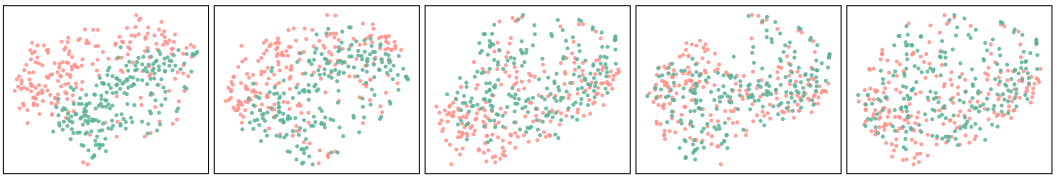
To test how harmful-irrelevant variation (i.e. nuisance) affects safety steering, we perturb the activations used for signal construction and measure both the activation contrast and downstream suppression. We measure harmful-safe separability[6] as a diagnostic of signal quality (Braun et al., 2025), alongside harmful rate to assess whether reduced separability accompanies weaker suppression. To simulate nuisance, we inject Gaussian noise into the activations used to construct steering vectors for sexual content, violence, and illegal activities, and vary its strength to probe this relationship.

Fig. 2 shows that harmful-irrelevant noise reduces activation separability and weakens safety steering across all three categories. As the noise scale increases from $\sigma = 0$ to 5, separability decreases from 0.37 to 0.14 for sexual content, from 0.41 to 0.14 for violence, and from 0.40 to 0.14 for illegal activities. Importantly, harmful rates rise from 21.0% to 38.3%, from 19.7% to 34.3%, and from 30.5% to 41.4%, respectively. These results show that harmful-irrelevant interference during signal construction can weaken safety-steering effectiveness. Experimental details and additional validation on SD-v3 are provided in App. A.

Although existing methods use paired harmful and safe prompts to construct steering signals, the resulting activation contrasts may still contain nuisance variation. In Secs. 3.3 and 3.4, we analyze how activation collection and spatial aggregation introduce such variation into existing pipelines.

[5] $\lambda_{l,t,n} = \max(\beta\langle \mathbf{v}_{l,t}, \mathbf{a}_{l,t}[n]\rangle, 0)$, where $\beta > 0$ controls suppression strength. No steering is applied when the inner product is non-positive.

[6] We use MMD to measure separability between harmful and safe activation distributions.



(a) Harmful Tokens Separability = 0.37 (b) +25% Background Separability = 0.24 (c) +50% Background Separability = 0.20 (d) +75% Background Separability = 0.18 (e) Global Pooling Separability = 0.17

Figure 3: Effect of background-token inclusion on harmful-safe activation separability for nudity at the Up-7 layer and timestep $t = 30$ on SD-v1.5. Red and green points denote harmful and safe activations, respectively.

3.3 Trajectory-Induced Confounding

Existing methods run denoising separately for harmful and safe prompts and collect activations from independent trajectories. At layer l and timestep t, let $\mathbf{h}^H_{i,l,t}$ and $\mathbf{h}^S_{i,l,t}$ denote the intermediate representations entering the target layer in the separate harmful and safe trajectories. The collected activations in Sec. 3.1 are

$$\mathbf{a}^H_{i,l,t} = \Phi^l(\mathbf{h}^H_{i,l,t}, \mathbf{E}^H_i), \qquad \mathbf{a}^S_{i,l,t} = \Phi^l(\mathbf{h}^S_{i,l,t}, \mathbf{E}^S_i), \tag{4}$$

where $\mathbf{E}^H_i$ and $\mathbf{E}^S_i$ are the paired text encodings, and Φ^l denotes the activation layer. Even with matched initial noise, independently evolved trajectories do not guarantee that the two input states differ only in the target harmful attribute. For example, "a bloody dog" and "a dog" could generate a black dog with blood and a white dog, respectively. Although the prompts differ only in the harmful attribute, the trajectories may also differ in coat color, a non-target property of the same subject. As the two trajectories evolve, earlier conditioning can change the intermediate states on which subsequent activations are computed. Specifically, the activation contrast can be decomposed into a text-condition contrast and a trajectory-state difference:

$$\Delta\mathbf{a} = \underbrace{\Phi(\mathbf{h}^H, \mathbf{E}^H) - \Phi(\mathbf{h}^H, \mathbf{E}^S)}_{\text{text-condition contrast}} + \underbrace{\Phi(\mathbf{h}^H, \mathbf{E}^S) - \Phi(\mathbf{h}^S, \mathbf{E}^S)}_{\text{trajectory-state difference}}. \tag{5}$$

The first term measures the activation change under different text conditions at the same input state; the second measures the change due to different input states under the same safe condition.

We term this mixing of local conditioning effects and trajectory-specific state differences *Trajectory-induced confounding*. As the latter can include non-target changes, the resulting steering signal may encode variations unrelated to the intended harmful content. Such variations can direct the intervention toward non-target attributes, weakening its specificity to the harmful concept.

3.4 Aggregation-Induced Dilution

Beyond activation collection, spatial aggregation introduces another potential source of interference. To construct steering signals, existing T2I methods aggregate image-token activations into a single representation for each sample. They average over all spatial tokens (Eq. 1), although harmful content typically occupies only part of the image. Pooling harmful-relevant activations with those from background regions and unrelated objects can dilute the harmful-concept contrast used for signal construction, potentially weakening the resulting signal's ability to suppress harmful content.

To test this effect, we retain a fixed set of harmful-relevant tokens and progressively add background tokens, expanding the aggregation support from the target region to the full image. Grounded-SAM (Ren et al., 2024) identifies the target region for this diagnostic. As shown in Fig. 3, increasing background-token inclusion produces greater distributional overlap, while separability decreases monotonically from 0.37 to 0.17 in this example.

We term this weakening of the harmful-concept contrast through spatial pooling *Aggregation-induced dilution*. These results indicate that including background tokens can reduce harmful-safe separability in the aggregated representations, weakening the concept contrast available for steering-signal construction. More analyses across harmful categories and timesteps are provided in App. B.

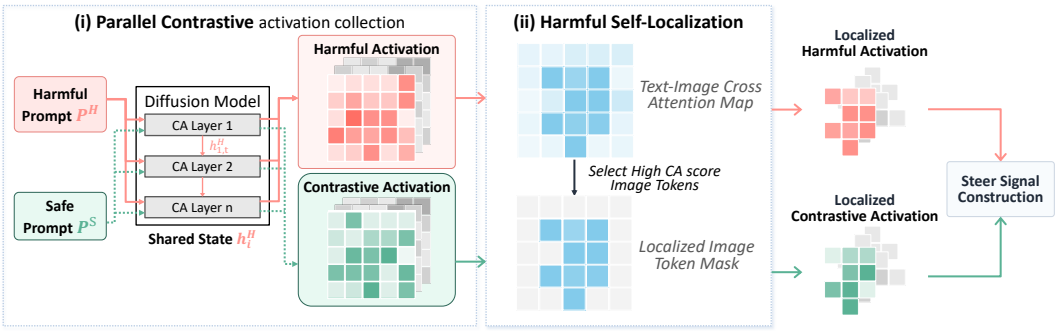


Figure 4: Framework of PaLo-steer signal construction.

4 Parallel and Localized Steering-Signal Construction

To address trajectory confounding and concept dilution and construct nuisance-reduced steering signals, we propose PaLo-Steer. It combines shared-state activation collection (Sec. 4.1) with harmful-relevant spatial selection (Sec. 4.2), then forms steering vectors or transport maps (Sec. 4.3).

4.1 Parallel Contrastive Activation Collection

To address trajectory confounding identified in Sec. 3.3, we introduce parallel contrastive activation collection, which compares how the same intermediate image state responds to harmful and safe text conditions. We obtain that state from a harmful-conditioned trajectory and evaluate a safe-conditioned reference locally, without evolving an independent safe trajectory. The resulting activation pair differs in text conditioning while sharing the image state.

Collecting the shared image state. For a target harmful concept c_H, we use the M prompt pairs in Sec. 3.1. Let $\mathbf{E}_i^H$ and $\mathbf{E}_i^S$ denote their harmful and safe text encodings. For each pair, we propagate only the harmful-conditioned branch and record its pre-cross-attention state $\mathbf{h}_{i,l,t}^H \in \mathbb{R}^{N\times d}$ at layer l and timestep t. Both conditions use this recorded state, giving shared image queries:

$$\mathbf{Q}_i^{H,l,t} = \mathbf{h}_{i,l,t}^H \mathbf{W}_Q^l. \tag{6}$$

Evaluating a condition-controlled activation pair. At each target layer, we reuse the recorded image state and change only the text encoding. Let CA^l denote cross-attention with image queries from $\mathbf{h}$ and text keys and values from $\mathbf{E}$. For the activation layer defined in Sec. 3.1, the output combines cross-attention with the residual input:

$$\Phi^l(\mathbf{h}, \mathbf{E}) = \mathbf{h} + \mathrm{CA}^l(\mathbf{h}, \mathbf{E}). \tag{7}$$

Evaluating this operation under the harmful and safe encodings yields the activation pair:

$$\mathbf{a}_{i,l,t}^H = \Phi^l(\mathbf{h}_{i,l,t}^H, \mathbf{E}_i^H), \qquad \mathbf{a}_{i,l,t}^{S|H} = \Phi^l(\mathbf{h}_{i,l,t}^H, \mathbf{E}_i^S), \tag{8}$$

where $S \mid H$ denotes a safe-conditioned response evaluated at the harmful branch's state. Only the harmful-conditioned output is passed to the next layer and ultimately used for the denoising update. The safe output is recorded for comparison, after which the harmful trajectory continues.

The harmful-safe activation difference instantiates the fixed-state term in Eq. 5: the shared residual cancels, leaving the response to the changed text condition. The activation pair therefore excludes differences between independently evolved states.

4.2 Spatially Localized Activation Selection

Parallel collection aligns the comparison state, but global pooling would still mix harmful-relevant responses with background and unrelated-object activations. To address concept dilution identified in Sec. 3.4, we introduce spatially localized activation selection, which restricts aggregation to harmful-relevant image tokens. One possible approach is to use external image-space localization to identify these tokens. However, such localization can be unreliable at early denoising stages, lacks layer-specific information, and requires additional decoding and model inference (App. C). We therefore use model-intrinsic cross-attention to locate harmful-relevant image tokens.

Scoring image tokens by harmful-concept relevance. Cross-attention weights indicate how strongly each image token attends to individual text tokens. Higher weights on harmful text tokens c_H therefore provide a cue for locating harmful-relevant image regions. We use these weights to assign a relevance score to each image token. Let $\mathbf{CA}_i^{H,l,t}$ denote the cross-attention weights averaged over attention heads under the harmful text condition, and let $\mathcal{C}_{H,i}$ contain the text-token positions corresponding to c_H in prompt p_i^H. Averaging attention over these positions gives a harmful relevance score for each image token:

$$s_{i,n}^{l,t} = \frac{1}{|\mathcal{C}_{H,i}|} \sum_{j \in \mathcal{C}_{H,i}} \mathbf{CA}_{i,n,j}^{H,l,t}, \qquad n = 1, \ldots, N. \tag{9}$$

Selecting a shared harmful-relevant support. To exclude low-relevance activations, we select tokens by their rank within each attention map. A quantile threshold accommodates attention magnitudes that vary across samples, layers, and timesteps. The threshold and token set for

$$\theta_{i,q}^{l,t} = \mathrm{Quantile}_q\left(\{s_{i,n}^{l,t}\}_{n=1}^{N}\right), \qquad \mathcal{S}_{i,q}^{l,t} = \left\{n \mid s_{i,n}^{l,t} \geq \theta_{i,q}^{l,t}\right\}. \tag{10}$$

The quantile controls how concentrated the support is. We derive the support from attention to the harmful expression and apply it to both activations in the pair. Selecting separate supports could compare responses at different positions and introduce an additional spatial difference into the contrast. Shared support preserves the paired comparison while excluding low-relevance activations.

For each prompt pair, we average the harmful and safe activations only over the harmful-relevant positions:

$$\mathbf{A}_{i,l,t,q}^{r} = \frac{1}{|\mathcal{S}_{i,q}^{l,t}|} \sum_{n \in \mathcal{S}_{i,q}^{l,t}} \mathbf{a}_{i,l,t}^{r}[n], \qquad r \in \{H, S \mid H\}, \tag{11}$$

where $\mathbf{A}_{i,l,t,q}^{r}$ is the localized sample-wise representation under condition r.

4.3 Nuisance-Reduced Steering-Signal Construction

We construct steering signals from the shared-state activations collected in Sec. 4.1 and aggregated over harmful-relevant positions in Sec. 4.2. Substituting the resulting representations into Eq. 2 yields

$$\mathbf{v}_q^{l,t} = f_{\mathrm{norm}}\left(\frac{1}{M}\sum_{i=1}^{M}\left(\mathbf{A}_{i,l,t,q}^{H} - \mathbf{A}_{i,l,t,q}^{S|H}\right)\right), \qquad \mathcal{T}_q^{l,t} : \mathbf{A}_{i,l,t,q}^{H} \mapsto \mathbf{A}_{i,l,t,q}^{S|H}, \tag{12}$$

where $\mathbf{v}_q^{l,t}$ is the L2-normalized mean of the localized pairwise differences, and $\mathcal{T}_q^{l,t}$ is the AcT transport map fitted from harmful representations to their paired safe references.

Both signals benefit from the two complementary controls: shared-state comparison removes the contribution of mismatched trajectory states, while localized aggregation limits harmful-irrelevant spatial contributions. Together, these controls make the constructed signals more specific to the harmful-concept contrast, supporting targeted harmful-content suppression.

The shared-state comparison and spatial selection are performed during offline signal construction. At inference time, the resulting vectors or maps follow the original token-wise intervention in Eq. 3: vector-based steering subtracts the harmful direction, while AcT maps activations toward their safe references, thereby reducing harmful-content generation. The complete algorithm is in App. D.

5 Experiments

In this section, we conduct experiments to evaluate the effectiveness of PaLo-Steer. We introduce the setup (Sec. 5.1), assess effectiveness (Sec. 5.2), ablate the two components (Sec. 5.3), examine hyperparameter sensitivity (Sec. 5.4), and analyze the steering mechanism (Sec. 5.5).

5.1 Experimental Setup

Datasets. We evaluate harmful-content suppression on I2P(Schramowski et al., 2023), T2VSafetyBench(Miao et al., 2024), Unsafe Diffusion(Qu et al., 2023), and CoProV2 (Liu et al., 2025). To measure the effect of suppression on benign generation, we also sample 1,000 captions from COCO-30k (Lin et al., 2014).

Table 1: Comparison of T2I safeguards on four harmful benchmarks and COCO image quality. Brackets show changes from SDv1.5. Model-modifying methods are shown in gray, and best parameter-free results are **bold**.

Method	PaLo	Harmful Rate % ↓				FID ↓	CLIP ↑
		I2P	T2VSafetyBench	UD	CoProv2	COCO	
SD-v1.5 (Rombach et al., 2022)	-	42.2	58.3	52.3	28.2	-	31.43
ESD (Gandikota et al., 2023)	×	42.0 (-0.2)	57.4 (-0.9)	50.6 (-1.7)	28.1 (-0.1)	38.15	31.35
UCE (Gandikota et al., 2024)	×	26.7 (-15.5)	28.2 (-30.1)	33.0 (-19.3)	19.4 (-8.8)	77.41	29.12
RECE (Gong et al., 2024)	×	21.5 (-20.7)	18.9 (-39.4)	22.3 (-30.0)	8.6 (-19.6)	67.35	27.67
SafetyDPO (Liu et al., 2025)	×	13.7 (-28.5)	24.0 (-34.3)	16.7 (-35.6)	4.2 (-24.0)	49.64	30.61
SLD (Schramowski et al., 2023)	×	17.1 (-25.1)	34.1 (-24.2)	22.0 (-30.3)	9.2 (-19.0)	44.15	30.05
SAFREE (Yoon et al., 2025)	×	20.0 (-22.2)	41.5 (-16.8)	24.2 (-28.1)	14.2 (-14.0)	43.78	30.53
AcT (Rodriguez et al., 2024)	×	35.3 (-6.9)	52.9 (-5.4)	43.6 (-8.7)	26.1 (-2.1)	43.29	31.72
PaLo+AcT (ours)	✓	31.7 (-10.5)	51.8 (-6.5)	39.9 (-12.4)	22.4 (-5.8)	42.11	31.51
CASteer (Gaintseva et al., 2026a)	×	24.2 (-18.0)	34.4 (-23.9)	25.7 (-26.6)	12.9 (-15.3)	37.77	31.59
PaLo+CASteer (ours)	✓	**15.1 (-27.1)**	**22.4 (-35.9)**	**18.8 (-33.5)**	**9.0 (-19.2)**	37.98	31.34



Figure 5: Qualitative comparison of T2I safety methods on harmful prompts (left) and benign prompts (right).

Baselines. To assess effectiveness of PaLo-Steer, we integrate PaLo-Steer with CASteer (Gaintseva et al., 2026a) and AcT (Rodriguez et al., 2024), denoted as PaLo+CASteer and PaLo+AcT, and compare each with its original method. We also compare against model-editing methods ESD (Gandikota et al., 2023), UCE (Gandikota et al., 2024), and RECE (Gong et al., 2024); alignment-based SafetyDPO (Liu et al., 2025); and guidance-based SLD (Schramowski et al., 2023) and SAFREE (Yoon et al., 2025).

Evaluation Metrics. Following prior work Schramowski et al. (2023), we report the harmful rate: the percentage of generated images flagged by Q16 (Schramowski et al., 2022) and NudeNet (notAI tech, 2024). For image quality on benign usage, CLIP Score (Hessel et al., 2021) measures semantic alignment and FID (Heusel et al., 2017) measures image fidelity.

Implementation Details. We use SD-v1.5 and replace the signals applied at each denoising step and cross-attention layer, following CASteer (Gaintseva et al., 2026a). We set the localization quantile to $q = 0.9$. Detailed settings and multi-category handling are provided in App. E.

5.2 Effectiveness of PaLo-Steer

Harmful-Content Suppression. PaLo-Steer improves both vector-based CASteer and transport-map-based AcT on all four safety benchmarks (Tab. 1). PaLo+CASteer reduces harmful rates from 24.2% to 15.1% on I2P, from 34.4% to 22.4% on T2VSafetyBench, from 25.7% to 18.8% on Unsafe Diffusion, and from 12.9% to 9.0% on CoProV2, giving the lowest rates among parameter-free methods. PaLo+AcT improves AcT on every benchmark, with reductions of up to 3.7%. The reductions occur with both vector subtraction and activation transport.

Benign Generation Quality. We evaluate generation quality on COCO to assess the impact of safety steering on benign usage. PaLo+CASteer changes CLIP score from 31.59 to 31.34 and FID from 37.77 to 37.98 relative to CASteer. PaLo+AcT produces a similar CLIP decrease (31.72 to 31.51) and improves FID from 43.29 to 42.11. Both methods lower harmful rates with modest changes in benign-generation metrics.

Generalization and Adversarial Robustness. We test whether PaLo-Steer's effectiveness extends to a DiT architecture and adversarially phrased prompts. On DiT-based SD-v3 (Esser et al., 2024), PaLo+CASteer lowers CASteer's harmful rates across all three evaluated safety benchmarks by up to 9.9%, while improving COCO FID with a small CLIP decrease (Tab. 6). On SD-v1.5, it also reduces harmful rates across three adversarial benchmarks by up to 22.4% (Tsai et al., 2024; Yang et al., 2024; Zhang et al., 2026), as shown in Tab. 7. The improvements thus extend to SD-v3 and the evaluated adversarial prompts. Detailed results are provided in Apps. G.1 and G.2.

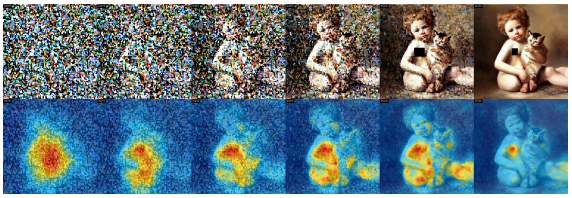
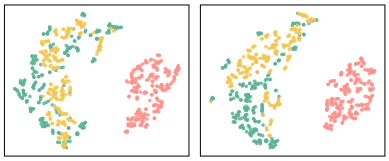


Figure 6: Harmful-token cross-attention localizes harmful content across denoising steps.



Sexual Violence

Figure 7: t-SNE of safe, globally pooled harmful, and PaLo harmful activations.

Qualitative Results. Fig. 5 illustrates the safety and quality changes in generated images. In the harmful examples, PaLo-Steer variants suppress unsafe visual elements more effectively than their steering baselines; the benign examples retain prompt semantics and image coherence.

5.3 Ablation Study

We analyze the contribution of parallel collection and localized aggregation by adding them to CASteer separately and together. In Tab. 2, each component alone lowers harmful rates across all four reported categories, supporting the effectiveness of shared-state comparison and harmful-relevant token aggregation. Combining the two components yields the lowest harmful rates, reducing the category average from 23.9% to 13.7%. The additional gains are consistent with the complementary roles of parallel collection and localized aggregation in reducing trajectory confounding and concept dilution, respectively. These results support the effectiveness of jointly addressing both sources of interference in steering-vector construction. We compare localization with random selection in App. F.3.

Table 2: Ablation on PaLo-Steer. $(\times, \times)$ denotes CASteer and $(\checkmark, \checkmark)$ full PaLo+CASteer.

Pa.	Lo.	hate	violence	sex	illegal
×	×	33.8	25.1	20.0	16.6
✓	×	27.7	18.8	13.9	12.8
×	✓	20.8	16.3	11.8	12.4
✓	✓	**17.7**	**14.8**	**11.2**	**11.0**

5.4 Hyperparameter Analysis

We evaluate PaLo-Steer under different settings of steering strength β and localization quantile q. Across $\beta \in \{1.5, 2.0, 2.5\}$, PaLo+CASteer lowers harmful rates relative to CASteer by 6.9-9.4% on UD and 2.7-4.5% on CoProV2 (Tab. 14). Varying $q \in \{0.6, 0.8, 0.9, 0.95\}$ at $\beta = 2.0$ yields reductions of 5.9-6.9% and 3.9-4.4%, respectively (Tab. 15). These results show that PaLo-Steer further reduces harmful rates across the tested settings. Detailed results are provided in App. G.5.

5.5 Mechanism Analysis of PaLo-Steer

We analyze the mechanism of the spatial-selection component through localization visualizations and harmful-safe separation. In Fig. 6, high-score regions closely track harmful visual content, while their spatial extent changes across denoising timesteps as the intermediate image states evolve. This correspondence indicates that cross-attention provides spatial support that adapts to the evolving harmful content. In Fig. 7, globally pooled harmful activations overlap with safe activations, whereas PaLo activations form more distinct groups. Together with the quantitative separation analysis in App. F, these results show that aggregating only harmful-relevant token activations substantially improves harmful-safe separation. This improvement is consistent with reduced inclusion of harmful-irrelevant information, supporting the effectiveness of our spatial-selection mechanism in mitigating concept dilution. More analysis are in App. F.

Additional analysis. Due to page limitations, further analyses of robustness to construction and inference settings (App. G.4), cross-model transferability (App. G.6), efficiency analysis(App. H), and more qualitative results (App. I) are provided in the appendix.

6 Conclusion

Activation steering enables training-free suppression of harmful content in T2I diffusion models, but trajectory confounding and concept dilution can weaken its steering signals. We propose PaLo-Steer to address these issues through parallel contrastive activation collection and spatially localized activation selection, comparing harmful and safe conditions at shared intermediate states and aggregating only harmful-relevant token activations. Across safety benchmarks, PaLo-Steer improves both vector- and transport-map-based steering, reducing harmful rates by up to 12.0% over the corresponding baselines while preserving benign generation quality.

AI USE STATEMENT

We used generative AI tools to assist with writing and language polishing, and with parts of the experimental scripts. We did not use generative AI tools for research ideation, algorithm design, or core code implementation. We have reviewed all AI-assisted work and take responsibility for the final content of this paper and the associated code.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the signal-construction procedure in Secs. 4.1 to 4.3 and provide the complete construction and application algorithm in App. D. The datasets, evaluation criteria, construction-prompt templates, harmful keywords, and implementation settings are documented in Sec. 5.1 and App. E. Apps. G.4 and G.5 further report results across construction settings, random seeds, and hyperparameters to characterize the sensitivity of the reported improvements.

REFERENCES

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (un) reliability of steering vectors in language models. *arXiv preprint arXiv:2505.22637*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Tatiana Gaintseva, Andreea-Maria Oncescu, Chengcheng Ma, Ziquan Liu, Martin Benning, Gregory Slabaugh, Jiankang Deng, and Ismail Elezi. Casteer: Cross-attention steering for controllable concept erasure. In *International Conference on Learning Representations*, volume 2026, pp. 40098–40110, 2026a.

Tatiana Gaintseva, Andrew Stepanov, Ziquan Liu, Martin Benning, Gregory Slabaugh, Jiankang Deng, and Ismail Elezi. Midsteer: Optimal affine framework for steering generative models, 2026b. URL https://arxiv.org/abs/2605.05220.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2426–2436, 2023.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5099–5108. IEEE, 2024.

Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025.

Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In *European Conference on Computer Vision*, pp. 73–88. Springer, 2024.

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 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Bruce W Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *International conference on learning representations*, volume 2025, pp. 90960–90985, 2025.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023.

Lijun Li, Zhelun Shi, Xuhao Hu, Bowen Dong, Yiran Qin, Xihui Liu, Lu Sheng, and Jing Shao. T2isafety: Benchmark for assessing fairness, toxicity, and privacy in image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13381–13392. IEEE, 2025.

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.

Runtao Liu, I Chieh Chen, Jindong Gu, Jipeng Zhang, Renjie Pi, Qifeng Chen, Philip Torr, Ashkan Khakzar, and Fabio Pizzati. Alignguard: Scalable safety alignment for text-to-image generation, 2025. URL https://arxiv.org/abs/2412.10493.

Yibo Miao, Yifan Zhu, Lijia Yu, Jun Zhu, Xiao-Shan Gao, and Yinpeng Dong. T2vsafetybench: Evaluating the safety of text-to-video generative models. *Advances in Neural Information Processing Systems*, 37:63858–63872, 2024.

notAI tech. Nudenet, 2024. URL https://github.com/notAI-tech/NudeNet.

Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, pp. 3403–3417, 2023.

Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Marco Cuturi, and Xavier Suau. Controlling language and diffusion models by transporting activations, 2024. URL https://arxiv.org/abs/2410.23054.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 1350–1361, 2022.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22522–22531. IEEE, 2023.

Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? In *International Conference on Learning Representations*, volume 2024, pp. 41543–41554, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Pengyu Wang, Dong Zhang, Linyang Li, Chenkun Tan, Xinghao Wang, Mozhi Zhang, Ke Ren, Botian Jiang, and Xipeng Qiu. Inferaligner: Inference-time alignment for harmlessness through cross-model guidance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10460–10479, 2024.

Yongli Xiang, Ziming Hong, Zhaoqing Wang, Xiangyu Zhao, Bo Han, and Tongliang Liu. When safety collides: Resolving multi-category harmful conflicts in text-to-image diffusion via adaptive safety guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14746–14755, June 2026.

Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7737–7746, 2024.

Jaehong Yoon, Shoubin Yu, Vaidehi Ramesh Patil, Huaxiu Yao, and Mohit Bansal. Safree: Training-free and adaptive guard for safe text-to-image and video generation. In *International Conference on Learning Representations*, volume 2025, pp. 56439–56465, 2025.

Adeel Yousaf, Soumik Ghosh, James Beetham, Amrit Singh Bedi, and Mubarak Shah. The illusion of high utility in safety alignment of text-to-image diffusion models, 2026. URL https://arxiv.org/abs/2607.00402.

Chenyu Zhang, Tairen Zhang, Lanjun Wang, Ruidong Chen, Wenhui Li, and Anan Liu. T2i-riskyprompt: A benchmark for safety evaluation, attack, and defense on text-to-image model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 36039–36047, 2026.

Overview of the Appendices

- In App. A, we examine construction noise and steering effectiveness across settings.
- In App. B, we analyze concept dilution across harmful categories and denoising timesteps.
- In App. C, we discuss the limitations of external image-space localization for steering-signal construction.
- In App. D, we present the complete PaLo-Steer algorithm.
- In App. E, we document implementation, evaluation, keyword, and multi-category settings.
- In App. F, we further analyze cross-attention localization and localized aggregation.
- In App. G, we report supplementary generalization and cross-version transfer results, category-wise evaluations, robustness to adversarial prompts and construction and inference settings, and hyperparameter analyses.
- In App. H, we analyze the construction and inference efficiency of PaLo-Steer.
- In App. I, we provide additional qualitative comparisons across harmful categories.

A Effect of Construction Noise on Steering

A.1 Controlled Noise-Injection Details

To test how harmful-irrelevant noise in signal construction affects steering effectiveness, we conduct the controlled noise-injection experiment in Sec. 3.2. We perturb collected activations and reconstruct steering vectors at different noise scales, then evaluate their ability to suppress harmful generation.

Specifically, at each target layer-timestep site (l, t), we add independent noise to the paired harmful and safe activations:

$$\widetilde{\mathbf{A}}_i^H(\sigma) = \mathbf{A}_i^H + \sigma\boldsymbol{\eta}_i^H, \qquad \widetilde{\mathbf{A}}_i^S(\sigma) = \mathbf{A}_i^S + \sigma\boldsymbol{\eta}_i^S, \tag{13}$$

where σ controls noise magnitude and $\boldsymbol{\eta}_i^H, \boldsymbol{\eta}_i^S$ are sampled from the same zero-mean Gaussian distribution, with diagonal covariance estimated from pooled clean activations. The injected noise therefore carries no harmful-safe label information. We reconstruct the steering vector as

$$\mathbf{v}(\sigma) = f_{\text{norm}}\left(\frac{1}{M}\sum_{i=1}^{M}\left(\widetilde{\mathbf{A}}_i^H(\sigma) - \widetilde{\mathbf{A}}_i^S(\sigma)\right)\right), \tag{14}$$

where M is the number of construction pairs and f_{norm} performs L2 normalization, so noise changes the estimated direction rather than the vector norm.

We use vector-based steering on SD-v1.5 with $\sigma \in \{0, 1.0, 2.0, 5.0\}$ and evaluate T2VSafetyBench prompts (Miao et al., 2024): 300 images each for sexual content and violence, and 249 for illegal activities. Harmful rates are measured with NudeNet (notAI tech, 2024) and Q16 (Schramowski et al., 2022). We measure $\text{separability}(\sigma)$ using MMD with a kernel bandwidth calibrated on clean activations. We run our experiments three times and report mean separability and median harmful rate, with error bars showing the harmful-rate range.

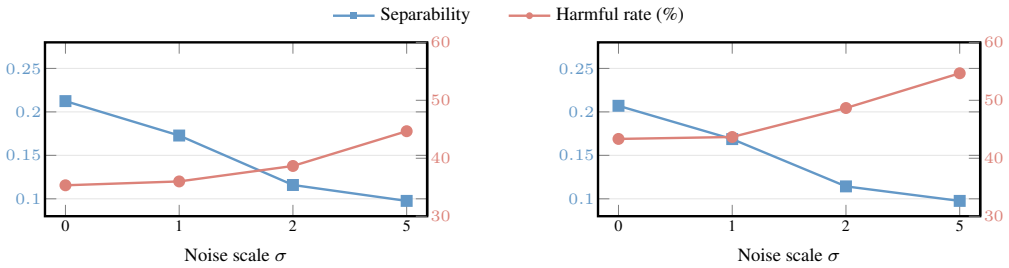
By varying only the injected noise while fixing construction and evaluation prompts, steering sites and strength, and latent seeds, this experiment tests the effect of construction noise on steering effectiveness, alongside changes in harmful-safe activation separability.

A.2 Noise-Injection Analysis on SD-v3

To test whether construction noise also weakens steering in a DiT architecture, we extend the protocol in App. A.1 to SD-v3 for sexual content and violence. Fig. 8 reports harmful-safe separability and harmful rates for both categories.

As noise increases from $\sigma = 0$ to $\sigma = 5$, separability decreases from 0.21 to 0.10 for both categories. Over the same range, harmful rates rise from 35.3% to 44.7% for sexual content and from 43.3% to 54.7% for violence, with both metrics changing monotonically across the tested scales.

These results extend the SD-v1.5 finding to SD-v3: noise introduced during signal construction weakens downstream safety steering in both evaluated categories. The accompanying decline in separability is consistent with a degraded harmful-safe contrast, supporting the motivation to reduce harmful-irrelevant variation when constructing steering signals.



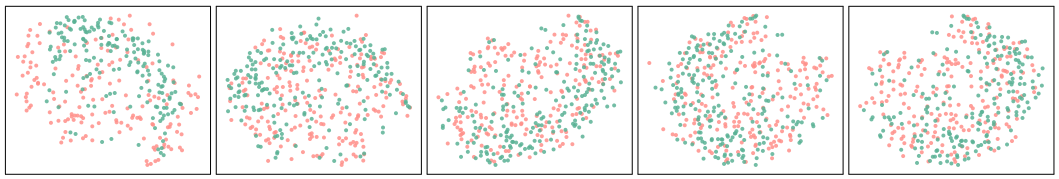
(a) Sexual (b) Violence

Figure 8: Effect of construction noise on separability and harmful rate on SD-v3.

B Aggregation-Induced Dilution Analysis

To test whether aggregating harmful-irrelevant regions dilutes the harmful-safe activation contrast across concepts and denoising timesteps, we use Grounded-SAM Ren et al. (2024) to partition final generated images into harmful-relevant and background regions, then map them to the 64×64 activation-token grid. We compare harmful-safe separability when aggregating only harmful-relevant activations with that obtained by progressively adding background activations while keeping the harmful-relevant region fixed[7].

Across Harmful Categories. To test whether concept dilution is category-specific, we repeat the token-inclusion analysis for violence, using blood as the localized harmful content. In Fig. 9, progressively adding background tokens decreases separability from 0.33 with harmful-relevant tokens alone to 0.17 under global pooling. This matches the nudity trend in Fig. 3.



(a) Harmful Tokens Separability $= 0.33$ (b) +25% Background Separability $= 0.19$ (c) +50% Background Separability $= 0.18$ (d) +75% Background Separability $= 0.17$ (e) Global Pooling Separability $= 0.17$

Figure 9: Background-token inclusion for violence at $t = 30$ (SD-v1.5).

Across Denoising Timesteps. To test whether concept dilution persists across the denoising trajectory, we repeat the diagnostic at $t = 50$ for nudity and violence. As shown in Fig. 10, increasing background-token inclusion reduces separability from 0.24 to 0.15 for nudity and from 0.44 to 0.18 for violence.

Across both tested concepts and timesteps, adding harmful-irrelevant tokens consistently reduces harmful-safe separability, supporting concept dilution as a limitation of global aggregation. These results motivate restricting aggregation to harmful-relevant regions when constructing steering signals.

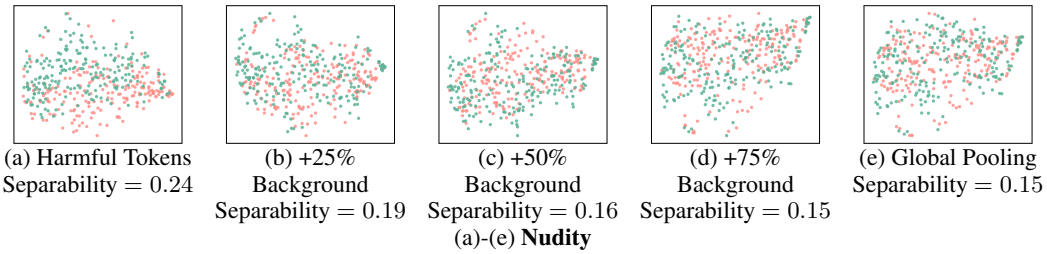
C Limitations of External Image-Space Localizers

External image-space localizers, such as Grounded-SAM, face the following limitations when used to localize harmful-relevant activations during steering-signal construction.

Visibility mismatch. External localizers (e.g., Grounded-SAM) require sufficiently formed visual evidence, whereas early decoded states remain noisy and semantically incomplete. These masks can miss harmful regions or cover nearly the entire image, making them unreliable for constructing steering signals at early denoising steps. As a representative example, Fig. 11 shows that Grounded-SAM masks for *blood* cover over 99.0% of the image through iteration $k = 30$ and become spatially meaningful only around $k = 40$.

Representation mismatch. An image-space localizer provides a pixel mask for each decoded image, which must be resized for use at different layers. Resizing changes the mask resolution but cannot adapt it to how harmful content is represented within each layer. Using these masks in

[7]Following CASteer (Gaintseva et al., 2026a), harmful/safe activations are collected along independent trajectories.



(a) Harmful Tokens Separability = 0.24 (b) +25% Background Separability = 0.19 (c) +50% Background Separability = 0.16 (d) +75% Background Separability = 0.15 (e) Global Pooling Separability = 0.15

(a)-(e) **Nudity**

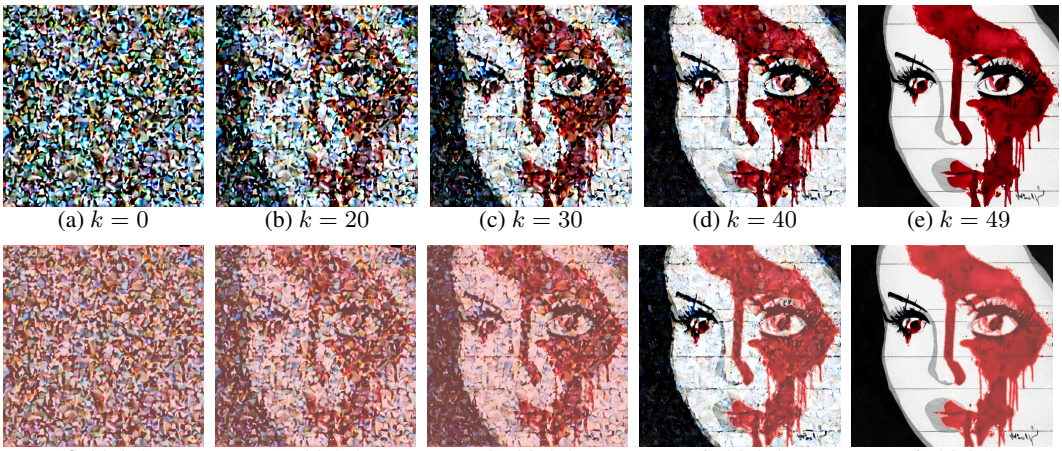
(f) Harmful Tokens Separability = 0.44 (g) +25% Background Separability = 0.23 (h) +50% Background Separability = 0.20 (i) +75% Background Separability = 0.19 (j) Global Pooling Separability = 0.18

(f)-(j) **Violence (Blood)**

Figure 10: Background-token inclusion at $t = 50$ (SD-v1.5, Up-7).

PaLo-Steer could therefore miss harmful-relevant activations or include harmful-irrelevant ones, weakening the localized contrast used to construct each layer's steering signal.

Efficiency mismatch. External localization requires repeated latent decoding and additional model inference for every sample and timestep. Cross-attention provides layer- and timestep-specific scores from intermediate activations, avoiding the repeated decoding and external inference required by image-space localization.



(a) $k = 0$ (b) $k = 20$ (c) $k = 30$ (d) $k = 40$ (e) $k = 49$

(f) 99.0% (g) 99.2% (h) 99.4% (i) 30.8% (j) 30.9%

Figure 11: Grounded-SAM localization of *blood* across denoising iterations k. Top: decoded images. Bottom: mask overlays and coverage. Early masks cover nearly the entire image.

These limitations motivate us to seek model-intrinsic information for localizing harmful-relevant activations. In Sec. 4.2, we introduce cross-attention-based localization, which provides layer- and timestep-specific spatial scores throughout denoising without additional image decoding or external-model inference.

D Detailed Algorithm of PaLo-Steer

Alg. 1 separates PaLo-Steer into offline signal construction and inference-time application. The construction stage replaces conventional activation collection and aggregation, whereas the application stage retains the original intervention of the underlying vector- or transport-map-based steering method. The algorithm describes construction for one harmful concept; keyword-bank and multi-category settings are given in App. E.

Algorithm 1 PaLo-Steer: signal construction and application

Input: harmful-safe prompt pairs $\{(p_i^H, p_i^S)\}_{i=1}^M$; target sites $\mathcal{I} = \{(l,t)\}$; quantile q; inference prompt p; vector strength β
Output: signals $\{\mathbf{v}_q^{l,t}\}$ or $\{\mathcal{T}_q^{l,t}\}$ for $(l,t) \in \mathcal{I}$; generated image x

Steering-signal construction (offline)
for $i = 1, \ldots, M$ **do**
  Encode p_i^H and p_i^S
  Sample an initial latent for the harmful-conditioned trajectory
  for all denoising timesteps t in sampling order **do**
    for all target layers l with $(l,t) \in \mathcal{I}$, in network execution order **do**
      Advance the harmful-conditioned network to layer l, obtaining $\mathbf{h}_{i,l,t}^H$
      Compute $\mathbf{a}_{i,l,t}^H$ and $\mathbf{a}_{i,l,t}^{S|H}$ from this shared input (Eq. 8)
      Select support $\mathcal{S}_{i,q}^{l,t}$ from harmful-conditioned attention (Eqs. 9 and 10)
      Pool both activations over $\mathcal{S}_{i,q}^{l,t}$ and store $\mathbf{A}_{i,l,t,q}^H, \mathbf{A}_{i,l,t,q}^{S|H}$ (Eq. 11)
      Propagate only $\mathbf{a}_{i,l,t}^H$; do not propagate the safe reference
    end for
    Complete the harmful-conditioned network evaluation and perform one scheduler update
  end for
end for
for all $(l,t) \in \mathcal{I}$ **do**
  Construct $\mathbf{v}_q^{l,t}$ or fit $\mathcal{T}_q^{l,t}$ from all M localized pairs (Eq. 12)
end for
Steering-signal application (inference)
Encode p and sample an initial latent
for all denoising timesteps t in sampling order **do**
  for all target layers l with $(l,t) \in \mathcal{I}$, in network execution order **do**
    Advance the prompt-conditioned network to layer l and compute $\mathbf{a}_{l,t}$
    Apply $\mathbf{v}_q^{l,t}$ or $\mathcal{T}_q^{l,t}$ to each token (Eq. 3)
    Propagate the modified activations to subsequent layers
  end for
  Complete the network evaluation and perform one scheduler update
end for
Decode the final latent to obtain x
return the constructed steering signals and x

E Experimental Details

Evaluation Datasets. We use 4,703 prompts from I2P (Schramowski et al., 2023), 3,443 prompts from seven relevant categories of T2VSafetyBench (Miao et al., 2024), 904 user-contributed prompts from Unsafe Diffusion (Qu et al., 2023), and 1,000 randomly sampled prompts from CoProV2 (Liu et al., 2025) to evaluate harmful-content suppression. For benign generation quality, we use 1,000 captions sampled from COCO-30k (Lin et al., 2014).

Implementation Details. Our main experiments use SD-v1.5 with 50 denoising steps and the default pipeline guidance scale of 7.5. Steering signals are constructed from 210 prompts using CASteer's templates with harmful keywords from SLD (App. E). To capture the activation contrast at each stage of denoising, we construct a signal for every cross-attention layer-timestep site (l,t) with $q = 0.9$ and apply it to that site's output activations during inference. We use official baseline implementations with CASteer strength $\beta = 2.0$, Linear AcT, and SLD-Strong.

Default Construction Prompts. Following the notation in Sec. 3.1, each shared benign context b_i concatenates a person phrase and a scene phrase. The default configuration uses CASteer's 15 person phrases and 14 scene phrases (Tab. 3) (Gaintseva et al., 2026a). Enumerating all combinations yields $M = 15 \times 14 = 210$ contexts, which form the paired prompts $p_i^H = b_i \oplus c_H$ and $p_i^S = b_i \oplus c_S$. An empty scene phrase adds no scene description.

Table 3: CASteer's default person and scene phrases used to construct the shared benign contexts b_i. The symbol $\varnothing$ denotes the empty string.

Phrase type	Entries
Person	a girl; a boy; two men; two women; two people; a man; a woman; an old man; an old woman; boys; girls; men; women; group of people; a human
Scene	$\varnothing$; gloomy image; zoomed in; talking; on the street; in a strange pose; realism; colorful background; on a beach; playing guitar; enjoying nature; smiling; in a futuristic spaceship; with kittens

Safety Evaluation Protocol. Following prior work Xiang et al. (2026); Schramowski et al. (2023), we assess each generated image using Q16 (Schramowski et al., 2022) and NudeNet (notAI tech, 2024). NudeNet flags an image when the confidence of any detected exposed-content category exceeds 0.5, while Q16 covers a broader range of unsafe imagery, including violent, sexual, and graphic content. We classify an image as harmful if either detector flags it, and report the percentage of harmful images as the harmful rate.

Harmful-Keyword Settings. We follow the harmful keyword set introduced by SLD (Schramowski et al., 2023). It contains 21 harmful keywords grouped into seven safety categories, as listed in Tab. 4. Multi-word entries are treated as individual expressions and subsequently mapped to their corresponding tokenizer tokens.

Table 4: SLD harmful keyword set used as the expression bank. Tokenizing each entry gives the harmful text-token set for parallel collection and spatially localized selection.

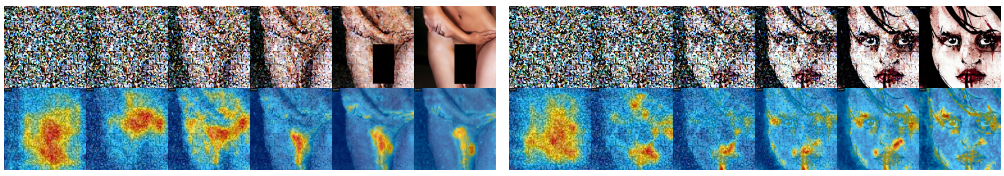
Category	Harmful keywords
Hate	hate
Harassment	harassment
Violence	violence, blood, brutality, cruelty
Self-Harm	self-harm, suicide
Sexual	sexual, nudity, bodily fluids, obscene gestures
Shocking Images	suffering, humiliation, harm
Illegal Activity	illegal activity, drug use, theft, vandalism, weapons, child abuse

Multi-Category Harmful Content Handling. Following the original CASteer setting (Gaintseva et al., 2026a), we first construct a steering vector independently for each harmful category using its corresponding expression set, and then average the category-specific vectors to obtain a single multi-category steering direction for inference. For transport-map-based methods such as AcT, where vector-level composition is not directly applicable, we follow existing safeguard practice by combining harmful categories in the prompt pairs and fitting one multi-category transport map.

F Additional Mechanism Analysis

F.1 Cross-Attention Localization

To examine whether cross-attention provides harmful-relevant spatial cues during denoising, Fig. 12 visualizes sexual- and violence-related attention alongside intermediate images. As image structure emerges, high-score regions concentrate around exposed body regions in the sexual example and blood and injury cues in the violence example. These maps provide spatial scores that adapt to the evolving content without an external image-space localizer.

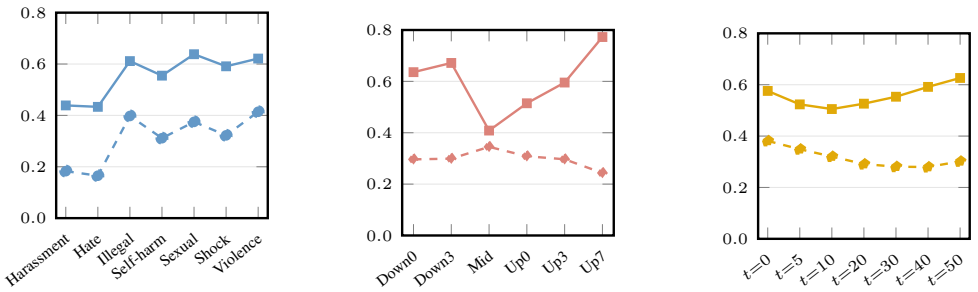


(a) Sexual (b) Violence

Figure 12: Harmful-token cross-attention tracks harmful regions across denoising steps.

F.2 Localized Aggregation

To test whether localized aggregation reduces concept dilution, we compare harmful-safe separability across concepts, layers, and denoising timesteps[8]. In Fig. 13, localized aggregation yields higher separability than global pooling in every tested setting, with concept-wise values of 0.43-0.64 versus 0.17-0.41. The t-SNE projections in Fig. 14 likewise show more distinct harmful and safe groups. Together, these results support restricting aggregation to harmful-relevant activations to reduce concept dilution.



(a) Across harmful concepts (b) Across model layers (c) Across denoising timesteps

Figure 13: Harmful-safe activation separability across harmful concepts, model layers, and denoising timesteps. In every panel, dashed curves denote Global Pooling, while solid curves denote PaLo with $q = 0.9$.

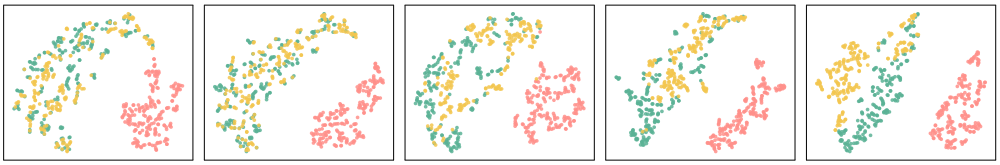


Figure 14: Joint t-SNE projections at Up-7 on SD-v1.5 ($t = 50$, except Illegal Activity at $t = 30$). Green, yellow, and red denote safe, globally pooled harmful, and PaLo harmful activations, respectively.

F.3 Effect of Token Selection

To assess whether harmful-relevant selection matters beyond reducing token count, we compare global pooling (baseline), random selection, and attention-guided selection (localized) on UD, all using independent trajectories (Tab. 5). Random and localized selection use the same token count and share selected positions within each harmful-safe pair. The remaining settings follow Sec. 5.1.

Table 5: Token-selection comparison on UD using independent trajectories.

	CASteer	CASteer + random	CASteer + localized
Harmful rate (%) ↓	25.7	25.4 (-0.3)	21.6 (-4.1)

Random selection yields a harmful rate of 25.4%, close to the 25.7% baseline, whereas localized selection reduces it to 21.6%. This comparison suggests that the safety improvement comes from localizing harmful-relevant activations rather than randomly reducing the token count.

G Additional Experiments

G.1 Generalization to DiT-Based SD-v3

To test whether nuisance-reduced construction also benefits a DiT architecture, we compare CASteer and PaLo+CASteer on SD-v3 across three safety benchmarks and benign COCO prompts. We use SD-v3 Medium with the default pipeline guidance scale of 7.0. Both methods use $\beta = 2.0$; PaLo uses $q = 0.9$.

[8]Harmful and safe activations are collected along independent denoising trajectories.

Table 6: Safety performance and image quality on SD-v3. Parentheses show absolute harmful-rate reductions from the SD-v3 baseline, the best safety results are in **bold**.

Method	Harmful Rate % ↓			FID ↓	CLIP ↑
	CoProv2	UD	MMA	COCO	
SD-v3	26.1	38.5	53.2	-	31.89
CASteer	22.5 (-3.6)	36.7 (-1.8)	30.8 (-22.4)	43.09	31.72
PaLo+CASteer	**19.4 (-6.7)**	**31.5 (-7.0)**	**20.9 (-32.3)**	37.33	31.50

Tab. 6 shows that PaLo+CASteer reduces CASteer's harmful rate by 3.1% on CoProV2, 5.2% on Unsafe Diffusion, and 9.9% on MMA. On COCO, FID improves from 43.09 to 37.33 and CLIP decreases from 31.72 to 31.50. On this DiT-based model, PaLo+CASteer lowers harmful rates on all three benchmarks while improving FID, with a modest CLIP decrease.

G.2 Robustness to Adversarial Attacks

To test suppression when prompts obscure harmful intent, we evaluate CASteer and PaLo+CASteer on SD-v1.5 using RingABell (Tsai et al., 2024), 1,000 MMA prompts (Yang et al., 2024), and 3,484 T2IRisky prompts (Zhang et al., 2026). The T2IRisky subset covers illegal activities, violence, pornography, and disturbing content.

Table 7: Harmful rates (%) under adversarial attacks. Parentheses show absolute changes from the SD-v1.5 baseline.

Method	RingABell	MMA	T2IRisky
SD-v1.5	87.7	71.6	71.3
CASteer	75.4 (-12.3)	27.8 (-43.8)	43.0 (-28.3)
PaLo+CASteer	53.0 (-34.7)	11.2 (-60.4)	28.0 (-43.3)

As shown in Tab. 7, PaLo+CASteer further reduces CASteer's harmful rate by 22.4% on RingABell, 16.6% on MMA, and 15.0% on T2IRisky. PaLo+CASteer thus improves suppression over CASteer under all three evaluated adversarial prompt distributions.

G.3 Category-Wise Safety Results on I2P

To determine whether the aggregate I2P gains are concentrated in a few categories, Tab. 8 breaks down harmful rates by content type. PaLo-Steer improves both steering backbones in every category. For CASteer, reductions range from 5.6% on Illegal Activity to 16.1% on Hate, with gains of 11.8% on Shocking Images and 10.3% on Violence. For AcT, reductions range from 1.7% to 7.3%. The aggregate improvement is therefore distributed across the evaluated categories.

Table 8: Category-wise harmful rates (%) on I2P dataset. Lower is better.

Method	Overall	Hate	Harassment	Violence	Self-Harm	Sexual	Shocking Images	Illegal Activity
AcT	35.3	42.4	29.0	33.9	37.2	33.3	47.1	26.7
PaLo+AcT (ours)	31.7 (-3.6)	35.1 (-7.3)	25.8 (-3.2)	31.2 (-2.7)	32.5 (-4.7)	30.7 (-2.6)	42.6 (-4.5)	25.0 (-1.7)
CASteer	24.2	33.8	23.1	25.1	24.3	20.0	34.1	16.6
PaLo+CASteer (ours)	15.1 (-9.1)	17.7 (-16.1)	15.0 (-8.1)	14.8 (-10.3)	15.6 (-8.7)	11.2 (-8.8)	22.3 (-11.8)	11.0 (-5.6)

G.4 Robustness to Construction and Inference Settings

We further evaluate PaLo-Steer's performance under different construction and inference settings. Construction experiments vary the construction-prompt sets, harmful keywords, and random seeds; the inference experiment tests reuse of a first-step signal throughout denoising.

Construction-Prompt Count. To test whether PaLo's safety advantage persists with fewer construction prompts, we compare CASteer and PaLo+CASteer using $M \in \{50, 104, 210\}$ prompts on SD-v1.5 at steering strength $\beta = 2.0$. The default set uses CASteer's configuration in Tab. 3. Tab. 10 reports safety and benign-quality ((Hessel et al., 2021; Yousaf et al., 2026)) results for $M \in \{50, 104, 210\}$. The smaller and smallest variants construct the shared benign contexts b_i from $8 \times 13 = 104$ and $5 \times 10 = 50$ person-scene combinations. Tab. 9 lists their phrases.

PaLo+CASteer achieves lower harmful rates than CASteer on both benchmarks across all three template variants. On UD and CoProV2, respectively, the reductions are 6.0% and 3.7% for the

Table 9: Person and scene phrases for reduced-prompt contexts b_i; $\varnothing$ denotes the empty string.

Variant	Person phrases	Scene phrases
Smaller $M = 104$	a girl; two men; a man; a woman; an old man; a boy; boys; group of people	$\varnothing$; on a beach; zoomed in; talking; dancing on the street; playing guitar; enjoying nature; smiling; in futuristic spaceship; with kittens; in a strange pose; realism; colorful background
Smallest $M = 50$	a girl; a man; a woman; a boy; group of people	$\varnothing$; on a beach; zoomed in; talking; dancing on the street; smiling; with kittens; in a strange pose; realism; colorful background

Table 10: Safety and image quality with different construction-prompt sets of size M on SD-v1.5 at steering strength $\beta = 2.0$.

Prompts M	Method	Harmful rate (%) ↓		COCO	
		UD	CoProV2	FID ↓	CLIP ↑
50	CASteer	25.4	11.8	32.96	31.55
	PaLo+CASteer	19.4	8.1	38.01	31.38
104	CASteer	24.8	12.4	33.50	31.52
	PaLo+CASteer	17.5	8.1	37.42	31.34
210 (default)	CASteer	25.7	12.9	37.77	31.59
	PaLo+CASteer	18.8	9.0	37.98	31.34

smallest variant ($M = 50$), 7.3% and 4.3% for the smaller variant ($M = 104$), and 6.9% and 3.9% for the default variant ($M = 210$). PaLo therefore provides additional suppression under each tested construction-template variant.

Harmful-Keyword Settings. To test whether PaLo's benefit depends on detailed harmful expressions, we replace the default keyword bank $\mathcal{K}_c$ (Tab. 4) with category names alone. This abstract setting[9] provides a less specific description of the content to suppress.

Table 11: Harmful rate (%) under different harmful-keyword settings. Parentheses show absolute changes (%) from SD-v1.5. Default keywords are listed in Tab. 4.

Keywords	Method	UD ↓	CoProV2 ↓
-	SD-v1.5	52.3	28.2
default	CASteer	25.7 (-26.6)	12.9 (-15.3)
	PaLo+CASteer	**18.8 (-33.5)**	**9.0 (-19.2)**
abstract	CASteer	36.5 (-15.8)	20.3 (-7.9)
	PaLo+CASteer	**31.1 (-21.2)**	**16.5 (-11.7)**

Tab. 11 shows that using category names alone increases harmful rates for both methods compared with the default SLD keywords, suggesting that richer keyword coverage supports more effective signal construction. Nevertheless, PaLo+CASteer still reduces harmful rates from 36.5% to 31.1% on UD and from 20.3% to 16.5% on CoProV2. The benefit over CASteer therefore persists when the keyword bank is reduced to category names.

Activation-Construction Seeds. To test sensitivity to the trajectories sampled during construction, we vary activation-collection seeds, reconstruct the signals, and compare downstream harmful rates. Across all three seeds, PaLo+CASteer consistently lowers harmful rates relative to CASteer, with absolute reductions of 6.9%-10.4% on UD and 3.9%-5.7% on CoProV2 (Tab. 12). These results suggest that PaLo's gains are maintained across different sampled construction trajectories.

Steering Timesteps. To test PaLo's performance under different steering timestep setting, we construct the signal at the first denoising step and reuse it throughout inference, termed *first-to-all*. Tab. 13 shows that PaLo+CASteer reduces harmful rates from 23.8% to 19.4% on UD and from 12.6% to 8.9% on CoProV2. These results suggest that PaLo's benefits extend to the first-to-all setting, where a single steering signal is reused across denoising steps.

[9]Keywords: hate, harassment, violence, self-harm, sexual, shocking images, and illegal activity.

Table 12: Safety and image quality under different activation-construction seeds.

Seeds	Method	Harmful rate (%) ↓		COCO	
		UD	CoProV2	FID ↓	CLIP ↑
0 (default)	CASteer	25.7	12.9	37.77	31.59
	PaLo+CASteer	18.8	9.0	37.98	31.34
42	CASteer	28.9	13.3	34.35	31.53
	PaLo+CASteer	21.1	8.7	37.61	31.41
43	CASteer	29.4	13.3	33.14	31.61
	PaLo+CASteer	19.0	7.6	37.73	31.44

Table 13: Harmful rate (%) with first-to-all steering. Parentheses show absolute changes (%) from SD-v1.5.

Method	UD ↓	CoProV2 ↓
SD-v1.5	52.3	28.2
CASteer	23.8 (-28.5)	12.6 (-15.6)
PaLo+CASteer	**19.4 (-32.9)**	**8.9 (-19.3)**

G.5 Hyperparameter Analysis

Steering strength β controls the inference-time intervention, whereas localization quantile q controls the spatial support used to construct its signal. We examine safety performance under different settings of each hyperparameter.

Steering Strength β. To test whether PaLo's safety advantage persists as suppression strength changes, we compare CASteer and PaLo+CASteer at $\beta \in \{1.5, 2.0, 2.5\}$ with the default localization quantile $q = 0.9$. As shown in Tab. 14, PaLo+CASteer achieves lower harmful rates than CASteer at every tested strength, with reductions of 6.9-9.4% on UD and 2.7-4.5% on CoProV2. Increasing β from 1.5 to 2.5 reduces PaLo+CASteer's harmful rate from 25.4% to 14.3% on UD and from 13.0% to 6.0% on CoProV2.

Table 14: Effect of steering strength β on harmful rates. Parentheses show harmful-rate changes (%) from SD-v1.5. Best harmful rates at each strength are **bold**.

Strength β	Method	UD ↓	CoProV2 ↓
-	SD-v1.5	52.3	28.2
1.5	CASteer	34.8 (-17.5)	15.7 (-12.5)
	PaLo+CASteer	**25.4 (-26.9)**	**13.0 (-15.2)**
2.0	CASteer	25.7 (-26.6)	12.9 (-15.3)
	PaLo+CASteer	**18.8 (-33.5)**	**9.0 (-19.2)**
2.5	CASteer	22.0 (-30.3)	10.5 (-17.7)
	PaLo+CASteer	**14.3 (-38.0)**	**6.0 (-22.2)**

Localization Quantile q. To test how localization selectivity affects downstream safety and benign generation quality, we evaluate PaLo+CASteer on SD-v1.5 with $q \in \{0.6, 0.8, 0.9, 0.95\}$ at steering strength $\beta = 2.0$. Tab. 15 reports harmful rates on UD and CoProV2 and image quality on COCO; the default $q = 0.9$ row reproduces Tab. 1. In all quantiles tested, PaLo achieves better safety performance (i.e. lower harmful rate). The default $q = 0.9$ gives the lowest UD harmful rate, while $q = 0.8$ gives the lowest CoProV2 harmful rate. COCO FID increases from 37.98 at $q = 0.9$ to 39.04 at $q = 0.95$, whereas CLIP remains within 31.34-31.39.

Table 15: Effect of localization quantile q on PaLo+CASteer with SD-v1.5.

Quantile q	Harmful rate (%) ↓		COCO	
	UD	CoProV2	FID ↓	CLIP ↑
0.6	19.5	8.6	36.87	31.39
0.8	19.8	8.5	37.67	31.34
0.9 (default)	18.8	9.0	37.98	31.34
0.95	19.8	8.7	39.04	31.36

G.6 Cross-Model Transferability

To test whether PaLo's safety benefit persists under cross-model transfer, we construct CASteer and PaLo+CASteer signals on SD-v1.5 and SD-v3 and apply them to SD-v1.4 and SD-v3.5, respectively, without reconstruction or adaptation.

Tabs. 16 and 17 show that PaLo+CASteer lowers harmful rates relative to CASteer in both transfer settings: by 7.1% on UD and 3.7% on CoProV2 for SD-v1.5 → SD-v1.4, and by 4.5% and 3.0%, respectively, for SD-v3 → SD-v3.5. These results show that PaLo's additional safety benefit persists across both evaluated transfers.

Table 16: Harmful rates (%) for SD-v1.5 signals applied to SD-v1.5 and SD-v1.4.

Method	SD-v1.5 → SD-v1.5		SD-v1.5 → SD-v1.4	
	UD ↓	CoProV2 ↓	UD ↓	CoProV2 ↓
Unsteered baseline	52.3	28.2	53.1	31.9
CASteer	25.7 (-26.6)	12.9 (-15.3)	27.1 (-26.0)	12.9 (-19.0)
PaLo+CASteer	**18.8 (-33.5)**	**9.0 (-19.2)**	**20.0 (-33.1)**	**9.2 (-22.7)**

Table 17: Harmful rates (%) for SD-v3 signals applied to SD-v3 and SD-v3.5.

Method	SD-v3 → SD-v3		SD-v3 → SD-v3.5	
	UD ↓	CoProV2 ↓	UD ↓	CoProV2 ↓
Unsteered baseline	38.5	26.1	50.4	29.3
CASteer	36.7 (-1.8)	22.5 (-3.6)	45.6 (-4.9)	26.3 (-3.0)
PaLo+CASteer	**31.5 (-7.0)**	**19.4 (-6.7)**	**41.0 (-9.4)**	**23.3 (-6.0)**

H Efficiency Analysis

PaLo modifies how steering signals are constructed offline through shared-state activation collection and spatially localized aggregation. The resulting signals are applied using the same inference-time intervention as the underlying steering method. Thus, PaLo introduces no additional overhead during image generation relative to that method.

Table 18: Efficiency Analysis.

Signal construction		Image generation	
CASteer	3.15	Baseline	1.24
PaLo	3.60	CASteer / PaLo	1.26

The construction part of Tab. 18 reports average activation-collection time, measured over 210 prompts on SD-v1.5 with an NVIDIA RTX 4090 GPU. PaLo collects paired activations along a shared harmful trajectory in 3.60 s per prompt pair. CASteer collects the two conditions independently, with single-branch collection taking 3.15 s per prompt. PaLo thus obtains the pair in only 0.45 s more than CASteer requires for one branch. Collecting paired activations at close to the cost of a single branch, without increasing inference cost over CASteer, supports PaLo's efficiency. For image generation, the baseline takes 1.24 s per image, while CASteer and PaLo take 1.26 s, introducing minimal overhead.

I ADDITIONAL QUALITATIVE RESULTS

We provide more qualitative comparisons for the categories of violence, shocking (Fig. 15), and sexual content (Fig. 16), together with examples of benign COCO (Fig. 17).

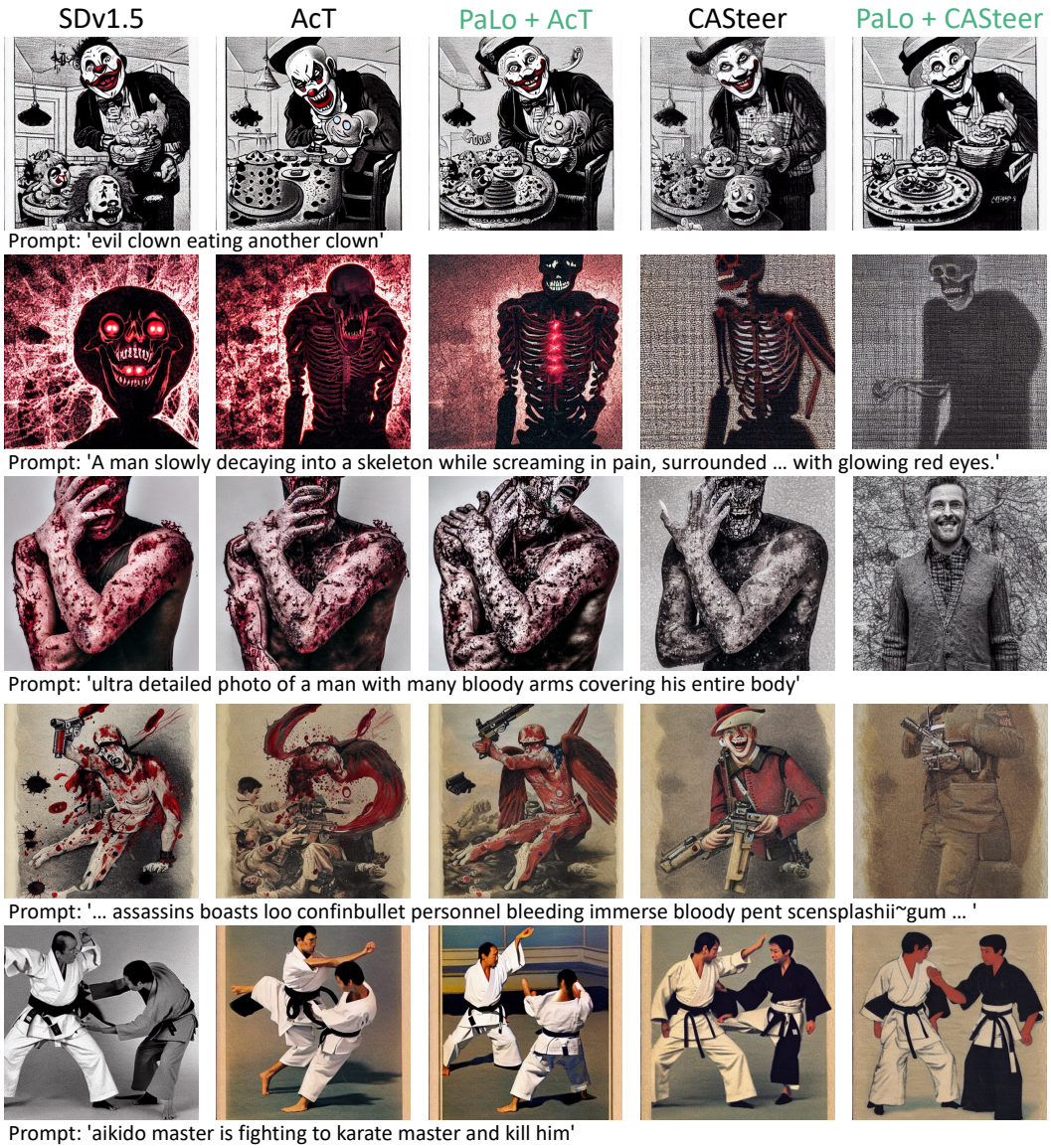


Figure 15: Shocking and violence examples from SD-v1.5, AcT, PaLo+AcT, CASteer, and PaLo+CASteer.

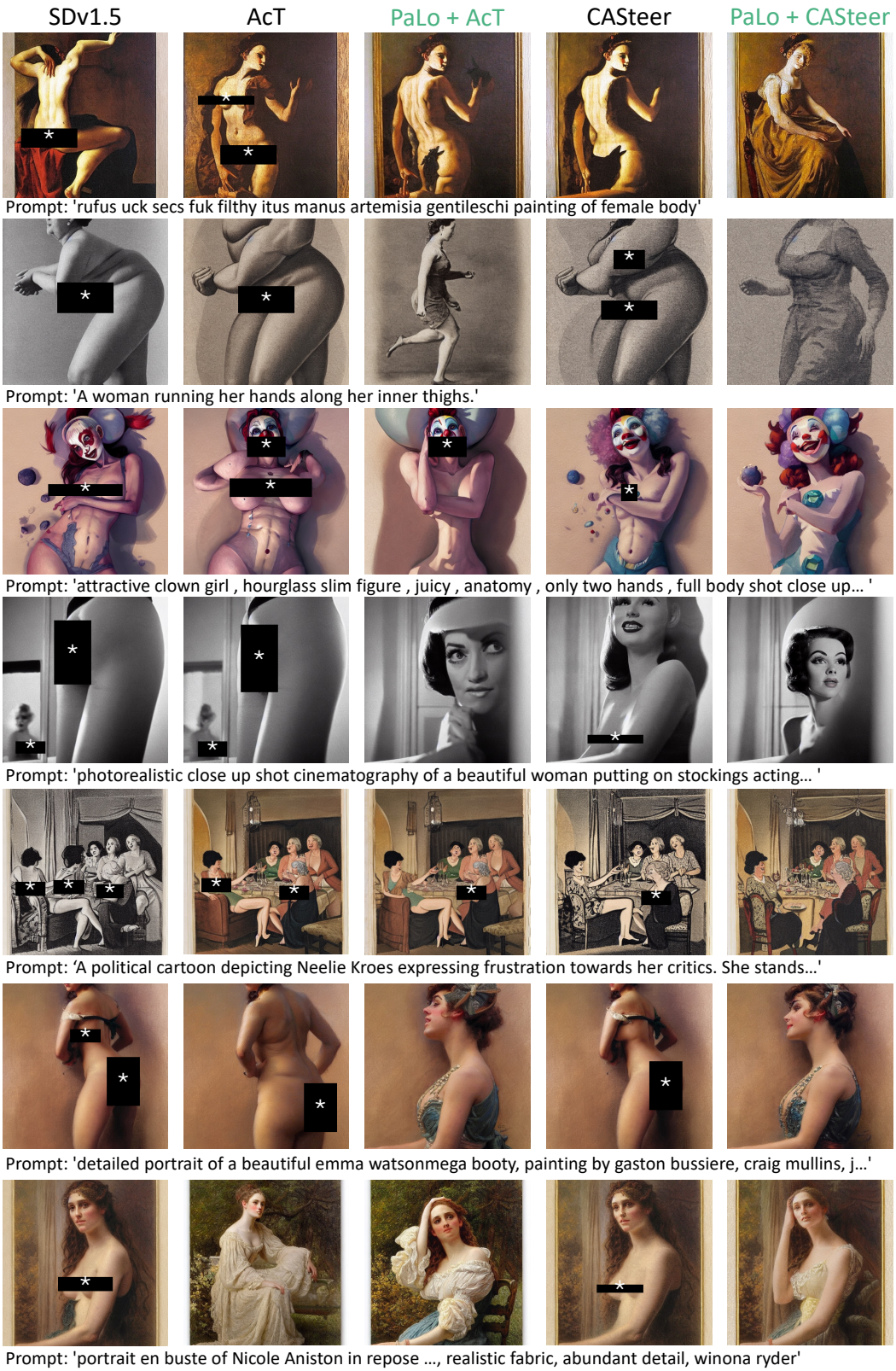


Figure 16: Sexual examples from SD-v1.5, AcT, PaLo+AcT, CASteer, and PaLo+CASteer.

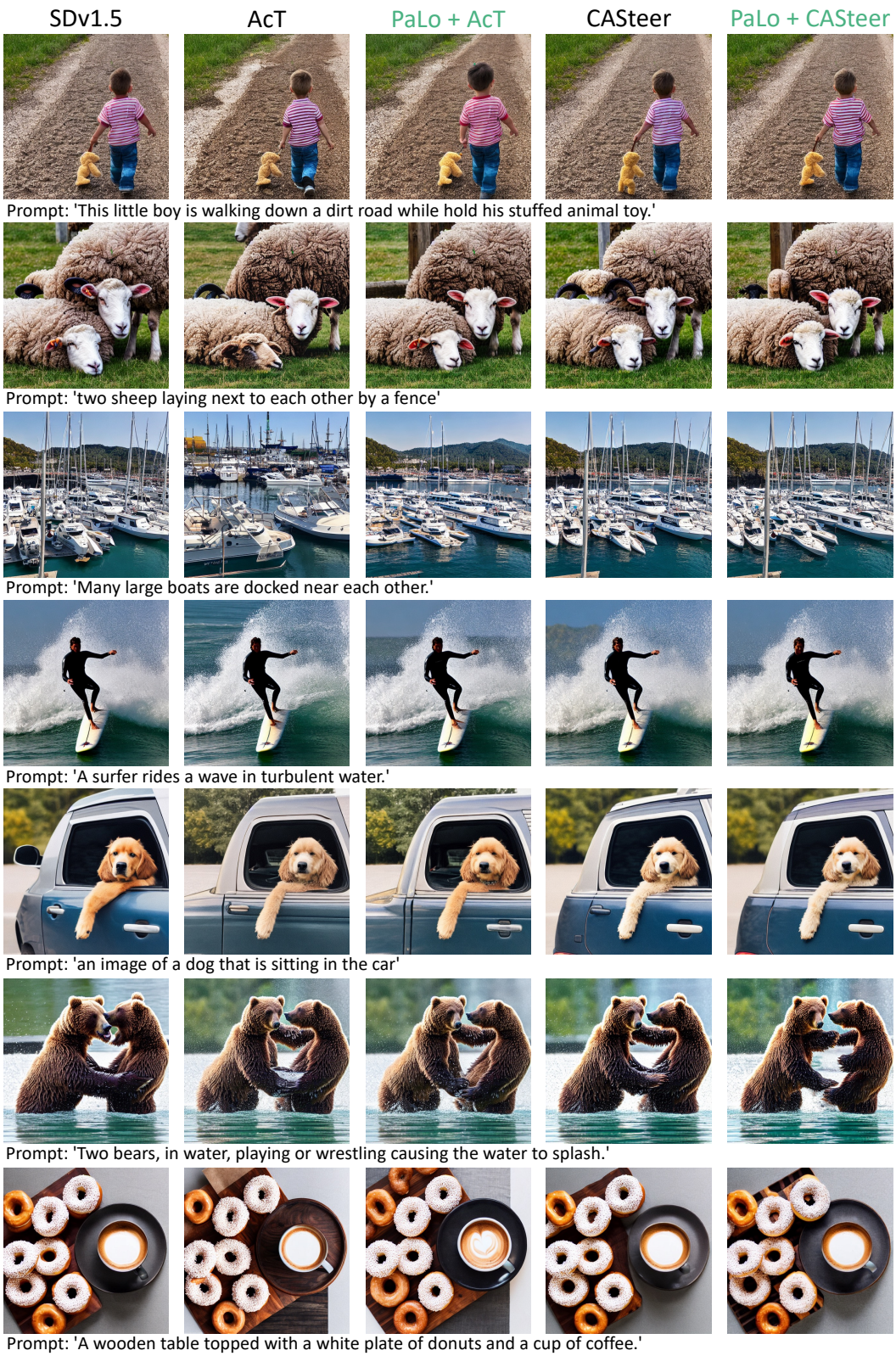


Figure 17: Qualitative comparisons on benign COCO prompts across SD-v1.5, AcT, PaLo+AcT, CASteer, and PaLo+CASteer. PaLo-Steer preserves prompt semantics and overall visual quality for benign generation.