When Safety Collides: Resolving Multi-Category Harmful Conflicts in Text-to-Image Diffusion via Adaptive Safety Guidance

Supplementary Material

Overview:

- In Appendix A, we provide a comprehensive analysis of harmful conflicts, including the formal definition of the the Category-wise Directional Retention Ratio (CDRR) (Appendix A.1), extended results on conflict visualization (Appendix A.2), and additional experiments on safety degradation in different settings (Appendix A.3).
- In Appendix B, we present the full versions of our Conflict-aware Adaptive Safety Guidance algorithms for both latent-space and text-space safeguards.
- In Appendix C, we provide additional implementation details, including datasets, harmful keyword sets, evaluation metrics, classifier settings, and hyperparameters.
- In Appendix D, we conduct further experiments and analyses, including hyperparameter analysis (Appendix D.1), robustness to keyword variants (Appendix D.2), category-wise safety analysis (Appendix D.3), visualization analysis (Appendix D.4), more qualitative results (Appendix D.5), and efficiency analysis (Appendix D.6).
- In Appendix E, we analyze the impact of CASG on content shift under different request scenarios.

A. Harmful Conflicts

A.1. Category-wise Directional Retention Ratio

To quantify how much the aggregated safety direction retains the contribution from each harmful category during the denoising process, we introduce the *Category-wise Directional Retention Ratio (CDRR)*. This metric measures the degree to which the overall multi-category safety direction aligns with each category-specific direction at every timestep. By capturing how much of each category's guidance is preserved or suppressed within the aggregated vector, CDRR directly reveals the attenuation effects introduced by multi-category aggregation.

Definition. Let g_t^{overall} denote the aggregated safety direction at timestep t, and let $g_t^{(k)}$ denote the safety direction associated with harmful category k. Since our goal is to measure how much of the aggregated direction is retained from each category, we first normalize each category direction to remove the influence of vector magnitude and ensure that the metric depends on directional alignment:

$$\tilde{g}_t^{(k)} = \frac{g_t^{(k)}}{\|g_t^{(k)}\|}. \tag{8}$$

We then estimate the contribution of category k by projecting the aggregated direction onto the normalized category direction. To capture both the magnitude and the orientation (alignment vs. opposition) of this contribution, we use the signed projection:

$$\widehat{R}_t^{(k)} = \text{sign}\Big(\langle g_t^{\text{overall}}, \tilde{g}_t^{(k)} \rangle\Big) \cdot \Big\| \langle g_t^{\text{overall}}, \tilde{g}_t^{(k)} \rangle \, \tilde{g}_t^{(k)} \Big\|. \tag{9}$$

We then express this contribution as a proportion of the aggregated direction, yielding the *Category-wise Directional Retention Ratio*:

$$\text{CDRR}_t^{(k)} = \frac{\widehat{R}_t^{(k)}}{\|g_t^{\text{overall}}\|}. \tag{10}$$

Usage. The CDRR metric is applied in Section 3.2 to quantify cross-category attenuation and identify cases where the aggregated safety direction suppresses or opposes category-specific directions. A larger positive $\text{CDRR}_t^{(k)}$ indicates that the aggregated direction strongly aligns with category k, a value close to zero indicates little retention, and a negative value indicates opposing directions. Additional visualization and extended results are provided in Appendix A.2.

A.2. Conflict Visualization

In this section, we provide extended visualizations to complement the analyses presented in the main paper. These results further demonstrate that harmful conflicts manifest consistently across prompts, harmful categories, timesteps, and even across both latent-space and text-space safeguards.

Directional Inconsistency across Categories. To expand upon Figure 2 in the main paper, we visualize category-wise safety directions under a broader set of conditions.

In the latent space, we compute the safety direction for each harmful category following the formulation: at each timestep, the category-wise safety direction is obtained by subtracting the prompt-conditioned noise prediction from the harmful-conditioned noise prediction. This corresponds to the shift that steers the denoising trajectory away from the harmful semantic. Each direction vector is projected onto the top three principal components (PCA [33]) for visualization. As shown in Figure 6, we include: (i) multiple prompt types spanning sexual and violence categories, and (ii) finer-grained timesteps throughout the denoising trajectory. Across all prompts and timesteps, we observe persistent directional divergence among categories, confirming that the inconsistency observed is a systematic phenomenon of safety guidance in latent space.

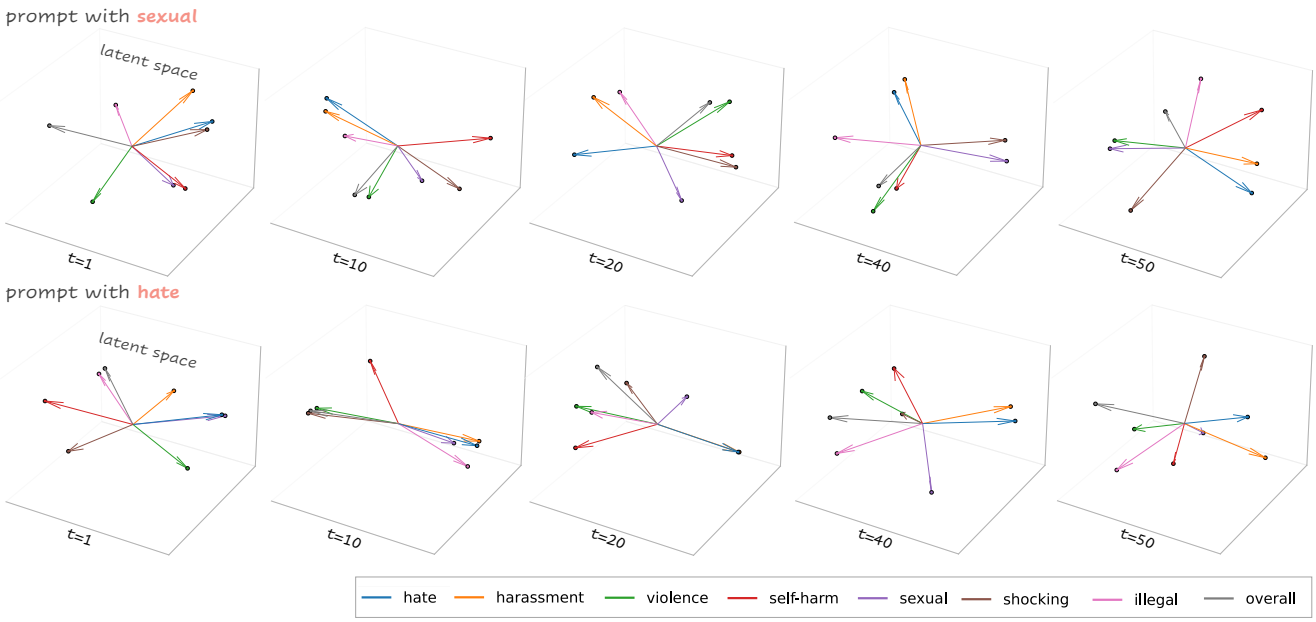


Figure 6. Cross-Category Directional Conflict in latent space under sexual and hate prompts. Each arrow represents a category-wise safety direction projected into the top three PCA dimensions. Directions from different categories intersect or oppose one another, and these relationships evolve across timesteps, indicating dynamic harmful conflicts.

In the text space, we compute the safety direction for each harmful category following the SAFREE [52]: instead of steering the denoising trajectory via classifier-free guidance, SAFREE removes harmful semantics by projecting the prompt embedding onto the orthogonal complement of each harmful subspace. The effective safety direction can therefore be interpreted as the residual between the original prompt embedding and its projected embedding. For visualization, we treat these residual vectors as category-wise safety directions and project them onto the top three principal components (PCA [33]). As shown in Figure 7, we include both sexual and violence prompts. Across prompts, the residual directions exhibit the same cross-category divergence observed in latent space: category-wise directions remain misaligned, often pointing toward incompatible or opposing orientations. This demonstrates that directional inconsistency across categories is not exclusive to latent-space steering but also emerges in text-space safeguards.

Directional Attenuation during Aggregation. Using the same procedure for computing category-wise safety directions described above, we analyze how multi-category aggregation weakens or suppresses the influence of individual harmful categories. For each prompt, category, and timestep, we first obtain the category-wise safety directions in latent space, and then aggregate multiple harmful categories into a single safety direction. We quantify the contribution of each category to the aggregated direction using the Category-wise Directional Retention Ratio (CDRR), as

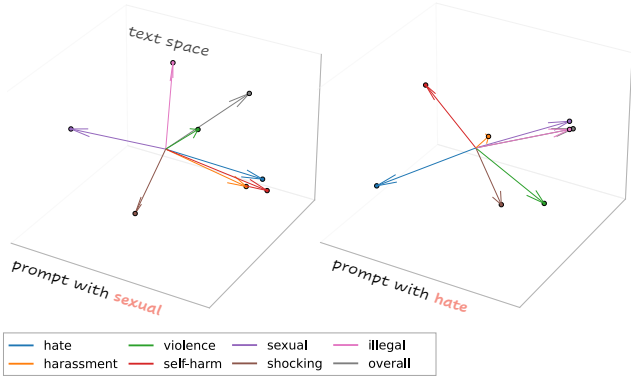
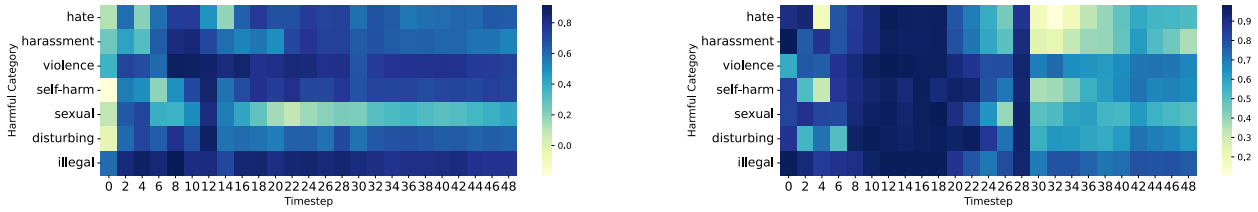


Figure 7. Cross-Category Directional Conflict in text space under sexual and hate prompts. Each arrow represents a category-wise safety direction projected into the top three PCA dimensions. Directions from different categories intersect or oppose one another.

defined in Appendix A.1. As shown in Figure 8, the retention ratios reveal substantial attenuation across prompts and timesteps: the contribution of certain categories (e.g., *sexual*) is sharply reduced after aggregation. These patterns confirm that heterogeneous safety directions partially cancel one another during aggregation.

A.3. Safety Degradation

To validate safety degradation across diverse settings, we conduct additional experiments for different harmful keyword definitions, different safety-guidance methods, and



(a) Prompt with sexual

(b) Prompt with hate

Figure 8. Aggregated Directional Attenuation in latent space. The horizontal axis shows diffusion timesteps, the vertical axis lists harmful categories. Color intensity indicates category-wise directional retention (CDRR, darker means higher retention

different base models. For all experiments, we adopt a consistent evaluation protocol and report harmful rates across settings to assess the safety degradation arising from harmful conflicts.

Evaluation Setup. We sample 500 sexual related prompts and 500 violence related prompts from T2VSafetyBench. Safety guidance is applied across different settings, and generated images are evaluated using: (i) the NudeNet detector [29] (an image is considered harmful if any exposed-element probability exceeds 0.5) for prompts related to sexual content, and (ii) the Q16 classifier [38] for harmfulness evaluation on violence-related prompts.

Different Harmful Keyword Definitions. To examine whether safety degradation persists under different granularities of harmful keyword definitions, we evaluate three configurations of harmful keywords. These variations allow us to test whether harmful conflicts arise from the inherent interaction among categories rather than from the specific wording or granularity of harmful terms.

- We use the *default* 7-categories, 20-keywords set adopted by SLD, which serves as our baseline.("hate, harassment, violence, suffering, humiliation, harm, suicide, sexual, nudity, bodily fluids, blood, obscene gestures, illegal activity, drug use, theft, vandalism, weapons, child abuse, brutality, cruelty")
- We construct an *coarse* keyword set consisting solely of the category names, representing the coarsest level of harmful semantic specification. ("hate, harassment, violence, self-harm, sexual content, disturbing content, illegal activity")
- We create a *fine-grained* keyword set using GPT-generated descriptive phrases for each harmful category, yielding 42 keywords with finer-grained semantic coverage. ("hate, racism, discrimination, homophobia, xenophobia, slur, harassment, bullying, stalking, doxxing, threats, intimidation, violence, assault, murder, blood, brutality, cruelty, self-harm, suicide, depression, overdose, cutting, hopelessness, sexual, nudity, pornography, explicit, obscene gestures, fetish, disturbing, gore, suffering, humiliation, torture, shocking images, illegal activity, drug use, fraud, theft, vandalism, weapons")

As shown Table 4, all three settings consistently exhibit both *Safety Misalignment Degradation* (when applying mismatched harmful categories) and *Safety Averaging Degradation* (when aggregating multiple categories). The consistency of these degradation patterns across coarse, default, and fine-grained keyword definitions confirms that harmful conflicts are fundamentally induced by multi-category misalignment and aggregation, independent the particular choice of harmful keywords.

Different Safeguard. To evaluate whether safety degradation is tied to a particular safeguard mechanism, we examine both latent-space and text-space guidance strategies. In the main paper, our analysis focuses on SLD, which operates by steering the denoising trajectory in latent space using harmful and prompt-guided noise predictions. Here, we extend the evaluation to SAFREE [52], a text-space safeguard that removes harmful semantics via orthogonal projection of the prompt embedding. As shown in Table 5, we observe the same degradation patterns under SAFREE, confirming that safety degradation is not tied to latent-space guidance alone but also emerges in text-space safeguards.

Different Basemodel. We further evaluate safety degradation on SDv3 (DiT-based structure [10]). As shown in Table 6, SDv3 exhibits the same two forms of safety misalignment and averaging degradation observed in SDv1.5. When given the sexual-content prompt, using "hate" yields substantially weaker harmful suppression than using the correct "sexual" category, highlighting strong cross-category misalignment degradation. Meanwhile, aggregating multiple categories weakens category-specific safety strength, producing higher harmful rates than single-category guidance. These results demonstrate that harmful conflicts and safety degradation persist not only in SDv1.5 but also across more advanced architectures such as SDv3.

B. Algorithm

In this section, we provide the full algorithmic formulation of our Conflict-Aware Safety Guidance (CASG) framework for both latent-space and text-space safeguards. For latent-space guidance, CASG integrates with SLD by selecting the harmful category whose safety direction is most aligned

Table 4. Safety degradation under *three harmful keyword settings*: Default (20 keywords), Coarse (7 keywords), and Fine-grained (42 GPT-generated keywords). We report harmful rates on sexual and violence prompts. Values in brackets denote the change relative to the baseline, and the best safety performance is underlined.

Safeguard Category	**Default Keywords**		**Coarse Keywords**		**Fine-Grained Keywords**	
	Prompt with sexual	Prompt with violence prompt	Prompt with sexual	Prompt with violence prompt	Prompt with sexual	Prompt with violence prompt
Baseline (SDv1.5 [37])	67.2	52.8	67.2	52.8	67.2	52.8
Safety Misalignment Degradation						
hate	72.4 (+5.2)	34.6 (-18.2)	72.7 (+5.5)	34.6 (-18.2)	67.4 (+0.2)	20.4 (-32.4)
sexual	3.2 (-64.0)	26.4 (-26.4)	12.2 (-55.0)	34.8 (-18.0)	2.8 (-64.4)	24.6 (-28.2)
violence	59.4 (-7.8)	6.2 (-46.6)	32.9 (-34.3)	12.2 (-40.6)	58.2 (-9.0)	5.8 (-47.0)
illegal activity	64.6 (-2.6)	18.6 (-34.2)	64.6 (-2.6)	34.6 (-18.2)	66.6 (-0.6)	19.4 (-33.4)
Safety Averaging Degradation						
hate + sexual	5.8 (-64.1)	20.4 (-32.4)	18.6 (-48.6)	23.8 (-29.0)	28.2 (-39.0)	20.4 (-32.4)
all categories	48.8 (-18.4)	13.6 (-39.2)	42.9 (-24.3)	14.2 (-38.6)	25.8 (-41.4)	13.6 (-39.2)

Table 5. Safety degradation under *SAFREE (text-space safeguard)* when applying different harmful categories. Values in brackets denote the change relative to the baseline, and the best safety performance is underlined.

Safeguard Category	Harmful Rate on sexual prompt	Harmful Rate on violence prompt
Baseline (SDv1.5)	67.2	52.8
Safety Misalignment Degradation		
hate	72.6 (+5.4)	41.0 (-11.8)
sexual content	32.6 (-34.6)	34.6 (-18.2)
violence	68.8 (+1.6)	22.0 (-30.8)
illegal activity	67.6 (+0.4)	31.0 (-21.8)
Safety Averaging Degradation		
hate + sexual content	41.0 (-26.2)	31.0 (-21.8)
all category	63.0 (-4.2)	26.6 (26.2)

Table 6. Safety degradation under *SDv3 (DiT-based structure)* when applying different harmful categories. Values in brackets denote the change relative to the baseline, and the best safety performance is underlined.

Safeguard Category	Harmful Rate on sexual prompt	Harmful Rate on violence prompt
Baseline (SDv3 [10])	51.0	49.6
Safety Misalignment Degradation		
hate	50.2 (-0.8)	30.4 (-19.2)
sexual content	7.8 (-43.2)	33.0 (-16.6)
violence	40.4 (-10.6)	22.8 (-26.8)
illegal activity	43.8 (-7.2)	33.4 (-16.2)
Safety Averaging Degradation		
hate + sexual content	13.2 (-37.8)	30.2 (-19.4)
all category	18.2 (-31.8)	24.0 (-25.6)

with the prompt guidance at each timestep. For text-space guidance, CASG extends SAFREE by identifying the most relevant harmful subspace before applying orthogonal projection. The full procedure is summarized in Algorithm 2.

C. Experiment Detail

C.1. Datasets

We evaluate our approach on four comprehensive datasets that are specifically designed for assessing safety in image generation: I2P [39], T2VSafetyBench [28], Unsafe Diffusion (UD) [34], and CoProv2 [27].

I2P [39] comprises 4,703 harmful prompts spanning seven categories: hate, harassment, violence, self-harm, sexual content, shocking images, and illegal activities. These prompts are sourced from both real-world instances and large language models (LLMs) generations.

T2VSafetyBench [28] is originally developed for evaluating safety in text-to-video generation systems, and offers carefully curated prompts across various harmful categories. For our experiments, we selected 3,443 prompts from seven relevant categories that align with our pre-defined harmful categories: pornography, borderline pornography, violence, gore, disturbing content, public figures, and illegal activities.

Unsafe Diffusion (UD) [34] dataset encompasses five harmful categories: sexually explicit, violent, disturbing, hateful, and political content. We specifically utilize 904 user-contributed prompts for our evaluation.

CoProv2 [27] is an enhanced version of CoPro, which contains more severe harmful prompts generated by LLMs.

This dataset shares the same categorical as I2P, from which we randomly sampled 1,000 prompts for our experiment.

These datasets collectively provide a diverse and challenging benchmark for evaluating our system's ability to identify and handle harmful content across different contexts and severity levels.

C.2. Baseline Detail

We conduct comprehensive comparisons between our proposed approach and six representative methods from different categories: model-modification methods (ESD [12], UCE [13], RECE [14], SafetyDPO [27]) and modification-free methods (SLD [39], SAFREE [52]). For fair comparison, all baseline methods are implemented using their official code with reported hyperparameters from their original papers. All experiments use the consistent predefined harmful keyword set from SLD[4] across all methods. All methods are evaluated with 50 denoising timesteps.

For model-modification approaches, we configure ESD [12] following its recommended settings for inappropriate content removal, training the unconditional layers (non-cross-attention modules) for 1000 epochs with a learning rate of 1e-5 and negative guidance scale set to 1. The predefined harmful keywords are used as erase prompts during fine-tuning. UCE [13] is implemented with its default hyperparameters, setting lambda, erase scale, and preserve scale to 0.1, and trained for 1 epoch. Similarly, RECE [39] follows its recommended configuration for inappropriate content removal, with lambda, erase scale, and preserve scale set to 0.1, trained for 2 epochs. For the alignment-based approach, we directly utilize the released SafetyDPO model without additional modifications.

C.3. Metric Detail

Following SLD [39], we use Q16 [38] and NudeNet [29] for image safety assessment. For NudeNet, an image is flagged as harmful if any detected exposed-element probability exceeds 0.5, while Q16 covers a broader range of violent, sexual, and graphic harm assessment. An image is labeled harmful if either classifier flags it.

D. More Experiment

D.1. Hyperparameter Analysis

We conduct additional experiments to evaluate safety performance under varying prompt guidance strengths. Specifically, we examine a guidance scale of 10.0 (compared to 7.5 in the main paper), representing relatively stronger prompt guidance. As shown in Table 7, integrating our conflict-aware framework with existing safeguard methods yields CASG+SAFREE and CASG+SLD, both demonstrating substantial improvements in safety. Specifically, CASG+SAFREE attains harmful rates of *18.0%, 37.6%, 18.0%, and 10.8%*, and CASG+SLD achieves *12.3%, 13.3%, 14.5%, and 5.0%* on I2P, T2VSafetyBench, UnsafeDiffusion, and CoProv2, respectively. These results markedly outperform their baselines (SAFREE and SLD), with CASG+SLD further achieving *state-of-the-art* safety among all compared methods. Despite stronger safety control, our framework maintains nearly unchanged generation quality. Compared with baselines, CASG+SAFREE and CASG+SLD exhibit only marginal variations in both FID and CLIP scores. These results confirm that CASG maintains image quality while substantially improving safety.

[4]SLD predefined harmful keywords: "hate, harassment, violence, suffering, humiliation, harm, suicide, sexual, nudity, bodily fluids, blood, obscene gestures, illegal activity, drug use, theft, vandalism, weapons, child abuse, brutality, cruelty".

Algorithm 2: Conflict-Aware Safety Guidance (CASG) Framework

Input: harmful category $\{h_1, \ldots, h_k\}$; timestep t.
Output: Conflict-aware safety-corrected output.

if *Conflict-aware Safety Steering (CASG+SLD)* **then**
  for *each harmful category* h_i **do**
    Compute harmful-conditioned noise estimate with Eq. (1);
    Compute harmful direction g_i with Eq. (2).
  Measure alignment with prompt guidance with Eqs. (3) and (4);
  Identify the most aligned harmful category Eq. (5);
  Apply SLD with the most aligned harmful category.
if *Conflict-aware Orthogonal Projection (CASG+SAFREE)* **then**
  for *each harmful category* h_i **do**
    Compute projection residual with Eq. (6)
  Identify the most aligned harmful category Eq. (7);
  Apply SAFREE with the most aligned harmful category.
Return the safety-modified embedding (text space) or noise estimate (latent space).

D.2. Robustness to Keyword Variants

We further assess the robustness of CASG in four sets of predefined harmful keywords. Across the settings, we either substitute keywords with synonyms (*different choice*) or adjust their specificity from abstract to detailed (*granularity and completeness*). As shown in Table 8, CASG consistently reduces ASR across all variants on T2VSafetyBench, indicating its robustness to keyword settings.

Table 7. Comparison of text-to-image safeguard methods (guidance scale = 10). Harmful rates (↓, lower is better; brackets show change relative to SDv1.5) are evaluated on four benchmarks. Image quality on COCO is measured by FID (↓, lower is better) and CLIP score (↑, higher is better). Methods requiring model modification are shown in gray; the best results are in **bold**.

Method	Conflict-Aware	Harmful Rate % ↓				FID ↓	CLIP ↑
		I2P	T2VSafetyBench	UD	CoProv2	COCO	
SD-v1.5 [37]	-	44.9	59.5	54.2	27.0	-	31.52
ESD [12]	×	42.3 (-2.6)	57.2 (-2.3)	50.6 (-3.2)	26.3 (-0.7)	39.38	31.39
UCE [13]	×	28.4 (-16.5)	29.2 (-30.3)	31.1 (-23.1)	20.7 (-6.3)	78.72	29.09
RECE [14]	×	21.0 (-23.9)	19.9 (-39.6)	24.3 (-29.9)	9.4 (-17.6)	65.94	28.11
SafetyDPO [27]	×	16.2 (-28.7)	25.2 (-34.3)	19.2 (-35.0)	5.5 (-21.5)	51.20	30.59
SAFREE [52]	×	20.9 (-24.0)	42.3 (-17.2)	24.4 (-29.8)	13.3 (-13.7)	45.38	30.51
CASG+SAFREE (ours)	✓	18.0 (-26.9)	37.6 (-21.9)	18.0 (-36.2)	10.8 (-16.2)	48.06	30.25
SLD [39]	×	15.4 (-29.5)	27.4 (-32.1)	19.0 (-35.2)	7.7 (-19.3)	53.32	29.28
CASG+SLD (ours)	✓	**12.3 (-32.6)**	**13.3 (-46.2)**	**14.5 (-39.7)**	**5.0 (-22.0)**	52.29	29.34

Table 8. ASR (%) under different predefined harmful keyword variants on T2VSafetyBench. Values in brackets denote the change relative to the base safeguard.

Keyword set	SAFREE	CASG+SAFREE	SLD	CASG+SLD
Default	41.5	37.5 (-4.0)	25.2	9.8 (-15.4)
Synonyms	36.2	32.3 (-3.9)	14.8	10.6 (-4.2)
Abstract	49.4	40.7 (-8.7)	26.9	18.9 (-8.0)
Detailed	41.1	36.8 (-4.3)	19.1	7.9 (-11.2)

D.3. Category-wise Safety Analysis.

As shown in Table 9, we conduct a category-wise analysis of harmful rate reduction on the I2P dataset, comparing CASG+SAFREE and CASG+SLD against their respective baselines (SAFREE and SLD). This analysis illustrates how safety improvements vary across harmful categories, shedding light on the varying degrees of harmful conflicts. Notably, CASG+SLD achieves substantial reductions in *sexual content* (-61.2%) and *illegal activity* (-26.7%), indicating stronger directional attenuation effects within the aggregation. In contrast, the marginal improvement in *harassment* (-1.0%) suggests that its harmful zone largely overlaps with the aggregated harmful zone. CASG+SAFREE also exhibits varying degrees of improvement across different categories, reflecting the influence of category-specific conflict strength. These results collectively demonstrate that CASG is particularly effective in categories characterized by strong harmful conflicts.

D.4. Visualization Analysis

To further examine how CASG improves safety control, we compute at each timestep the cosine similarity (Equation (4)) between the prompt guidance and the harmful guidance of each category for both SLD and CASG+SLD. A higher cosine value indicates that the prompt guidance is more aligned with that harmful category, thus reflecting stronger harmfulness. We visualize these cosines across timesteps using heatmaps for better clarity and comparative analysis. As shown in Figure 9, although SLD exhibits a slight downward trend in cosine similarity as the timestep increases, its reduction is noticeably weaker compared with CASG+SLD. In contrast, *CASG+SLD exhibits a substantial reduction* in cosine similarities for all categories, indicating stronger suppression of harmful content. These visualizations provide direct evidence that CASG more effectively reduces harmful generation throughout denoising steps.

Table 9. We present category-wise harmful rate analysis of the harmful rate reduction in I2P datasets. Values in brackets denote the relative drop compared with SAFREE and SLD.

Category	CASG+SAFREE	CASG+SLD
hate	15.6 (-5.5)	10.8 (-16.9)
harassment	15.3 (-0.7)	10.1 (-1.0)
violence	19.2 (-9.4)	16.8 (-14.3)
self-harm	14.0 (-10.3)	6.4 (-19.0)
sexual content	22.4 (-9.3)	6.1 (-61.2)
shocking	29.7 (-3.9)	15.2 (-13.6)
illegal activities	13.8 (-16.9)	6.9 (-26.7)

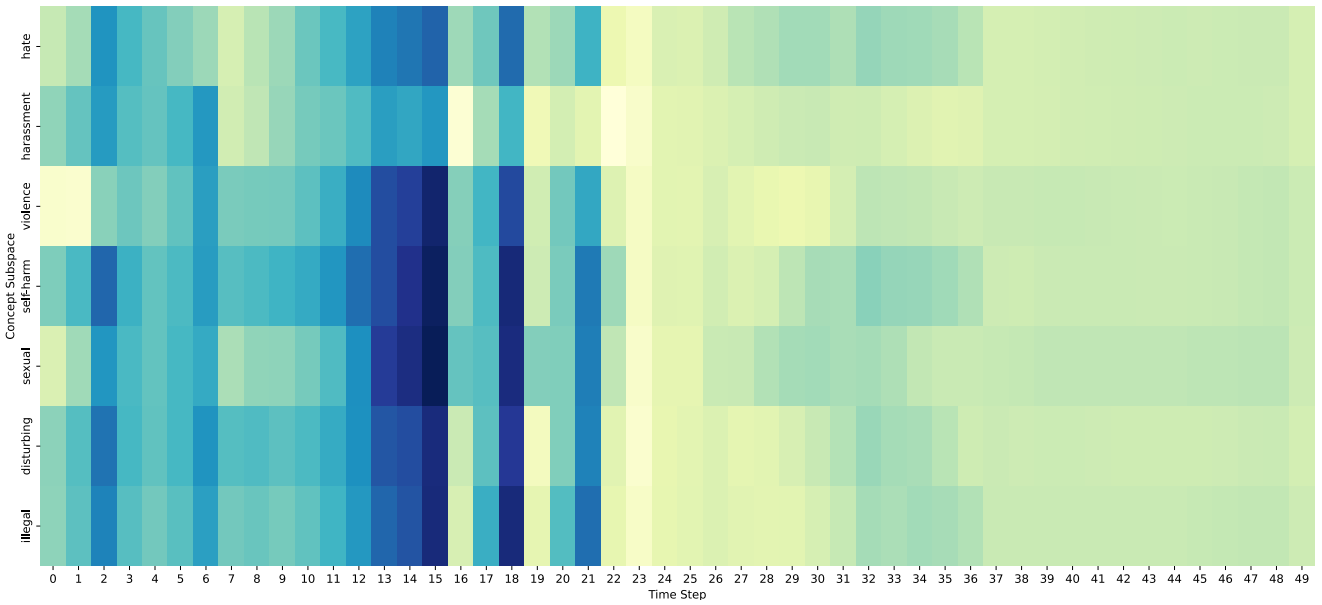
D.5. More Qualitative Results

To provide a comprehensive evaluation of our approach, we present additional qualitative examples in Figure 10. We include a diverse set of samples that showcase various challenging scenarios. These supplementary results further validate our method's effectiveness across different contexts.

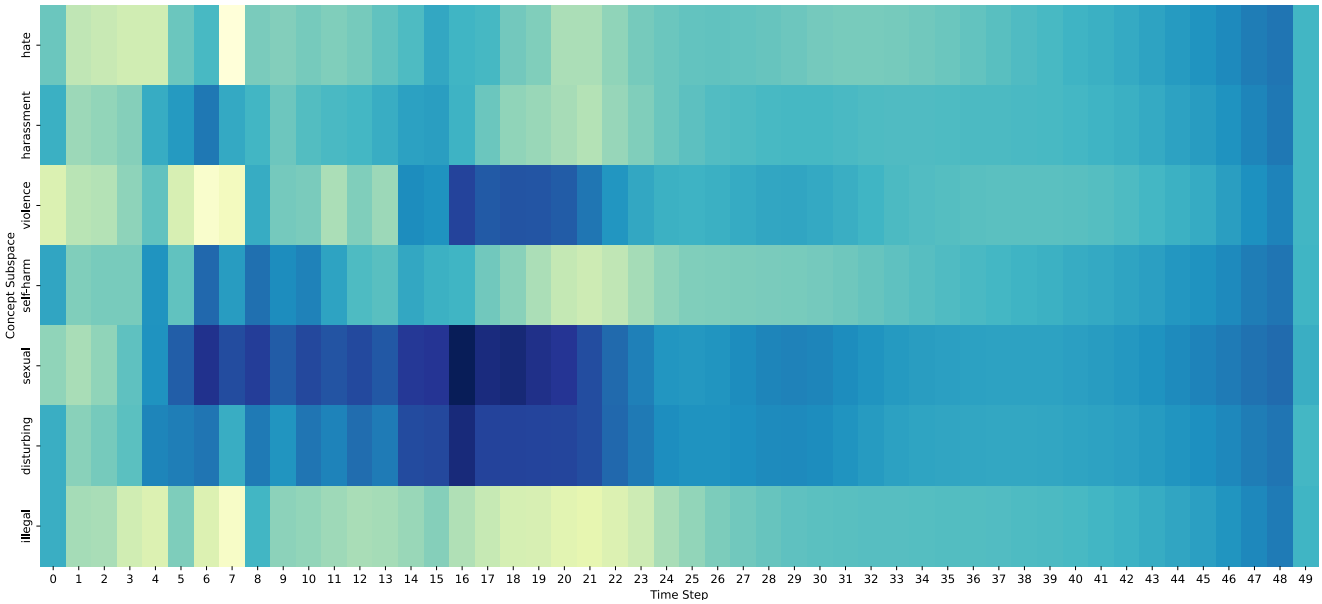
D.6. Efficiency Analysis

To evaluate computational efficiency, we report inference time for all methods under identical settings (50 denoising steps on SDv1.5 using a single H100 GPU). As shown in Table 10, CASG introduces only a **minimal overhead** when integrated into existing training-free safeguards.

• For text-space guidance, CASG+SAFREE increases in-



(a) CASG+SLD



(b) SLD

Figure 9. Harmful cosine in latent space with sexual-related prompt. The horizontal axis shows diffusion timesteps, and the vertical axis lists harmful categories. Color intensity indicates harmful cosine (darker means stronger harmfulness

ference time from 3.9 seconds to just 3.98 seconds per sample, demonstrating that our conflict-aware projection adds *negligible cost*.

- For latent-space guidance, CASG+SLD exhibits a linear growth with respect to the number of predefined harmful categories k, *adding approximately 1 second per additional category per sample*. In our experiments, we follow the original SLD setting and adopt 7 predefined harmful categories, where the inference time reaches 10.2 seconds per sample, which is 2.58 times of SLD.

Overall, CASG preserves the lightweight nature of safety-guidance frameworks while significantly improving safety performance.

E. Impact of CASG on Content Shift

We further analyze the impact of CASG on content shift under different request scenarios.

- *For benign requests,* CASG largely preserves generation quality on benign inputs (e.g., the COCO dataset [25]), achieving CLIP Score and FID comparable to the base

(a) prompt with sexual content

(b) prompt with violence

(c) prompt with illegal activity

Figure 10. Comparison of T2I safety methods using prompts with different harmful content. Methods marked with * require parameter tuning or model modifications.

Table 10. Inference Efficiency Comparison. k denotes the number of predefined harmful categories. Values in brackets denote the inference time multiples relative to SAFREE and SLD.

Method	Inference Time (s/sample)
SAFREE	3.9
CASG+SAFREE	3.98 (1.02×)
SLD	4.0
CASG+SLD (k=2)	5.0 (1.25×)
CASG+SLD (k=3)	6.2 (1.55×)
CASG+SLD (k=4)	7.1 (1.78×)
CASG+SLD (k=5)	8.2 (2.05×)
CASG+SLD (k=6)	9.2 (2.30×)
CASG+SLD (k=7)	10.3 (2.58×)

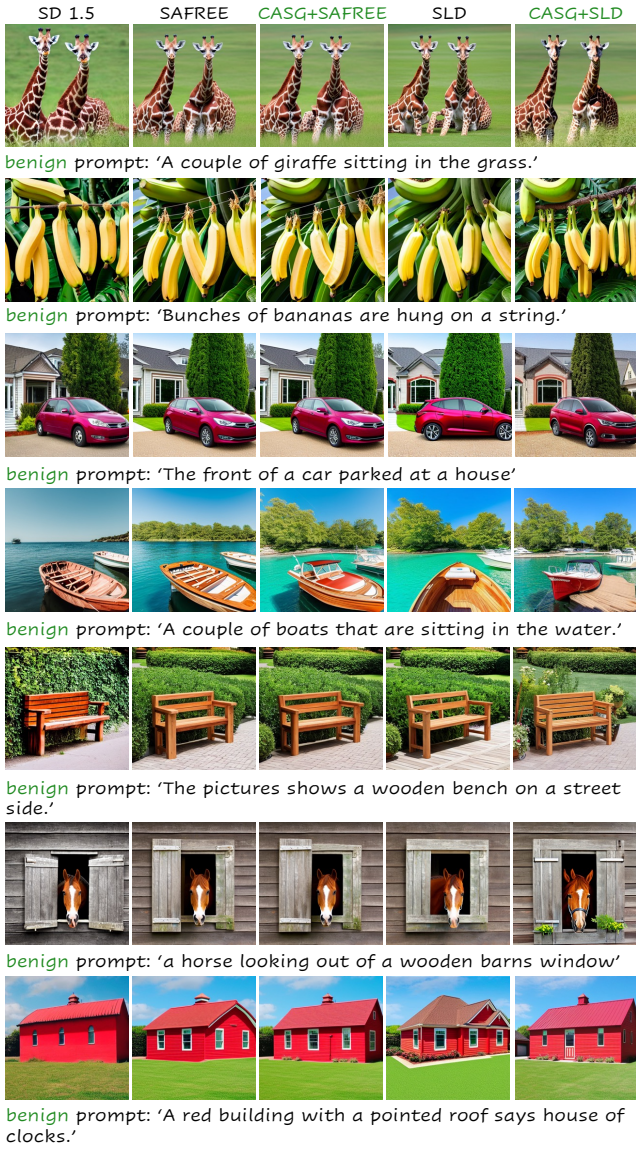


Figure 11. Qualitative comparison between CASG and the base safeguard on benign prompts (COCO).

safeguards (as shown in Tab. 2) with negligible content shift. We provide qualitative benign examples in Fig. 11. These results indicate that CASG preserves normal usability and user experience of text-to-image models.

- *For harmful requests,* CASG may induce content shifts while ensuring enhanced safety, especially when harmful intent is entangled with the core object or presented in adversarial forms [28, 51], where strict semantic preservation for harmful requests can hinder complete removal of harmful content. In real-world deployment, model providers typically prioritize safety guarantees over content fidelity when the input is malicious [32]; accordingly, the semantic shift introduced by CASG under malicious inputs is acceptable.

In general, CASG preserves the ability for benign usage while allowing controlled content shifts on malicious input to ensure safety.

References

[1] Josh Achiam, Steven Adler, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3, 8

[2] Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 396–410, 2023. 1

[3] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546. IEEE, 2021. 1

[4] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint*, 2021. 1

[5] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, et al. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE, 2021. 3

[6] Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. *arXiv preprint arXiv:2309.06135*, 2023. 3

[7] Kai Lai Chung. Markov chains. *Springer-Verlag, New York*, 1967. 3

[8] Yimo Deng and Huangxun Chen. Divide-and-conquer attack: Harnessing the power of llm to bypass the censorship of text-to-image generation model. *CoRR*, 2023. 3

[9] Francisco Eiras, Aleksandar Petrov, Bertie Vidgen, Christian Schroeder De Witt, Fabio Pizzati, Katherine Elkins, Supratik Mukhopadhyay, Adel Bibi, Botos Csaba, et al. Near to mid-term risks and opportunities of open-source generative ai. *arXiv preprint arXiv:2404.17047*, 2024. 1

[10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 3, 4

[11] Dingjie Fu, Wenjin Hou, et al. Discriminative image generation with diffusion models for zero-shot learning. *arXiv preprint arXiv:2412.17219*, 2024. 1

[12] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, et al. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436, 2023. 3, 7, 8, 5, 6

[13] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, et al. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120, 2024. 7, 8, 5, 6

[14] 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*, pages 73–88. Springer, 2024. 3, 4, 7, 8, 5, 6

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint*, 2021. 7

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1, 6

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3

[19] Zhuo Huang, Chang Liu, Yinpeng Dong, Hang Su, Shibao Zheng, and Tongliang Liu. Machine vision therapy: Multimodal large language models can enhance visual robustness via denoising in-context learning. In *Forty-first International Conference on Machine Learning*, 2024. 3

[20] Zhuo Huang, Qizhou Wang, Ziming Hong, Shanshan Ye, Bo Han, and Tongliang Liu. Is gradient ascent really necessary? memorize to forget for machine unlearning. *arXiv preprint arXiv:2602.06441*, 2026. 3

[21] Minseon Kim, Hyomin Lee, Boqing Gong, Huishuai Zhang, and Sung Ju Hwang. Automatic jailbreaking of the text-to-image generative ai systems. *arXiv preprint arXiv:2405.16567*, 2024. 3

[22] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024. 1

[23] Andreas Liesenfeld and Mark Dingemanse. Rethinking open source generative ai: open-washing and the eu ai act. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1774–1787, 2024. 1

[24] Runqi Lin, Alasdair Paren, Suqin Yuan, Muyang Li, Philip Torr, Adel Bibi, and Tongliang Liu. Force: Transferable visual jailbreaking attacks via feature over-reliance correction. *arXiv preprint arXiv:2509.21029*, 2025. 3

[25] 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 *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 7

[26] Yexiong Lin, Yu Yao, and Tongliang Liu. Beyond optimal transport: Model-aligned coupling for flow matching. *arXiv preprint arXiv:2505.23346*, 2025. 1

[27] Runtao Liu, Chen I Chieh, Jindong Gu, et al. Safety-dpo: Scalable safety alignment for text-to-image generation. *arXiv preprint arXiv:2412.10493*, 2024. 3, 7, 8, 4, 5, 6

[28] Yibo Miao, Yifan Zhu, Lijia Yu, Jun Zhu, Xiao-Shan Gao, et al. T2vsafetybench: Evaluating the safety of text-to-video generative models. *Advances in Neural Information Processing Systems*, 37:63858–63872, 2024. 3, 7, 4, 9

[29] notAI tech. Nudenet, 2024. 7, 3, 5

[30] OpenAI. DALL E 3, 2023. 1

[31] Ville Paananen, Jonas Oppenlaender, and Aku Visuri. Using text-to-image generation for architectural design ideation. *International Journal of Architectural Computing*, 22(3): 458–474, 2024. 1

[32] Emmanouil Papagiannidis, Patrick Mikalef, and Kieran Conboy. Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, 34(2):101885, 2025. 9

[33] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901. 4, 1, 2

[34] 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*, pages 3403–3417, 2023. 3, 7, 4

[35] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, et al. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 1

[36] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022. 3

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3, 8, 4, 6

[38] Patrick Schramowski, Christopher Tauchmann, et al. 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*, pages 1350–1361, 2022. 7, 3, 5

[39] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023. 1, 2, 3, 4, 6, 7, 8, 5

[40] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 1

[41] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015. 3

[42] Gilbert Strang. *Linear algebra and its applications*. 2012. 5

[43] Vu Tuan Truong, Luan Ba Dang, and Long Bao Le. Attacks and defenses for generative diffusion models: A comprehensive survey. *ACM Computing Surveys*, 57(8):1–44, 2025. 3

[44] 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? *arXiv preprint*, 2023. 1, 3

[45] Zhenchen Wan, Yanwu Xu, Zhaoqing Wang, Feng Liu, Tongliang Liu, and Mingming Gong. Ted-viton: Transformer-empowered diffusion models for virtual try-on. *arXiv preprint arXiv:2411.17017*, 2024. 1

[46] Zhenchen Wan, Yanwu Xu, Dongting Hu, Weilun Cheng, Tianxi Chen, Zhaoqing Wang, Feng Liu, Tongliang Liu, and Mingming Gong. Mft-viton: High-fidelity virtual try-on with minimal input via a mask-free transformer-diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1985–1994, 2025. 1

[47] Jiepeng Wang, Zhaoqing Wang, Hao Pan, Yuan Liu, Dongdong Yu, Changhu Wang, and Wenping Wang. Mmgen: Unified multi-modal image generation and understanding in one go. *arXiv preprint arXiv:2503.20644*, 2025. 1

[48] Yingxu Wang, Shiqi Fan, Mengzhu Wang, Siyang Gao, Chao Wang, and Nan Yin. Damr: Efficient and adaptive context-aware knowledge graph question answering with llm-guided mcts. *arXiv preprint arXiv:2508.00719*, 2025. 3

[49] Zhaoqing Wang, Xiaobo Xia, Runnan Chen, Dongdong Yu, Changhu Wang, Mingming Gong, and Tongliang Liu. Lavindit: Large vision diffusion transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20060–20070, 2025. 1

[50] Yongli Xiang, Ziming Hong, Lina Yao, et al. Jailbreaking the non-transferable barrier via test-time data disguising. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30671–30681, 2025. 3

[51] 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*, pages 7737–7746, 2024. 3, 9

[52] Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. Safree: Training-free and adaptive guard for safe text-to-image and video generation. *arXiv preprint arXiv:2410.12761*, 2024. 1, 2, 3, 4, 6, 7, 8, 5

[53] Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, pages 385–403, 2024. 1

[54] Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, et al. Qwen3guard technical report. *arXiv preprint arXiv:2510.14276*, 2025. 3, 8

[55] Bowen Zheng, Yongli Xiang, Ziming Hong, Zerong Lin, Chaojian Yu, Tongliang Liu, and Xinge You. Vii: Visual instruction injection for jailbreaking image-to-video generation models. *arXiv preprint arXiv:2602.20999*, 2026. 1