Image Generation from Contextually Contradictory Prompts

Saar Huberman[1] Or Patashnik[1] Omer Dahary[1] Ron Mokady[2] Daniel Cohen-Or[1]
[1]Tel Aviv University [2]BRIA AI

https://tdpc2025.github.io/SAP/

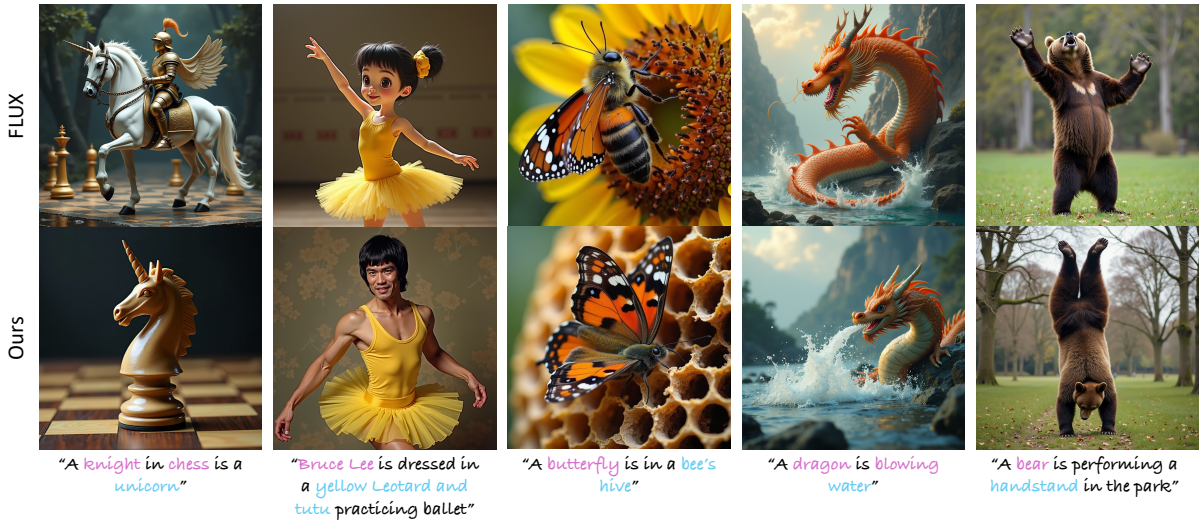


Figure 1. *Our method addresses contextual contradictions in text-to-image generation. These contradictions arise when one concept implicitly conflicts with another due to the model's learned associations. For example, if a concept like "butterfly" is strongly entangled with "flowers", it may conflict with another concept in the prompt such as "bee's hive", leading the model to ignore or distort part of the semantic meaning.*

Abstract

Text-to-image diffusion models excel at generating high-quality, diverse images from natural language prompts. However, they often fail to produce semantically accurate results when the prompt contains concept combinations that contradict their learned priors. We define this failure mode as contextual contradiction, where one concept implicitly negates another due to entangled associations learned during training. To address this, we propose a stage-aware prompt decomposition framework that guides the denoising process using a sequence of proxy prompts. Each proxy prompt is constructed to match the semantic content expected to emerge at a specific stage of denoising, while ensuring contextual coherence. To construct these proxy prompts, we leverage a large language model (LLM) to analyze the target prompt, identify contradictions, and generate alternative expressions that preserve the original intent while resolving contextual conflicts. By aligning prompt information with the denoising progression, our method enables fine-grained semantic control and accurate image generation in the presence of contextual contradictions. Experiments across a variety of challenging prompts show substantial improvements in alignment to the textual prompt.

1. Introduction

Text-to-image diffusion models have demonstrated remarkable capabilities in generating high-quality and diverse visual content from natural language prompts [15, 24, 27]. However, achieving precise semantic alignment between the generated image and the conditioning prompt remains an open challenge. This issue becomes particularly pronounced when the target prompt lies outside the model's training distribution, such as prompts that combine semantically plausible yet statistically uncommon or unconventional concepts. For example, as shown in Figure 1, generating an image from the prompt *"A butterfly is in a bee's hive"* often results in a butterfly on a flower. This is due to the model's prior that entangles butterflies with flowers, which implicitly contradicts the notion of a bee's hive.

We refer to this phenomenon as *Contextual Contradiction*, a conflict between two concepts that arises not from direct semantic opposition, but from the model's associations learned during training. More precisely, we say that concept A contextually contradicts concept B if the model's prior entangles A with concept C, and B contradicts C (see Figure 2). In Figure 1, we illustrate this with a blowing dragon, which stands in contextual contradiction with the water due to its entanglement with fire.

The phenomenon of contextual contradiction in text-to-

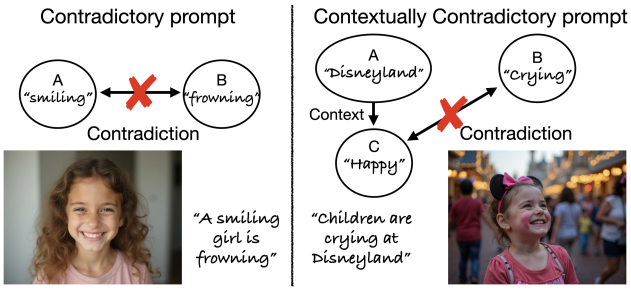


Figure 2. *On the left is a direct contradiction, since a girl cannot smile and frown at the same time. On the right is a contextual contradiction: while Disneyland and crying are not directly opposed, the model's prior associates Disneyland with happiness, which conflicts with crying.*

image models relates to the broader issue of spurious correlations in deep learning. Models often exploit shortcuts, relying on correlations in the training data that are statistically strong but semantically misleading [13]. In this paper, we identify a similar bias in generative models: contextual contradictions occur when prompts combine concepts that individually align with the model's priors but conflict when combined, revealing the model's reliance on such correlations rather than robust compositional reasoning.

To address this issue, we introduce *Stage-Aware Prompting (SAP)*, which builds on the observation that the denoising process follows a coarse-to-fine progression, during which different semantic attributes (e.g., background, pose, shape, and texture) emerge at distinct stages [4, 9, 23]. Our key idea is to guide the model at each stage of denoising with the information most relevant to the type of content being formed at that point. To achieve this, we decompose the original prompt into a sequence of proxy prompts, each aligned with the attributes expected at a specific stage and designed to avoid contextual contradictions.

Ensuring that proxy prompts preserve the original intent while avoiding contextual contradictions requires a broad understanding of the real world. For example, it involves understanding that a bear is entangled with specific poses, such as walking on all fours or standing upright, which in turn, contradicts the handstand pose. To achieve this, SAP leverages a large language model (LLM) to analyze the target prompt, identify contextual contradictions, and construct suitable proxy prompts.

We demonstrate that, by using in-context examples and prompting the LLM to follow a reasoning process through a brief explanation, it can identify contextually contradictory concepts in a prompt and determine the appropriate stage of denoising at which each attribute should be introduced. It can also suggest alternative, non-conflicting concepts that preserve the intended attributes and use them to construct stage-specific proxy prompts. In doing so, the LLM effectively guides the model toward the intended meaning of the original prompt.

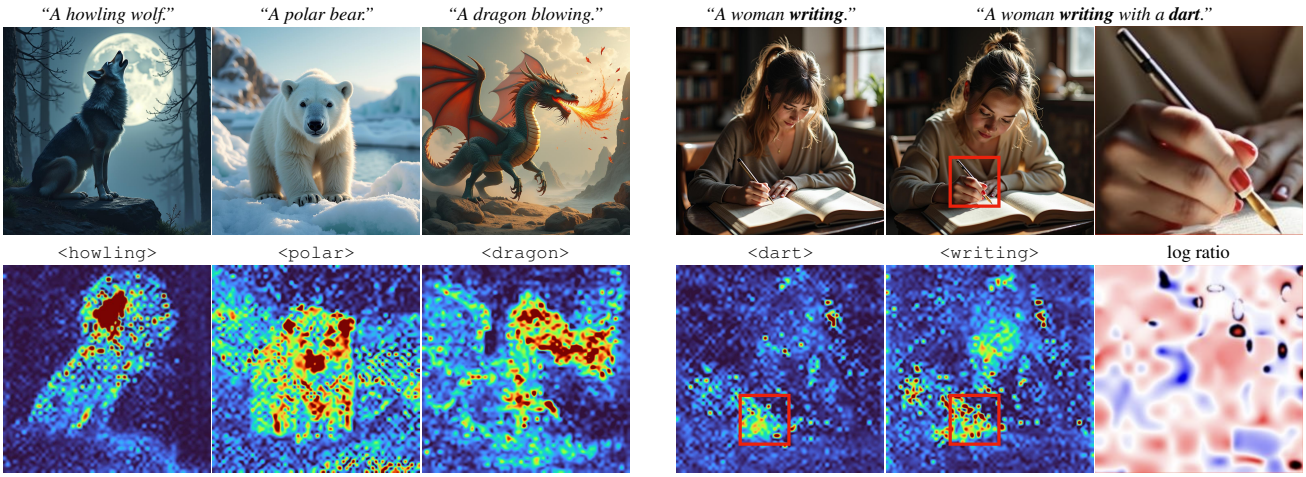
Through extensive experiments, we demonstrate the effectiveness of SAP in generating images from contextually contradictory text prompts. By introducing prompt information at targeted stages, SAP generates precise combinations of semantic attributes while avoiding undesired entanglement. Compared to previous methods, SAP's stage-dependent prompt decomposition leads to more faithful and semantically aligned generations.

2. Related Work

Learned Spurious Correlations. Machine learning models are known to be sensitive to spurious correlations in their training data [13, 20, 33], leading to performance drops when training-time associations do not hold at test time. Prior work has extensively studied this in discriminative vision tasks, showing, for example, that recognition models tend to rely heavily on background cues [6, 28, 31]. In our work, we show that text-to-image models exhibit similar behavior. When given contextually contradictory prompts – combinations of concepts that conflict with correlations seen during training – diffusion models often fail to generate images that accurately reflect the prompt. We evaluate this behavior using the Whoops! dataset [7], which contains prompts constructed by first describing two co-occurring elements, and then replacing one with a less compatible alternative. This results in scenarios that are unlikely to occur in the real world.

Semantic Alignment in Text-to-Image Synthesis. Text-to-image models often struggle to fully capture the semantic intent of input prompts, particularly when prompts involve complex or internally conflicting concepts. Previous works have analyzed common failure cases and proposed targeted improvements across various stages of the generation pipeline, including enhanced text embedding representations [12, 25, 29], refined attention mechanisms [10], guidance strategies that leverage attention maps for loss heuristics [2, 9, 21, 26] and dynamic guidance scheduling via annealed classifier-free guidance [34]. Despite these advances, existing methods often fail to handle prompts containing contradictory concepts arising from the model's learned associations. Our work directly addresses this underexplored challenge, focusing on contextually contradictory prompts.

Multi-Prompt Conditioning Techniques. Conditioning diffusion models on multiple prompts has emerged as an effective strategy for improving control and compositionality. One line of work, primarily focused on personalization, introduces distinct learned tokens at different layers of the model and at various denoising timesteps [3, 30]. This design allows each token to capture different attributes of the personalized concept, leading to improved identity preservation. Other approaches vary the prompt across timesteps



(a) Entangled concepts. (b) From entanglement to contradiction.

Figure 3. *(a) By examining attention maps, we observe that textual tokens embed contextual associations, leading to the generation of concepts not explicitly mentioned in the prompt. For example, the token 'howling' encourages the presence of the moon, as indicated by its strong attention connection. (b) In prompts with contextual contradictions, two tokens may overlap in attending to the same region, as seen with 'dart' and 'writing'. The token 'writing' dominates this region, as shown in the log-ratio map, where red areas indicate stronger attention to 'writing' relative to 'dart'.*

to modulate specific visual properties, such as object shape [18, 23], or alternate between rare and frequent object descriptions to improve attribute binding [22]. Fine-grained spatial control has also been achieved by assigning subprompts to separate image regions [32]. Additionally, some methods leverage multiple diffusion models, each conditioned on different prompt attributes, and combine their outputs into a unified prediction [5, 19]. Unlike most prior works, we focus on utilizing multiple prompts to settle internal semantic tension, where concept combinations lead to contextual contradictions.

LLM-Guided Diffusion. LLMs have demonstrated strong capabilities in language understanding. They also capture broad world knowledge through large-scale training on diverse text. Recent approaches have leveraged these capabilities to guide diffusion model generation, often incorporating planning and reasoning to improve semantic alignment [16, 22, 32]. In our work, we employ LLMs with in-context learning to identify contextual contradictions, generate proxy prompts, and determine the corresponding timestep ranges for conditioning, while encouraging reasoning through brief explanatory outputs.

3. From Entangled Concepts to Contextual Contradictions

Diffusion models inherit strong distributional biases from their training data, where objects are frequently tied to specific contexts. For example, prompts like *"a duck"* almost always result in water backgrounds, and *"a polar bear"* appears in snow. These reflect *entangled concepts*, learned associations that go beyond the explicit text. While often helpful, such priors hinder generation when prompts require unusual or contradictory combinations.

To study this effect, we analyze attention maps (Figure 3a), which indicate the spatial regions influenced (attended) by each token. We find that text tokens often attend not only to the image regions directly corresponding to the object but also to contextually linked elements. For example, in *"a howling wolf"*, the token 'howling' influences both the mouth and the moon. Similarly, 'dragon' attends to flames even when fire is not mentioned. These patterns reveal that the model encodes distributional correlations beyond literal semantics.

These distributional correlations contribute to the difficulty of generating images with *contextual contradictions*. For example, when generating an image from the prompt *"a woman writing with a dart"*, the model fails to replace the pen with a dart. The attention maps shed light on this failure: both 'writing' and 'dart' attend to the same spatial area (the hand/tool), but 'writing' dominates (see log-ratio maps in Figure 3b), suppressing the influence of 'dart'. This reflects a failure mode in which entrenched associations override less familiar ones, preventing proper integration of conflicting concepts.

4. Stage Aware Prompting (SAP)

In this section, we begin by analyzing the coarse-to-fine behavior of the denoising process (Section 4.1). Building on the insights from our analysis, we introduce our training-free framework for resolving contextual contradictions in text-to-image generation. As illustrated in Figure 5, our approach consists of two main components: (i) prompt decomposition (Section 4.2, top part of the figure), and (ii)

Figure 4. *Coarse-to-fine denoising with stage-aware prompting. We show x_0 predictions at initial, intermediate, and final steps.* `A` *is the target prompt,* `B` *a suitable proxy, and* `C` *an unsuitable proxy. Using* `A` *alone locks in night and moon despite "midday". Using* `B` *first and then switching to* `A` *preserves a daytime layout and later adapts the identities to wolf and bat. Using* `C` *first sets a layout without a flying object, so the final image fails to produce the intended subjects.*

stage-aware prompt injection (Section 4.3, bottom part of the figure). In the following, we describe each of these components.

4.1. Coarse-to-Fine Denoising

Diffusion models generate images in a coarse-to-fine manner: early steps determine low-frequency structures, while high-frequency details emerge in later steps [4, 9, 14, 23]. From this behavior, we draw two key observations: (i) *Irreversibility of details.* As denoising progresses, the model sequentially commits to different levels of detail, beginning with layout and overall shape. Once these are formed, they cannot be revised in later stages, even if they conflict with the prompt. (ii) *Flexibility in high-frequency details.* In early stages, high-frequency details are absent and unaffected by the prompt, enabling flexible guidance without influence on fine details.

As shown in Section 3, contextual contradictions stem from concept entanglement in the diffusion prior. Since different concepts emerge at different levels of detail during the coarse-to-fine denoising process, they can be decomposed across the denoising stages. We illustrate this in Figure 4 by examining the model's x_0 predictions across denoising steps. In the top-left row, the prompt *"a howling wolf and a flying bat at midday"* shows that early steps already impose entangled nighttime and moon structures, contradicting "midday". In the bottom-left row, starting with a *proxy prompt* containing "dog" and "bird" (instead of "wolf" and "bat", respectively) and later switching to the target prompt produces a correct midday scene with the intended objects. The bottom-left row demonstrates both the *irreversibility of coarse details* (the scene remains daytime) and the *flexibility of high-frequency details* (the object identities adapt). The top-right row highlights the importance of selecting an appropriate proxy prompt. Since rats do not fly, the layout determined in early steps lacks a flying object, resulting in a sitting wolf with bat wings.

These observations motivate a stage-aware prompting strategy that fixes structural decisions early while keeping later attributes flexible. Building on this, we now describe the two main components of our method: prompt decomposition and stage-aware prompt injection.

4.2. Prompt Decomposition

Given a prompt P that contains contextually contradictory concepts, we aim to construct a sequence of proxy prompts $\{p_1, p_2, ..., p_n\}$ and corresponding timestep intervals $\{I_1, I_2, ..., I_n\}$ that together reflect the intended semantics of P. Each proxy prompt p_i is designed to (i) preserve the relevant semantics of P for attributes typically formed during its interval I_i, which is crucial due to the *irreversibility of details*, and (ii) avoid contradictions likely to emerge at that stage, leveraging the *flexibility in high-frequency details*. This decomposition conditions the diffusion model on contextually coherent content that evolves in tandem with the coarse-to-fine denoising process.

To generate proxy prompts and their intervals, we use a large language model (LLM) that detects contextual contradictions and proposes suitable substitutes for conflicting concepts. It also infers the appropriate staging of these concepts within the proxy prompts. We implement this using a structured prompt template containing instructional text, in-context examples, and explanations of contextual contradictions. The examples demonstrate both successful decompositions and cases requiring no decomposition, enabling the LLM to generalize. The full instruction prompt is provided in the Appendix.

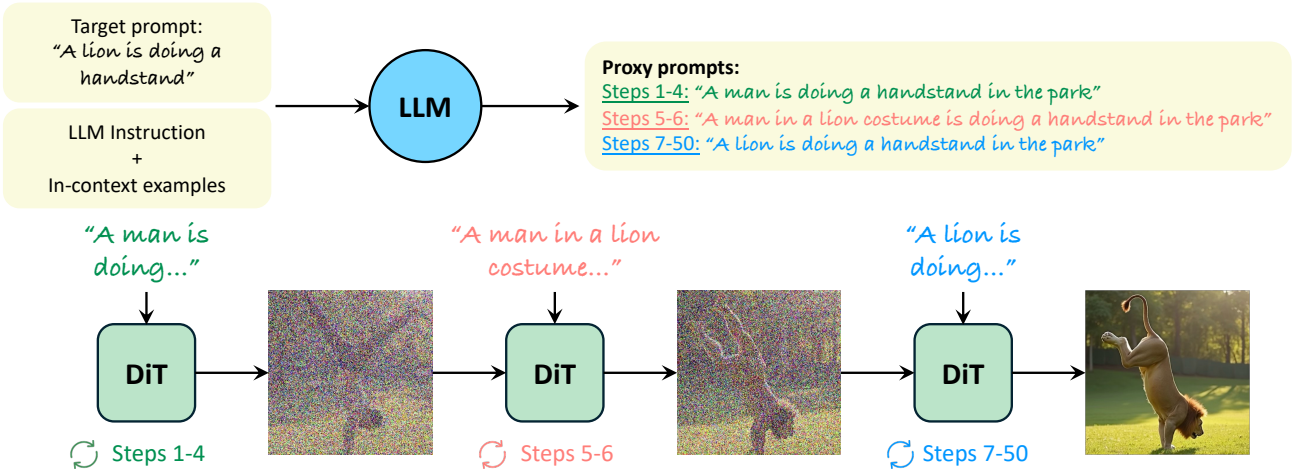


Figure 5. *SAP generates images from contextually contradictory prompts using time-dependent proxy prompts. (Top) A large language model (LLM) decomposes the target prompt into a sequence of proxy prompts with corresponding timestep intervals. (Bottom) These proxy prompts are injected into the diffusion process at their designated intervals to guide generation.*

In-context Examples. Our in-context examples take a target prompt as input and output proxy prompts with timestep intervals, along with a brief explanation of the contradiction. Requiring the LLM to provide this explanation encourages reasoning about conflicts and ensures coherent substitutions. These examples were created by identifying contextually contradictory prompts that fail under the base model (FLUX) and manually decomposing them into proxy prompts with corresponding intervals. The explanations were auto-generated by an LLM. We include 20 examples that demonstrate diverse strategies for handling contradictions. We next elaborate on one of the most frequent strategies in our method.

Concept Substitution. In this strategy, a conflicting concept is temporarily replaced with a structurally appropriate proxy (Figure 4). A simpler alternative is to omit the conflicting concept, but introducing an object only in later stages without a placeholder can distort its perceived size or cause it to be omitted entirely. In Figure 6, we demonstrate this by comparing decompositions that differ only in when the second interval begins. The first proxy specifies the background, while the second adds the foreground. Introducing the foreground early allows the model to allocate more space, whereas delaying it constrains the layout and produces a smaller object. Substitution resolves these issues and yields stable layouts. In Figure 7, we show the effect of misplacing intervals across denoising stages. Using two proxies with different intervals, where the second is the full prompt, we observe two failure modes: introducing the full prompt too early prevents disentangling contradictions, while introducing it too late alters only fine details. Earlier in Figure 4 we illustrated the importance of careful proxy selection.

4.3. Stage-aware Prompt Injection

Given a sequence of proxy prompts $\{p_1, p_2, ..., p_n\}$, their corresponding timestep intervals $\{I_1, I_2, ..., I_n\}$, and a text-to-image (T2I) diffusion model, we condition the model using different prompts throughout the denoising process. At each timestep t, we apply the prompt p_i such that $t \in I_i$. By aligning each proxy prompt with its interval, the denoising process is guided by concepts appropriate to the level of detail emerging at that stage. This enables gradual image construction while avoiding conflicts with the model's learned priors. The injection mechanism integrates seamlessly into existing inference pipelines without architectural modifications and is compatible with a range of pretrained diffusion models.

5. Experiments and Results

In this section, we evaluate SAP through qualitative (Section 5.1) and quantitative (Section 5.2) comparisons. We further conduct ablation studies (Section 5.3) on component contributions and robustness, followed by a discussion on the limitations of our method.

Implementation Details. We use FLUX.1 [dev] [17] as the base T2I model and GPT-4o [1] for prompt decomposition. In all experiments, inference is performed using $T = 50$ steps and the LLM is restricted to at most three proxies per prompt. Baseline hyperparameters follow their original papers or implementations, with $T = 50$ steps for fair comparison. To further demonstrate robustness, we also report results with SD3.0 using the same LLM-generated proxy prompts and intervals.

Baselines. We compare SAP against the following approaches: (1) base models FLUX-dev (denoted by

Figure 6. *Early insertion of the foreground allows the model to allocate more space to it, while late insertion confines the object within the existing layout, making it appear smaller (e.g., the snowman introduced at Step 5). Step labels indicate when the foreground object is introduced.*

Figure 7. *Effect of interval assignment. Introducing the full prompt too early fails to disentangle contextual contradictions, while introducing it too late alters only fine details. Top proxy: "A pillow fort in a bedroom"; bottom proxy "A mother duck guards three ducklings".*

FLUX) [17] and SD3.0 [11]; (2) R2F [22], a training-free method reported under three settings: SD3.0 (original), FLUX-schnell (official), and our reimplementation on FLUX; (3) Annealing Guidance [34], which trains a small MLP to predict the classifier-free guidance scale at each step; and (4) Ella [16], a fine-tuned model on SD1.5. (5) VL-DNP [8], a VLM-guided negative prompting method, evaluated using SD3 with Qwen2.5-VL-7B.

Datasets We evaluate SAP using three datasets: Whoops! [7], Whoops-Hard, and ContraBench. *Whoops!* consists of 500 prompts paired with commonsense-defying images, designed to test visual reasoning and compositional understanding. While relevant to our task, many of its prompts are relatively easy for modern T2I models and do not consistently expose model limitations.

To address this, we curate *Whoops-Hard*, a subset of 100 particularly difficult prompts from Whoops!, providing a stronger benchmark for evaluating current state-of-the-art models. To further probe semantic alignment under contradictory conditions, we introduce *ContraBench*, a curated set of 40 prompts capturing contextual contradictions. The dataset was constructed in two steps: (1) ChatGPT generated candidate prompts based on the definition of contextual contradictions, and (2) human annotators filtered them to retain only those that clearly expressed contradictions. The full prompt lists for both datasets are provided in the Appendix.

5.1. Qualitative Results

Figure 8 presents qualitative comparisons on the Whoops! and ContraBench datasets (see Appendix for more results). Across both benchmarks, baseline methods consistently exhibit characteristic failure modes when handling contradictory prompts. In contrast, SAP successfully generates challenging cases such as a blue Shrek or a monkey juggling tiny elephants (Figure 8). In addition to FLUX, SAP also improves SD3 generations under contradictory prompts (Figure 9).

For SD3 and FLUX, contradictory prompts expose conflicts with learned priors, resulting in prompt misalignment. Ella and Annealing Guidance, not designed for contradictions, perform less effectively on such cases. R2F alternates between prompts at predefined timesteps, a strategy designed for attribute binding rather than addressing contextual contradictions. While it can reinforce rare concepts, it does not align prompts with the stages at where semantic features emerge during denoising. As a result, it often produces hybrid concepts that merge incompatible elements from conflicting concept (see the bodybuilder in Figure 9, and the owl and SpongeBob results in the Appendix).

In contrast, SAP produces semantically coherent outputs by introducing proxy prompts at denoising stages where corresponding features emerge. This enables effective handling of conflicting concepts. Across both benchmarks, SAP consistently generates coherent images for contradictory prompts. Robustness is further demonstrated in the Appendix, comparing generations across multiple seeds.

5.2. Quantitative Results

We evaluate prompt alignment using GPT-4o vision–language model (VLM). For each generated image, GPT-4o assigns a score from 1–5 based on adherence to

Figure 8. *Qualitative comparison of SAP with baseline methods. Our method resolves contextual contradictions, whereas baselines struggle to produce text-aligned images. Additional examples are provided in the Appendix.*

Figure 9. *SAP is robust to the base model, as shown by the results obtained with SAP$_{SD3}$.*

the prompt. Scores are averaged across three fixed random seeds per prompt and scaled to 20–100. The evaluation prompt is provided in the Appendix.

As shown in Table 1, SAP outperforms all baselines across the three benchmarks. Between the base models, SD3.0 tends to yield stronger alignment under contradictions, while FLUX offers higher visual quality (Figure 8). SAP improves both backbones, enhancing prompt adherence while maintaining the visual fidelity of the underlying models (Figures 8 and 9).

User study. VLM-based metrics often miss subtle semantic inconsistencies and do not adequately assess image quality. To complement them, we conducted a user study evaluating both prompt adherence and overall visual appearance. We randomly sampled 24 prompts from the Whoops! and ContraBench benchmarks. For each prompt, participants compared two images, one generated by SAP and the other by a baseline, and answered: (1) which most accurately reflected the prompt, and (2) which had higher visual quality. In total, we collected responses from 61 users, yielding 1,464 individual evaluations. Table 2 summarizes the win rates of SAP against each baseline.

These results highlight the superiority of our approach in handling contextually contradictory prompts, achieving both stronger prompt alignment and higher visual quality.

5.3. Ablation Studies

We evaluate SAP through ablations that assess design choices in prompt decomposition and robustness under different conditions, such as timestep perturbations. Additional results on alternative LLMs and non-contradictory prompts are in the Appendix.

Prompt decomposition components. We conduct ablations on the Whoops-Hard benchmark, where each variant isolates a design choice to quantify its effect on alignment within our method (Table 3). In-context examples significantly improve the LLM's ability to decompose contradictory prompts, leading to better text–image alignment. Removing the explanation requirement impairs reasoning and causes a notable drop, showing that generating explicit explanations encourages more coherent semantic decisions. Restricting decomposition to two proxies performs close to

Table 1. *Quantitative evaluation on various benchmarks using the GPT-4o vision-language model. We report average alignment, where alignment reflects how well the image matches the prompt semantics, independent of visual quality. SAP achieves the best results, regardless of the base model.*

Models	Benchmarks		
	Whoops	Whoops-Hard	Contra-Bench
SD3.0	82.63	55.73	57.5
FLUX	78.85	44.3	57.16
Ella	69.09	45.19	55.16
Annealing	79.59	59.06	58.33
VL-DNP	82.79	56.26	58.83
R2F	83.50	57.06	59.16
$R2F_{schnell}$	79.58	54.80	59.33
$R2F_{FLUX}$	48.68	32.80	25.33
$SAP_{SD3.0}$	**85.87**	**64.40**	65.33
SAP	85.10	62.13	**66.16**

Table 2. *User study results. Win rates of SAP in text-image alignment and image quality, compared against each baseline method.*

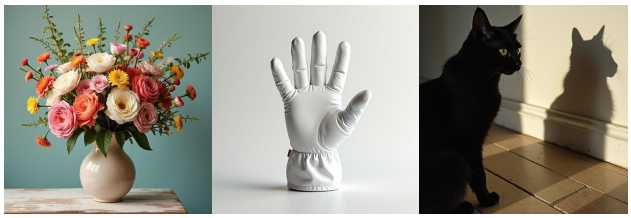
	SD3	FLUX	Ella	Annealing	R2F
Alignment	70%	81%	81%	73%	75%
Quality	72%	63%	74%	79%	68%

Table 3. *Ablation on Whoops-Hard. We evaluate our prompt decomposition method by (1) removing in-context examples, (2) removing the explanation requirement, and (3) limiting decomposition to two proxy prompts.*

	static	w/o in-context	w/o reasoning	2 proxy	Full
Alignment	44.3	48.0	46.46	60.26	62.13

the full method, while allowing up to three proxies provides extra flexibility for harder cases and yields further gains.

Robustness to LLM-predicted timestep intervals. Our method relies on LLM-predicted intervals to schedule proxy prompts, but these boundaries do not require exact placement. The earlier results (Figure 7) highlight that placing proxy prompts at the wrong *stage* of denoising (e.g., too early or too late) can harm alignment. Here we show that within the correct coarse stage, the method is robust to moderate boundary shifts. Specifically, we perturb interval boundaries while keeping the proxy prompts fixed, uniformly shifting them within windows of varying size. As shown in Table 4, small shifts of up to ± 1 step (window=3) have almost no effect on alignment, and even larger shifts of up to ± 2 steps (window=5) cause only minor degradation, despite representing a substantial perturbation relative to the full method effective range ($\sim$12 steps; see Appendix). These results confirm that SAP is sensitive to the stage at which information is introduced, but largely insensitive to exact step boundaries within that stage.



"A bouquet of flowers is upside down in a vase" *"A white glove has 6 fingers"* *"The shadow of a cat is facing the opposite direction"*

Figure 10. Failure cases of SAP.

Table 4. *Effect of perturbing LLM-predicted timestep intervals. Boundaries are uniformly shifted within window w. $SAP_{w=i}$ denotes evaluation with window i.*

Models	Benchmarks		
	Whoops	Whoops-Hard	Contra-Bench
FLUX	78.85	44.3	57.16
SAP	85.10	62.13	66.16
$SAP_{w=3}$	84.18	62.06	65.5
$SAP_{w=5}$	81.46	58.46	62.5

Limitations. In Figure 10, we present representative failure cases of our method on examples from the Whoops! benchmark. Since our approach relies on guiding the model through text alone, it cannot control properties that the underlying model inherently struggles with, such as generating specific quantities or enforcing precise orientations; a more detailed categorization and analysis of failure modes is provided in the supplementary.

6. Conclusions

We introduced a training-free framework for resolving contextual contradictions in text-to-image generation, cases where seemingly plausible prompts fail due to strong, hidden model biases. At its core, our method leverages the coarse-to-fine generation process to separate contradictions across denoising stages, enabling faithful rendering of prompts that would otherwise yield semantically inconsistent outputs. The introduction of proxy prompts steers the generative process in line with the model's priors, enabling it to resolve conflicts and preserve semantic fidelity without the need for retraining or fine-tuning.

We argue that since our approach already leverages the broad world knowledge of vision–language models, integrating them more tightly with generative models holds promise for addressing contextual contradictions directly. As a next step, we plan to explore emerging compound architectures that combine VLMs and generative models, with the aim of understanding how to effectively harness them to resolve conflicts in open-ended generation.

Acknowledgments

This research was supported in part by the Israel Science Foundation (grants no. 2492/20 and 1473/24), Len Blavatnik and the Blavatnik family foundation.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5

[2] Aishwarya Agarwal, Srikrishna Karanam, KJ Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2283–2293, 2023. 2

[3] Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization. *ACM Transactions on Graphics (TOG)*, 42(6):1–10, 2023. 2

[4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2, 4

[5] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023. 3

[6] Sara Beery, Grant Van Horn, and Pietro Perona. *Recognition in Terra Incognita*, page 472–489. Springer International Publishing, 2018. 2

[7] Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 2616–2627. IEEE, 2023. 2, 6

[8] Hoyeon Chang, Seungjin Kim, and Yoonseok Choi. Dynamic vlm-guided negative prompting for diffusion models. *arXiv preprint arXiv:2510.26052*, 2025. 6

[9] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023. 2, 4

[10] Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. Be yourself: Bounded attention for multi-subject text-to-image generation. In *European Conference on Computer Vision*, pages 432–448. Springer, 2024. 2

[11] 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. 6

[12] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. 2

[13] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 2

[14] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 4

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. 1

[16] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 3, 6

[17] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 5, 6

[18] Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. Magicmix: Semantic mixing with diffusion models. *arXiv preprint arXiv:2210.16056*, 2022. 3

[19] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022. 3

[20] Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019. 2

[21] Tuna Han Salih Meral, Enis Simsar, Federico Tombari, and Pinar Yanardag. Conform: Contrast is all you need for high-fidelity text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9005–9014, 2024. 2

[22] Dongmin Park, Sebin Kim, Taehong Moon, Minkyu Kim, Kangwook Lee, and Jaewoong Cho. Rare-to-frequent: Unlocking compositional generation power of diffusion models on rare concepts with llm guidance. *arXiv preprint arXiv:2410.22376*, 2024. 3, 6

[23] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 23051–23061, 2023. 2, 3, 4

[24] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 1

[25] Royi Rassin, Shauli Ravfogel, and Yoav Goldberg. Dalle-2 is seeing double: Flaws in word-to-concept mapping in text2image models. *arXiv preprint arXiv:2210.10606*, 2022. 2

[26] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36:3536–3559, 2023. 2

[27] 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

[28] Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020. 2

[29] Hazarapet Tunanyan, Dejia Xu, Shant Navasardyan, Zhangyang Wang, and Humphrey Shi. Multi-concept t2i-zero: Tweaking only the text embeddings and nothing else. *arXiv preprint arXiv:2310.07419*, 2023. 2

[30] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. p+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 2

[31] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021. 2

[32] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Forty-first International Conference on Machine Learning*, 2024. 3

[33] Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*, 2024. 2

[34] Shai Yehezkel, Omer Dahary, Andrey Voynov, and Daniel Cohen-Or. Navigating with annealing guidance scale in diffusion space. *arXiv preprint arXiv:2506.24108*, 2025. 2, 6