Guiding What *Not* to Generate: Automated Negative Prompting for Text-Image Alignment

Supplementary Material

A. Additional Benchmark

T2I-CompBench T2I-CompBench [20] provides compositional prompts with multiple objects. We evaluate six tasks—attribute binding (color, shape, texture), object relationships (spatial, non-spatial), and a complex setting combining them—using 300 images per task (1,800 total). Metrics follow the benchmark: BLIP-VQA [20] for attribute binding, UniDet [78] for spatial relations, and CLIPScore [15, 51] for non-spatial relations; all three are applied to the complex task. In this setup, we compare our approach against recent alignment-enhancing methods in three groups: (1) *attention map–based* (Attend-and-Excite [4], SynGen [53]); (2) *layout-based* (LMD [32], InstanceDiffusion [61]); and (3) *feedback-based* (GORS [20], CoMat [23]). Only Attend-and-Excite is implemented on Stable Diffusion 2; the others (and ours) are implemented on SDXL.

In this experiment, we implement NPC on SDXL for a fair comparison. As shown in Table S3, NPC attains the best attribute-binding performance and nearly a 15%p gain on the complex task. While layout-based methods improve spatial controllability, their dependence on LLM outputs leads to failures with small or overlapping boxes [46]. In contrast, NPC avoids these issues, requires no additional training (unlike feedback-based methods), and needs no gradient updates (unlike attention-based methods), yielding both efficiency and superior average performance across tasks. Qualitative samples of alignment methods are also provided in Figure S2.

DPG-Bench DPG-Bench [18] comprises 1,065 long, dense prompts designed for compositional text–image evaluation, drawn from COCO, PartiPrompts, DSG-1k, and Objects365 and expanded via GPT-4, then human-verified (5 level-1 and 13 level-2 categories). DPG-Bench assesses text-to-image models on complex prompts, using mPLUG-large [29] to evaluate entities, attributes, and relationships with a focus on fine-grained details like object properties and spatial arrangements.

Table S4 reports DPG-Bench results: NPC attains state-of-the-art relation accuracy (93.34) and the highest overall score, while remaining strong on entities. It lags top systems on global/other and is slightly behind on attributes, yet still captures fine-grained compositional details efficiently. Figure S3 further shows robust alignment in complex scenes.

	pre-check	caption	propose	salient score
Time (s)	2.77	4.55	0.78	23.48

Table S1. Computation time by stage.

B. Experimental Details

Software and Hardware All experiments used PyTorch and an NVIDIA A40 GPU.

Inference time and Memory Results averaged over 10 randomly sampled cases (as in Table S1): pre-check about 2.77 s, caption about 4.55 s, proposal about 0.78 s, and scoring about 23.48 s. Scoring is the longest stage, but the overall increase remains marginal in practice. Memory changes were only marginal: we use the GPT API for captioning/proposal (no local GPU growth), and the salient-score stage ran on CPU, so GPU usage remained effectively unchanged.

C. Additional Results for Section 3

For further validation of our analysis in Section 3, we compute salient attention scores on a broader set of prompts, as summarized in Table S5. Salient tokens are manually selected from the generated images to pinpoint regions where text–image alignment fails; this manual curation is necessary to precisely measure how much attention increases on the misaligned parts, providing a more rigorous and reliable verification of our analysis. Across ten additional prompts, both the targeted and untargeted settings consistently yield higher salient attention scores than the base, reinforcing the robustness of our findings.

D. Model-Agnostic Effectiveness of NPC

To assess the generality of NPC beyond a single model or component choice, we evaluate its effectiveness across diverse generative architectures and MLLM or LLM configurations. As shown in Table S6, across all three base models—SDXL [49], SD3 [11], and FLUX [26]—applying NPC consistently improves performance on every evaluated attribute. The gains are substantial and uniform, demonstrating that NPC is not model-specific but transfers effectively across architectures with different capacities. These results confirm the general robustness and broad applicability of NPC.

Method	Attribute Binding			Object Relationship		Complex ↑	AVG ↑
	Color ↑	Shape ↑	Texture ↑	Spatial ↑	Non-Spatial ↑		
DPO-Diff [60]	0.4405	0.4381	0.4949	0.1445	0.3182	0.4119	0.3746
NPC$_{SDXL}$	**0.7950**	**0.5681**	**0.6889**	**0.2545**	**0.3191**	**0.5171**	**0.5237**

Table S2. Comparison with negative prompt optimization method (Section E).

Figure S1. Examples of DPO-Diff (Section E).

We further evaluate NPC using an alternative component stack in which the verifier and captioner are implemented with Qwen2.5-VL [2] and the proposer with Qwen3 [72]. As shown in Table S7, the Qwen-based configuration achieves performance comparable to the GPT-based setup across all attributes, demonstrating that NPC remains effective even when substituting the underlying LLM components. This again highlights the model-agnostic nature of NPC and its robustness across different verifier, captioner, and proposer architectures.

E. Comparison with Negative Prompt Optimization Method

Many prompt optimization approaches for enhancing image quality have been explored [14, 40]. Since NPC can also be considered a negative prompt optimization method, we compare it with DPO-Diff [60], a recently proposed approach that designs a compact space and employs shortcut gradient search to optimize negative prompts. As shown in Table S2, NPC outperforms DPO-Diff in T2I-CompBench since DPO-Diff is not specifically designed for alignment (examples of negative prompts from DPO-Diff are provided in Figure S1). Moreover, unlike DPO-Diff, which requires training, NPC enhances alignment purely through inference, making it a more efficient alternative.

F. LLM Prompts

In Table S8, Table S9, and Table S10, we provide the full LLM prompts used for the verifier, captioner, and proposer. Table S11 shows the LLM prompt used in the ablation study of the prompt-only proposer in Section 5.3.

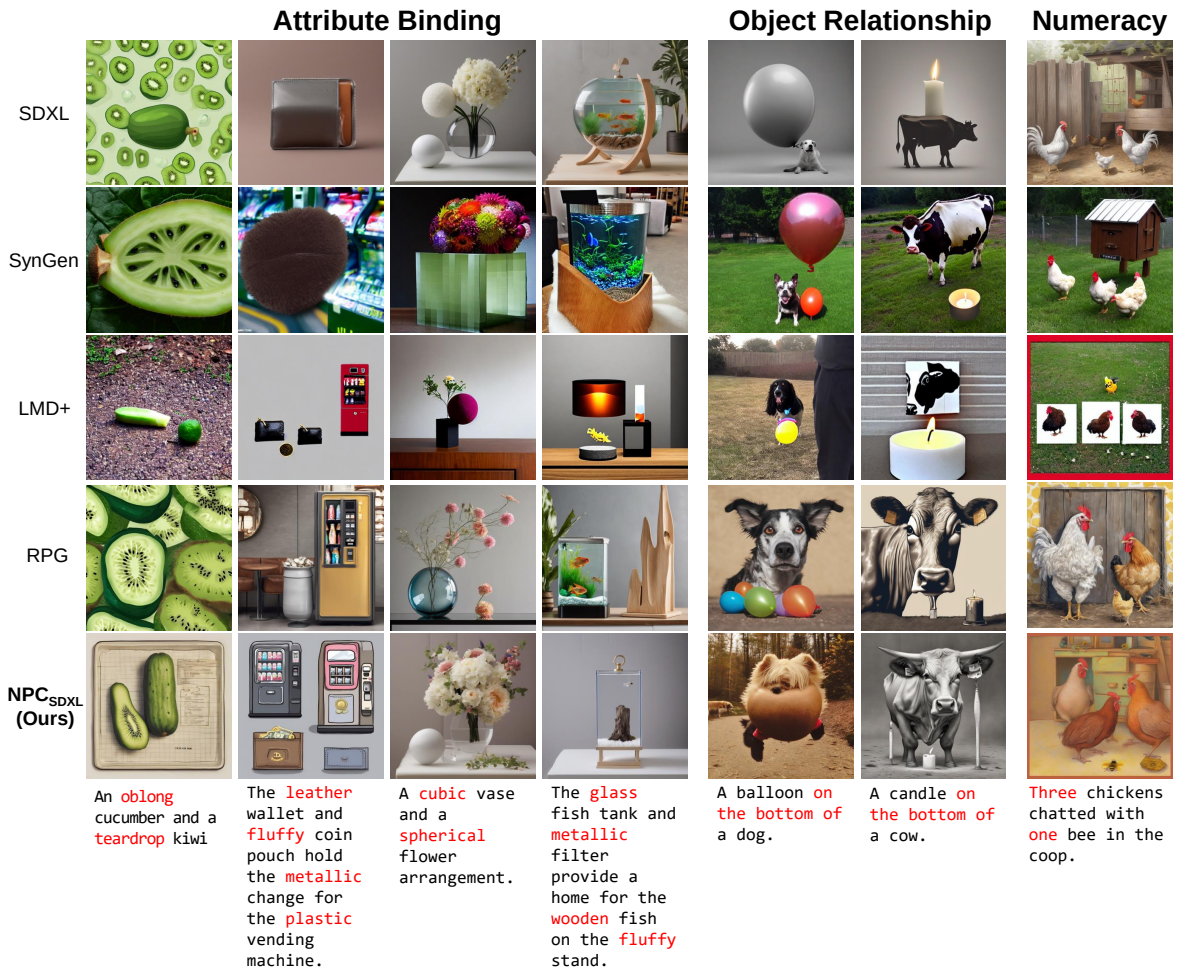

Figure S2. Qualitative comparison with SDXL [49] and other text-image alignment methods [32, 53, 73]. NPC consistently aligns across various tasks, while other methods fail in some tasks. Samples for NPC are generated using SDXL.

Model	Attribute Binding			Object Relationship		Complex ↑	AVG ↑
	Color ↑	Shape ↑	Texture ↑	Spatial ↑	Non-Spatial ↑		
Stable Diffusion 2	0.5065	0.4221	0.4922	0.1342	0.3096	0.3386	0.3672
SDXL	0.6369	<u>0.5408</u>	0.5637	0.2032	0.3110	0.4091	0.4441
●Attn-Exct v2	0.6400	0.4517	0.5963	0.1455	0.3109	0.3401	0.4140
●SynGen	0.7000	0.4550	0.6010	0.2260	0.3100	0.3339	0.4376
●LMD	0.4814	0.4865	0.5699	0.2537	0.2828	0.3323	0.4011
●InstanceDiffusion	0.5433	0.4472	0.5293	**0.2791**	0.2947	0.3602	0.4089
●GORS	0.6603	0.4785	0.6287	0.1815	**0.3193**	0.3328	0.4335
●CoMat	<u>0.7827</u>	0.5329	<u>0.6468</u>	0.2428	0.3187	<u>0.3680</u>	<u>0.4819</u>
NPC$_{SDXL}$ (Ours)	**0.7950**	**0.5681**	**0.6889**	<u>0.2545</u>	<u>0.3191</u>	**0.5171**	**0.5237**

Table S3. T2I-CompBench. Comparison of various methods across different attributes. The best performances are highlighted in bold and the second-best are underlined. The red circle (●) represents attention-map-based approaches, the blue circle (●) indicates layout-based approaches, and the green circle (●) denotes feedback-based approaches. Performance for the baselines is based on reported values.

Figure S3. Examples using DPGBench, comparing NPC with SD3.0 and DALL-E 3, demonstrate that NPC is also effective for long, dense-context prompts. Quantitative comparison results are provided in the Table S4.

Method	Global↑	Entity↑	Attribute↑	Relation↑	Other↑	Overall↑
SDXL [49]	83.27	82.43	80.91	86.76	80.41	74.65
Hunyuan-DiT [31]	84.59	80.59	88.01	74.36	86.41	78.87
DALLE3 [3]	<u>90.97</u>	89.61	88.39	90.58	<u>89.83</u>	83.50
SD3-medium [11]	87.90	<u>91.01</u>	88.83	80.70	88.68	84.08
FLUX.1-dev [26]	82.10	89.50	88.70	<u>91.10</u>	89.40	84.00
OmniGen [68]	87.90	88.97	88.47	87.95	83.56	81.16
Show-o [70]	79.33	75.44	78.02	84.45	60.80	67.27
EMU3 [62]	85.21	86.68	86.84	90.22	83.15	80.60
TokenFlow-XL [50]	78.72	79.22	81.29	85.22	71.20	73.38
Janus Pro [7]	86.90	88.90	89.40	89.32	89.48	84.19
T2I-R1 [24]	**91.79**	90.23	89.05	90.13	89.48	84.76
UniWorld-V1 [34]	83.64	88.39	88.44	89.27	87.22	81.38
OmniGen2 [65]	88.81	88.83	<u>90.18</u>	89.37	**90.27**	83.57
BAGEL [10]	88.94	90.37	**91.29**	90.82	88.67	<u>85.07</u>
NPC$_{\text{FLUX}}$	83.88	**91.12**	87.69	**93.34**	83.03	**85.38**

Table S4. DPG-Bench comparison (higher is better, ↑). Bold indicates the best value in each column; <u>underline</u> indicates the second-best.

Prompt	Salient	Targeted	Untargeted	Base↑	Targeted↑	Untargeted↑
A photo of two trains, three bottles, and two tennis rackets	three	2 bottles	toy stream locomotives	0.0831	**0.1018**	**0.0839**
					+0.0187	+0.0008
A photo of three bowls and two apples	three	stacked bowls	shiny apples	0.0746	**0.1018**	**0.0860**
					+0.0272	+0.0114
A photo of two teddy bears and three laptops	three	one laptop	wooden table	0.0583	**0.0642**	**0.0703**
					+0.0059	+0.0120
A photo of a purple chair, a blue bottle, and a white vase	white	purple vase	bulbous object	0.0427	**0.0512**	**0.0473**
					+0.0085	+0.0046
A photo of a larger donut on the above and a smaller zebra on the below	zebra	two donuts	colorful sprinkles	0.0344	**0.0371**	**0.0428**
					+0.0027	+0.0084
A photo of seven bears	seven	six bears	wildflowers	0.0926	**0.1031**	**0.1070**
					+0.0105	+0.0144
A photo of six vases	six	seven vases	glass vases	0.0863	**0.0947**	**0.0963**
					+0.0084	+0.0100
A photo of three dining tables and three horses	three	two horses	chandelier	0.1571	**0.1638**	**0.1630**
					+0.0067	+0.0059
A photo of two stop signs and one birds	two	one stop sign	television sets	0.0790	**0.0850**	**0.0820**
					+0.0060	+0.0030
A photo of two green tvs and two pink potted plant	two	one green TV	antennas	0.1178	**0.1464**	**0.1289**
					+0.0286	+0.0111

Table S5. Additional results for Section 3. results (higher is better, ↑). In each pair of rows, the first row lists *Base*, *Targeted*, and *Untargeted* scores (cells in **green, bold** exceed *Base*); the second row shows unlabeled absolute improvements in the last two columns *(Targeted−Base, Untargeted−Base)*. All prompts show *Targeted* and *Untargeted* > *Base*.

Model	Color	Count	Color/Count	Color/Pos	Pos/Count	Pos/Size	Multi-Count	Overall
SDXL [49]	0.050	0.375	0.000	0.000	0.000	0.000	0.000	0.061
NPC_{SDXL}	**0.500**	**0.575**	**0.200**	**0.400**	**0.325**	**0.425**	**0.500**	**0.417**
SD3 [11]	0.550	0.500	0.125	0.350	0.175	0.150	0.225	0.296
NPC_{SD3}	**0.650**	**0.675**	**0.325**	**0.550**	**0.575**	**0.600**	**0.625**	**0.571**
FLUX [26]	0.350	0.625	0.150	0.275	0.200	0.375	0.225	0.314
NPC_{FLUX}	**0.550**	**0.675**	**0.350**	**0.525**	**0.550**	**0.725**	**0.625**	**0.571**

Table S6. Comparison of metric scores across three models before and after applying NPC. In all cases, NPC improves performance across all evaluated attributes.

Method	Color	Count	Color/Count	Color/Pos	Pos/Count	Pos/Size	Multi-Count	Overall
Qwen	0.650	0.575	0.400	0.500	0.575	0.650	0.600	0.564
GPT	0.550	0.675	0.350	0.525	0.550	0.725	0.625	0.571

Table S7. Comparison of metric scores between Qwen-based and GPT-based components for NPC.

```
<System Prompt>:  You are an expert image evaluator.

Decide if a single candidate image fully satisfies a natural-language INSTRUCTION and, if
provided, an expectation CHECKLIST.

Strict rules:
1) All expectations must be satisfied:
   - Object classes, counts, colors, spatial relations (above/below/left/right by pixel
position), size/relative scale.
2) The image must be a single coherent, natural, photo-like scene.
   - Reject stylized (cartoons, sketches, anime), collages, multi-panels, or text-only.
3) Be strict and conservative.  If uncertain, mark incorrect.

Return a strict JSON object only:
{"correct":  1 or 0, "score":  float in [0,1], "reason":  "brief explanation"}
No extra text.  "score" should reflect how well the image meets the INSTRUCTION (and
CHECKLIST), even if correct=0.

<User Prompt>:  POSITIVE PROMPT:
$positive_prompt

CHECKLIST (optional):
$checklist

Judge the attached image against the INSTRUCTION (and CHECKLIST if present).
Return exactly:
{"correct":  1 or 0, "score":  float in [0,1], "reason":  "brief explanation"}
```

Table S8. LLM prompt for *Verifier*.

```
<System Prompt>:  You are a meticulous visual describer.
Write a comprehensive caption (90--160 words) describing concrete, falsifiable details only.
Cover:  main objects (counts, colors, materials), textures/patterns, spatial relations,
background elements,
lighting, style/mood, any visible text.
Avoid speculation, emotions, and camera metadata.  Output plain text only.

<User Prompt>:  Describe this image.
```

Table S9. LLM prompt for *Captioner*.

```
<System Prompt>:  You are a precise prompt editor.
From a given POSITIVE PROMPT and an image CAPTION, extract concise negative prompt
candidates (1--6 words, all lowercase) that would most effectively steer generation away
from elements that contradict the POSITIVE PROMPT.
Return a strict JSON object as specified.  Do not include explanations.

<User Prompt>:  POSITIVE PROMPT:
$positive_prompt

CAPTION:
$caption

GENERAL_FALLBACKS (use only if needed to reach K):
$fallbacks

K (max total candidates):  $k

Task
1) From the CAPTION, find elements that are NOT requested by the POSITIVE PROMPT (consider
synonyms, inflections, paraphrases).
2) Prioritize contradictions to the POSITIVE PROMPT (object identity, attributes like
color/material/count/shape, relations/pose), and if contradictions are weak or absent, pick
items at least tightly related to the POSITIVE PROMPT's objects (foreground-first).
3) Produce a flat list "candidates" of up to K concise negative phrases.
4) If you have fewer than K items, supplement with the most relevant items from
GENERAL_FALLBACKS while still obeying all constraints and avoiding anything implied by the
POSITIVE PROMPT.
5) Return exactly K items if possible (use fallbacks to fill).  Only return fewer than K if
no valid item remains.

Constraints
- Use ONLY elements explicitly present in the CAPTION.
- Exclude anything present in (or synonymous with) the POSITIVE PROMPT.
- Phrases:  lowercase, 1--6 words, no "no/without/not", no quotes, no trailing punctuation,
no duplicates.

Return exactly:
{"candidates":  ["...", "...", "..."]}
```

Table S10. LLM prompt for *Proposer*.

<System Prompt>: You are a precise prompt editor.
Given only a POSITIVE PROMPT (no caption) and K, produce up to K concise negative prompt candidates (1--6 words, all lowercase) that would most effectively steer image generation away from elements that contradict the POSITIVE PROMPT.
Return only strict JSON as specified. Do not include explanations.

Task
1) Parse the POSITIVE PROMPT to identify primary subjects/objects, key attributes (color/material/count/shape/pose), relations, scene/setting/time/lighting, viewpoint/composition, and medium/style/era.
2) Generate concise negative phrases that capture the most likely contradictions or near-misses to those elements, prioritized in this order: foreground object identity → attributes → relations/pose → composition/viewpoint → style/medium/era.
3) If strong contradictions are scarce, choose tightly related alternatives for the same foreground objects (e.g., sibling species, alternate colors/counts, opposite angles).

Constraints
-- Phrases: lowercase, 1--6 words, no ‘‘no/without/not’’, no quotes, no trailing punctuation, no duplicates.
-- Exclude anything present in (or synonymous with) the POSITIVE PROMPT.
-- Use only concepts reasonably inferred from the POSITIVE PROMPT’s domain; avoid unrelated items.
-- Foreground-first: prefer conflicts about the main subject before background or style.

<User Prompt>: POSITIVE PROMPT: $positive_prompt
K (max total candidates): $k

Task:
Propose $k negative prompt tokens likely to reduce undesirable or off-topic elements that could appear when generating this scene.
Return exactly:
{candidates: [..., ...]}

Table S11. LLM prompt for prompt ablation in Section 5.3.