Copy2Prompt: Copyright-Compliant Image Synthesis via Text-Mediated Generation

Lucas Lee
Department of Data Science
UCSD
San Diego, CA, 92092
l2lee@ucsd.edu

Anandhini Rajendran
Department of Computer Science
UCSD
San Diego, CA, 92092
anrajendran@ucsd.edu

David Lurie
Department of Data Science
UCSD
San Diego, CA, 92092
dlurie@ucsd.edu

Abstract

We introduce Copy2Prompt, a novel framework for copyright-compliant image-to-image synthesis that resolves the fundamental tension between semantic fidelity and ethical content generation by operating entirely in the textual domain. Our three-stage pipeline extracts detailed descriptions from reference images using vision-language models like LLaVA, refines these descriptions through large language models or RL agents, and synthesizes variants using diffusion models without direct pixel reuse. Evaluation across multiple backends (DALL-E 3, Stable Diffusion XL) demonstrates it is possible to utilize just textual description to generate faithful image while eliminating copyright concerns. Copy2Prompt represents a principled solution for ethical generative AI that balances creative flexibility with legal compliance through text-mediated synthesis.

1 Introduction

Diffusion models have transformed both text-to-image and image-to-image (I2I) synthesis, achieving photorealism and stylistic flexibility once reserved for GANs or classical renderers [1, 2]. In the I2I setting, a user supplies a reference picture and requests an edited or stylistically altered version. State-of-the-art pipelines typically fall into two families:

1. **Pixel/latent inversion.** Methods such as Stable Diffusion XL (SDXL) [3] invert the reference image into the diffusion latent space, then sample variants conditioned on that latent. While they preserve fidelity, they often reproduce protected pixels verbatim, raising copyright and ethical concerns [4].
2. **Conditional control.** Approaches like ControlNet inject edge, pose or depth hints into the denoising process [5]. They guarantee a geometric match but cannot retain fine appearance details without, again, copying large latent fragments of the original.

Thus, increasing noise during inversion avoids plagiarism but drifts semantically; decreasing noise retains detail but risks near-duplicates. The community therefore lacks a principled mechanism to balance *semantic fidelity* and *copyright safety*.

Text as a safe, controllable bridge. Recent vision–language models—e.g. BLIP-2 [6] and LLaVA [7]—can translate an image into a rich, grounded description. Concurrently, text-to-image back-ends such as SDXL [3] and single-step distillations like ADD/SD-Turbo [8] accept hundreds of tokens and expose fine-grained prompt weighting [9, 10]. These developments suggest a new paradigm: *convert the reference image to text, edit or weight that text, then re-synthesise from scratch*. Because generation is conditioned solely on words, no protected pixels are reused.

Submitted to Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Thus, we introduce Copy2Prompt, a text-mediated I2I pipeline simultaneously offers fine image semantic control and strong copyright compliance. The pipeline can be visualized in Figure 1.

2 Background

Our Copy2Prompt pipeline builds on recent advances in multimodal learning that enable high-fidelity visual-to-text translation through CLIP-based encoders and instruction-tuned language models. On the generation side, we leverage state-of-the-art diffusion and transformer-based models—particularly DALL-E 3 and SDXL—augmented with Compel's prompt weighting mechanisms to regain semantic and stylistic alignment with the original image. Finally, we enhance prompts using LLMs, and also experiment with reinforcement learning using CLIP-based scoring to evaluate prompt-generated images. Together, these components allow for controlled, copyright-compliant image regeneration guided by precise textual representations.

Visual-to-Textual Translation

The landscape of image-to-text translation has evolved significantly with the emergence of various multimodal architectures. Contrastive Language-Image Pretraining (CLIP) learns image representations that capture semantics and style, allowing it to serve as the basis for prompting image regeneration that is similar to the original image [11]. Recent work, however, has demonstrated that CLIP models inherently possess the ability to convert images into text prompts through linear projection matrices calculated in closed form [12]. Liu et al.'s visual instruction tuning established that language models can process visual information through frozen CLIP encoders, achieving 85.1% relative performance on reasoning tasks [13]. This foundation enables pixel-accurate visual grounding in OMG-LLaVA (Zhang et al.) [14], which integrates universal segmentation with large language models (LLMs).

Text-to-Image Generation

Contemporary work in text-to-image generation has evolved from more rigid models to complex multi-component pipelines that combine specialized generators, adapters, and upscaling modules [14] [15]. Diffusion models remain dominant in this space, with some providing highly controllable frameworks [16] [17].

Methods to optimize these frameworks, like policy optimization, reduce computation and improve quality by training ensemble reward models that predict image quality directly from prompt-workflow combinations [18]. Prompt engineering has also been explored, revealing that different timesteps in diffusion processes exhibit varying sensitivity to conditioning, with noisy later steps showing greater responsiveness to prompt modifications [19].

Compel Integration for Prompt Weighting Compel [20] is a library that enables direct manipulation of text embedding tensors in diffusion models. Compel parses weighted prompt syntax and modifies the corresponding CLIP text encoder embeddings before they condition the diffusion process. The library applies scalar multipliers to specific token embeddings: + operators scale embeddings by 1.1 $\times$ while - operators apply 0.9 $\times$ scaling, with compound operations enabling fine-grained attention control (e.g., ++ results in $1.1^2 = 1.21$). The .and() function concatenates separately weighted embedding segments, allowing heterogeneous conditioning where different prompt components receive distinct emphasis levels during the denoising process.

DALL-E 3 DALL-E 3's native ChatGPT integration automatically refines prompts, expanding descriptions into detailed generation instructions that complement LLaVA's annotations. The model's enhanced nuance understanding enables more accurate translation of textual descriptions into visual outputs. Additionally, DALL-E 3's sophisticated content filtering and copyright compliance measures align with Copy2Prompt's objectives, providing built-in safeguards against copyrighted material generation while maintaining creative flexibility.

3 Methodology

We propose Copy2Prompt, a framework designed to generate high-fidelity image variations by converting source images into detailed textual prompts, which are then used to guide advanced text-to-image diffusion models. The pipeline integrates vision-language models and prompt engineering

to achieve copyright-compliant, semantically rich image synthesis, with the latter vision-language diffusion model never having no pixel-level knowledge of the source image, thereby avoiding copyright claims.

3.1 Pipeline Overview

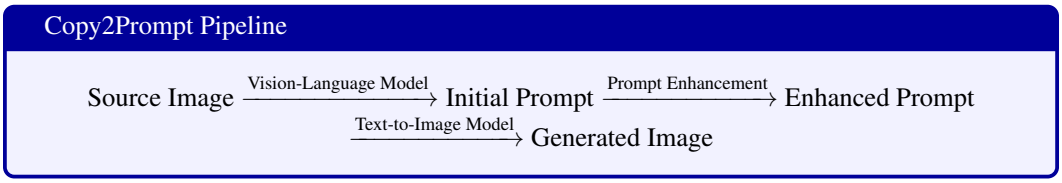


Figure 1: Overview of the Copy2Prompt pipeline for generating image variations through text-mediated synthesis.

As seen in Figure 1, Copy2Prompt works in three steps.

1. Image understanding (vision language model):
   - Input: A source image
   - Output: A detailed textual prompt describing the visual content, objects, and relationships present in the image.
2. Prompt Enhancement:
   - Input: Output from previous step
   - Output: Enhanced textual prompt with improved coherence, tailoring the description for downstream image synthesis
3. Image Synthesis (text-to-image diffusion model):
   - Input: Output from previous step
   - Output: Generated image, replicating the source image through pure textual description

3.2 Image Understanding

The image understanding component transforms visual information into rich textual representations that capture both semantic content and spatial relationships. We employ vision-language models, specifically LLaVA, to generate comprehensive descriptions that go beyond simple object detection, incorporating contextual details, visual attributes, and compositional elements essential for high-fidelity image reconstruction through text-to-image synthesis.

LLaVA yields multi-sentence annotations that exceed traditional captioning approaches in both depth and granularity. The architecture converts images into text through a two-stage process: first, CLIP's vision encoder extracts visual features from the input image, then these features are passed to a modified LLaMA language model, which generates comprehensive descriptive captions. The complete LLaVA processing pipeline is illustrated in Figure 2.

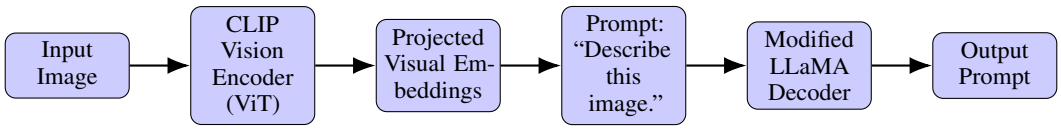


Figure 2: LLaVA architecture for converting input images to detailed textual descriptions through CLIP visual encoding and LLaMA language generation.

3.3 Prompt Enhancement

The initial prompt generated by LLaVA often requires strategic refinement to optimize compatibility with text-to-image generation models. We employ large language models (ChatGPT) to enhance

these descriptions. The system prompt used by ChatGPT to format the detailed prompt is shown in Appendix A.

The prompt enhancement process addresses several key limitations of raw vision-language outputs: improving linguistic coherence, expanding contextual details, and restructuring descriptions to align with the inductive biases of diffusion models. This refinement ensures that critical visual elements are preserved while adapting the textual representation for optimal downstream synthesis.

3.4 Image Generation

The final stage of our pipeline leverages state-of-the-art text-to-image diffusion models to synthesize high-fidelity images from the enhanced textual descriptions. We primarily employ Stable Diffusion XL (SDXL) as our generation backbone, though our framework supports multiple diffusion architectures to accommodate different quality-speed trade-offs.

SDXL and diffusion models in general can only take in a textual inputs of 77 tokens. This is because CLIP, the underlying text encoder, is limited to processing 77 tokens per prompt.

To address this limitation, we implement two complementary strategies to maximize prompt utilization and preserve the richness of information generated during the enhancement stage.

First, we integrate Compel with SDXL to enable advanced prompt processing capabilities, including weighted attention mechanisms, prompt segmentation for fine-grained control, and intelligent chunking with padding to retain maximum textual information within model constraints.

Second, we support DALL-E 3 as an alternative generator, which allows for prompts up to 4000 tokens — significantly expanding the context window compared to base SDXL for the ability to read in detailed prompts.

3.5 Reinforcement as an Alternative for Prompt Enhancement

Instead of relying on fixed or manually engineered prompts and having an LLM revise prompts, we formulate the problem as a prompt optimization for image generation as an RL problem, where the agent iteratively modifies text prompts to maximize visual similarity between the target image and generated image. The agent interacts with the text-to-image generation and is guided by a CLIP-based reward. This approach makes no assumptions on the image generator, making it scalable for larger pipelines.

RL Formulation: The agent interacts with the Stable Diffusion XL generator by sampling prompt mutations (actions) to generate images from text (states), aiming to maximize a CLIP-based similarity score (reward) with a reference image. The policy randomly selects stylistic, technical, and compositional modifiers to evolve prompts toward higher visual-semantic alignment.

Optimization Loop The agent iteratively changes the best prompt, generates an image via the SDXL pipeline, evaluates it with CLIP similarity to the reference, and updates the best prompt and image if the reward improves.

4 Evaluation

In this section, we outline both the similarity metrics used to assess image quality and the experimental results that demonstrate the effectiveness of our pipeline. This comprehensive evaluation provides insight into how well our approach preserves visual fidelity and semantic alignment, as well as parameters and different components affect the generated images.

4.1 Similarity Metrics:

To assess the quality of the generated images, we evaluate their similarity to the original images using a combination of perceptual and semantic similarity metrics:

- **Structural Similarity Index (SSIM)** measures the perceptual similarity in structure and texture between two images, reflecting how well the layout and content are preserved.

- **CLIP Score** evaluates semantic similarity between images and text, capturing high-level content alignment and conceptual consistency.
- **Learned Perceptual Image Patch Similarity (LPIPS)** quantifies perceptual distance using deep features, providing a measure of human-perceived visual similarity.

4.2 Experiments and Results

4.2.1 RL Agent for Prompt Enhancement

We use different type of prompt modifiers in the RL agent to evaluate the impact on the generated image. The prompt modifiers were categorized from three groups:

- **Quality Enhancers:** These include descriptors that improve image fidelity and realism, such as "hyperrealistic" and "8k resolution."
- **Technical Specs:** Modifiers reflecting photographic parameters, for example, "shot on Canon EOS R5" and "f/1.4 aperture."
- **Composition Rules:** Terms guiding the visual arrangement, like "rule of thirds" and "symmetrical composition."

The RL agent employs an iterative optimization strategy, sampling from predefined modifier pools to progressively enhance prompt effectiveness. We demonstrate this process using a representative example with the base prompt: *"Sunset at the beach with a group of people."* as shown in Table 1 and Figure 3

Table 1: RL Agent Optimization Trajectory

Iteration	Enhanced Prompt	CLIP Score	Key Modifiers
Base	"Sunset at the beach with a group of people"	–	–
1	Base + "intricate details, shot on Canon EOS R5, balanced composition"	0.7515	Technical + Composition
2	Base + "ultra realistic, long exposure, golden ratio"	0.7129	Quality + Technical + Composition
3	Base + "fine art quality, film grain texture, leading lines"	0.7593	Quality + Technical + Composition
4	Base + "award winning, bokeh background, diagonal composition"	**0.8228**	Quality + Technical + Composition
5	Base + "professional photography, f/1.4 aperture, negative space"	0.7905	Quality + Technical + Composition

Convergence Pattern: The optimization process demonstrates non-monotonic convergence, with peak performance achieved at iteration 4 (CLIP score: 0.8228), followed by a slight decline in iteration 5.

Optimal Modifier Combination: The highest-performing configuration combined quality descriptors ("award winning"), technical specifications ("bokeh background"), and compositional guidelines ("diagonal composition").

Implications: This case study demonstrates that RL agents can systematically transform simple base prompts into sophisticated generation instructions through strategic modifier addition. The 9.5% improvement from baseline to optimal configuration ($0.7515 \rightarrow 0.8228$) validates the effectiveness of compositional and stylistic enhancement for image-text alignment tasks. The non-monotonic optimization trajectory highlights the importance of allowing sufficient exploration in the RL training process, as premature convergence could miss optimal modifier combinations.

Qualitatively, the source and generated image pair in Figure 1 are semantically similar to humans.

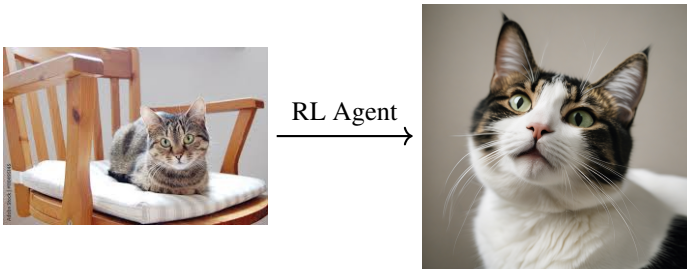
Figures 4 and Figure 5 show additional experiments for the RL agent. Qualitatively, the lake images are compositionally similar, where as the cat images are not. This is in contrast to what CLIP score

CLIP Score: 0.8228
Final Prompt:
Sunset at the beach with a group of people, award winning, bokeh background, diagonal composition

Figure 3: Transformation from input image (left) to output image (center) generated by the RL agent, with the final CLIP similarity score and prompt shown on the right.

CLIP Score: 0.7622
Final Prompt:
Lake, highly detailed, 35mm film aesthetic, symmetrical composition.

Figure 4: Transformation from input image (left) to output image (center) generated by the RL agent, with the final CLIP similarity score and prompt shown on the right.



CLIP Score: 0.8681
Final Prompt:
A cat, studio quality, medium format camera, dynamic composition

Figure 5: Transformation from input image (left) to output image (center) generated by the RL agent, with the final CLIP similarity score and prompt shown on the right.

is suggesting. This suggests that CLIP score may value color accuracy more than compositional accuracy, and this prioritization does not always align with human judgement.

4.2.2 GPT Model and Prompt Length into SDXL + Compel

We evaluate the effectiveness of Copy2Prompt on a subset of the Unsplash dataset (N = 15) containing photographs akin to background wallpapers. Our experimental setup in Table 2 systematically varies the ChatGPT model version (GPT-4.0 vs GPT-4.1) and target prompt length (77, 200, 500, and 1000 words) to assess their impact on generation quality.

For each image, we generate initial descriptions using LLaVA, enhance these prompts using ChatGPT with specific length constraints, then feed the enhanced prompts to SDXL for image generation. We evaluate results using SSIM (structural similarity), LPIPS (perceptual distance), CLIP score (semantic alignment), and actual output prompt length.

Initially, we used GPT-4.0 with explicit word count constraints but observed significant length mismatches in actual output, particularly for longer targets (500-1000 words), prompting a switch to GPT-4.1 which demonstrated better constraint adherence. The results reveal that longer enhanced prompts do not consistently improve generation quality: GPT-4.0 shows decreasing SSIM scores

Table 2: Comparison of model performance on varying Prompt Enhancement Model and Prompt Length

Model/Technique	Mean SSIM ↑	Mean LPIPS ↓	Mean CLIP Score ↑	Mean output length from ChatGPT
Copy2Prompt 4o (77)	0.19	0.67	0.80	72
Copy2Prompt 4o (200)	0.19	0.68	0.79	180
Copy2Prompt 4o (500)	0.17	0.66	0.79	383
Copy2Prompt 4o (1000)	0.16	0.68	0.79	**211**
Copy2Prompt 4.1 (77)	0.18	0.68	0.80	73
Copy2Prompt 4.1 (200)	0.20	0.67	0.81	182
Copy2Prompt 4.1 (500)	0.18	0.67	0.81	450
Copy2Prompt 4.1 (1000)	0.18	0.70	0.80	890

from 0.19 (77 words) to 0.16 (1000 words) with stable CLIP scores (0.79-0.80), while GPT-4.1 maintains consistent performance across lengths (SSIM: 0.18-0.20, CLIP: 0.80-0.81). Notably, both models achieve similar performance with concise 77-word prompts, suggesting well-crafted shorter descriptions can be as effective as verbose ones.

4.2.3 LLaVa Prompt Length + ChatGPT Prompt Length:

Table 3: Comparison of model performance on varying LlaVA Output Length

Model/Technique	Mean SSIM ↑	Mean LPIPS ↓	Mean CLIP Score ↑	Mean output length from ChatGPT
Copy2Prompt 4.1 (77)	0.18	0.68	0.79	75
Copy2Prompt 4.1 (200)	0.16	0.69	0.81	186
Copy2Prompt 4.1 (500)	0.17	0.67	0.81	460
Copy2Prompt 4.1 (1000)	0.17	0.70	0.83	820

To investigate whether richer initial descriptions improve generation quality, we configured LLaVA with extended token limits (1800-2400 tokens) compared to (600-1000) from Table 2 and repeated the experiment using the same dataset (N = 15) and GPT-4.1 enhancement with identical target lengths.

The results in Table 3 compared to those in Table 2 show modest improvements in semantic alignment, where CLIP scores demonstrate a clearer upward trend from 0.79 (77 words) to 0.83 (1000 words). On the other hand, SSIM and LPIPS values remain largely unchanged.

While longer LLaVA outputs slightly improve semantic metrics, the effect size is minimal and unlikely to significantly impact practical image generation quality.

4.2.4 Image Generators + ChatGPT Prompt Length

Table 4: Comparison of model performance across different image generators and prompt length

Model/Technique	Mean SSIM ↑	Mean LPIPS ↓	Mean CLIP Score ↑
SDXL Source Image to Image Baseline	**0.62**	**0.16**	**0.92**
Source image → ChatGPT → Dalle-3 Baseline	0.20	0.67	**0.82**
BLIP-2 → SDXL Baseline	0.11	0.75	0.66
Copy2Prompt (Dalle-3, 77)	0.20	0.72	**0.81**
Copy2Prompt (Dalle-3, 200)	0.19	0.70	0.78
Copy2Prompt (Dalle-3, 500)	**0.22**	0.69	0.79
Copy2Prompt (Dalle-3, 1000)	0.21	0.70	0.77
Copy2Prompt (SDXL+C, 77)	0.20	0.68	0.79
Copy2Prompt (SDXL+C, 200)	0.17	0.69	0.77
Copy2Prompt (SDXL+C, 500)	0.16	0.67	0.78
Copy2Prompt (SDXL+C, 1000)	0.16	**0.66**	0.79

To evaluate different image generation backends, we compared Copy2Prompt using DALL-E 3 and SDXL (with Compel) against three baselines: SDXL image-to-image (direct pixel-level generation), ChatGPT→DALL-E 3 (image-to-text-to-image), and BLIP-2→SDXL (captioning baseline).

As shown in Table 4, SDXL image-to-image. as expected, achieves superior performance (CLIP: 0.92, SSIM: 0.62), reflecting the advantages of direct pixel manipulation. Among text-mediated approaches, Copy2Prompt variants consistently outperform the BLIP-2 captioning baseline and perform comparably to the ChatGPT baseline. The similarity between ChatGPT→DALL-E 3 and Copy2Prompt performance suggests that LLaVA (llava-hf/llava-v1.6-mistral-7b-hf), despite being designed for multi-granular image description, may lack sufficient detail for optimal prompt generation. Notably, the ChatGPT baseline can be viewed as an alternative Copy2Prompt implementation, as both perform image-to-text-to-image translation.

Our results reveals distinct backend-specific responses to prompt length variations. SDXL+Compel shows degrading structural similarity as prompt length increases (SSIM: 0.22→0.16) while maintaining stable perceptual metrics (LPIPS, CLIP). Conversely, DALL-E 3 preserves structural and perceptual consistency across lengths but exhibits declining semantic alignment (CLIP: 0.81→0.77). These contrasting behaviors demonstrate that optimal prompt strategies are model-dependent, with prompt quality and relevance proving more critical than length alone.

A full run of Copy2Prompt can be seen in Appendix B.

5 Conclusion

5.1 Discussion

Copy2Prompt demonstrates an image-text-image pipeline that enhances image generation through prompt optimization and model selection. The RL agent successfully alters prompts by adding details and wording that results in generated images that have greater semantic similarity to the original image. Furthermore, our experiments show that increasing the length of the image-understanding prompt improves image generation quality. However, increasing the prompt enhancement prompt length does not always result in improved quality, and the effect of the prompt length depends on the image generation model used. When comparing candidate image generators, DALLE-3 performs marginally better than SDXL with Compel in our Copy2Prompt framework. Our results also highlight the limitations of captioning models like LLaVA for comprehensive image descriptions, where their textual descriptions appear no better than general purpose multimodal models like ChatGPT 4o.

Overall, we believe that it's possible to generate semantically similar photos to the human eyes without using any pixel level information based on the experiments. However, significant effort has to go into the prompting of the image understanding model, the prompt enhancer model, or the compute spent on training RL agents.

5.2 Future Work

Future research directions include:

1. **Enhanced RL Integration**: Extending RL optimization to the vision-language understanding stage and integrating with LLMs for dynamic prompt generation rather than sampling from curated modifier lists
2. **Advanced Description Strategies**: Since longer descriptions did not improve fidelity, exploring ensemble methods where LLaVA generates multiple complementary descriptions, hierarchical frameworks separating structural and semantic elements, and more sophisticated vision-language models, like OMG-LlaVA
3. **Copyright Validation**: Systematic evaluation on real copyright infringement datasets to establish empirical guidelines for transformative use and validate legal compliance.

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.

[3] A. Podell and S. AI, "Stable diffusion xl," https://stability.ai/blog/stable-diffusion-xl, 2023.

[4] N. Yu, C. Meng, M. Fritz *et al.*, "Synthetic imagery can be traced back to its generative source," *Nature*, vol. 620, pp. 104–111, 2023.

[5] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *arXiv preprint arXiv:2302.05543*, 2023.

[6] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *ICLR*, 2023.

[7] H. Liu, Q. Ji *et al.*, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023.

[8] A. Sauer, N. Frühstück, M. Sun, and A. Geiger, "Adversarial diffusion distillation," *ICLR*, 2023.

[9] A. Hertz, Y. Bahat, R. Giryes, A. A. Efros, and E. Shechtman, "Prompt-to-prompt image editing with cross attention control," in *SIGGRAPH Asia*, 2022.

[10] S. Mahajan, T. Rahman, K. M. Yi, and L. Sigal, "Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models," 2023. [Online]. Available: https://arxiv.org/abs/2312.12416

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[12] Y. Ding, C. Tian, H. Ding, and L. Liu, "The clip model is secretly an image-to-prompt converter," 2024. [Online]. Available: https://arxiv.org/abs/2305.12716

[13] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023. [Online]. Available: https://arxiv.org/abs/2304.08485

[14] C. Zhang, C. Zhang, M. Zhang, I. S. Kweon, and J. Kim, "Text-to-image diffusion models in generative ai: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2303.07909

[15] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," 2022. [Online]. Available: https://arxiv.org/abs/2205.11487

[16] S. Hartwig, D. Engel, L. Sick, H. Kniesel, T. Payer, P. Poonam, M. Glöckler, A. Bäuerle, and T. Ropinski, "A survey on quality metrics for text-to-image generation," 2025. [Online]. Available: https://arxiv.org/abs/2403.11821

[17] X. Zhang, H. Zhou, H. Qin, X. Lu, J. Yan, G. Wang, Z. Chen, and Y. Liu, "Enabling versatile controls for video diffusion models," 2025. [Online]. Available: https://arxiv.org/abs/2503.16983

[18] U. Gadot, R. Gal, Y. Ziser, G. Chechik, and S. Mannor, "Policy optimized text-to-image pipeline design," 2025. [Online]. Available: https://arxiv.org/abs/2505.21478

[19] S. Mahajan, T. Rahman, K. M. Yi, and L. Sigal, "Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models," 2023. [Online]. Available: https://arxiv.org/abs/2312.12416

[20] damian0815, "Compel," 2025. [Online]. Available: https://github.com/damian0815/compel

Appendix

A System Prompts

This section details the system prompts used for different components of the Copy2Prompt pipeline.

A.1 LLaVA Enhancement Prompt

The following system prompt was used to enhance initial LLaVA-generated descriptions:

LLaVA Enhancement System Prompt

```
You are an expert at converting detailed image descriptions into concise,
effective prompts for Stable Diffusion XL.

Your task is to transform a long, detailed description into a prompt of
20% more or less than EXACTLY {target_length} words.

Key requirements:
1. Preserve the most visually important elements
2. Use comma-separated phrases for clarity
3. Front-load the most important elements (subject, style, composition)
4. Remove redundant or verbose descriptions
5. Keep specific visual details that define the image
6. MUST be {target_length} words or less
```

A.2 ChatGPT Base Prompt

The base system prompt for ChatGPT as the prompt enhancer:

ChatGPT Base System Prompt

```
You are an expert at converting detailed image descriptions into concise,
effective prompts for Stable Diffusion XL.

Your task is to transform a long, detailed description into a prompt of
20% more or less than EXACTLY {target_length} words.

Key requirements:
1. Preserve the most visually important elements
2. Use comma-separated phrases for clarity
3. Front-load the most important elements (subject, style, composition)
4. Remove redundant or verbose descriptions
5. Keep specific visual details that define the image
6. MUST be {target_length} words or less
```

A.3 ChatGPT Style Instructions

Additional style instructions provided to ChatGPT for structured prompt generation:

ChatGPT Style Instructions

```
Include: main subject with key features, composition, artistic style,
atmosphere, important details

Structure: [detailed subject], [composition], [style/medium],
[atmosphere/mood], [key details]
```

A.4 ChatGPT Fallback Prompt

When the ChatGPT formatter produces output that deviates from the target length by more than 20%, the following fallback prompt is used:

ChatGPT Length Correction Fallback

```
Convert the description to EXACTLY {target_length} words.
Count every word. Be extremely concise.
```

B Case Study: Complete Copy2Prompt Pipeline

This section demonstrates a complete Copy2Prompt execution using GPT-4.1 for prompt formatting with a target length of 77 words. The LLaVA model was configured to generate descriptions between 800-1200 tokens to ensure comprehensive image analysis, and SDXL with Compel is used to generate the final image.

B.1 Pipeline Execution

Stage 1: Source Image Analysis

Figure 6 shows the input image from the Unsplash dataset, depicting a winter forest scene with a cross-country skier.

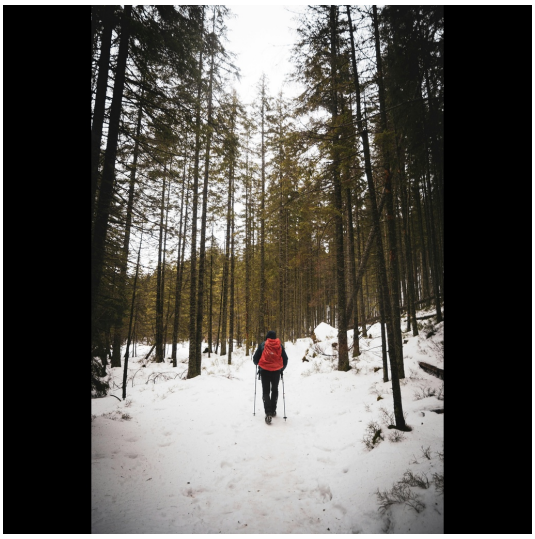


Figure 6: Source image: Winter forest scene with cross-country skier

Stage 2: LLaVA Description Generation

The vision-language model produced a detailed description scene composition, lighting conditions, and photographic elements. Key extracted features include:

- *Subject*: Lone figure in red jacket with ski poles on snow-covered path
- *Environment*: Dense forest with snow-dusted coniferous and deciduous trees
- *Composition*: Off-center positioning following rule of thirds, blurred background for depth
- *Lighting & Color*: Soft diffused lighting, muted monochromatic palette with red accent
- *Style*: Professional landscape photography with attention to texture and detail

The full prompt below:

> *"The image captures a serene winter scene in a forest. The main subject is a lone figure, clad in a vibrant red jacket and black pants, walking away from the viewer on a snow-covered path. The individual is equipped with ski poles, suggesting they are on a cross-country skiing adventure. The person's posture is relaxed, with their head turned slightly to the side, as if they are taking in the beauty of their*

surroundings. The forest is a dense thicket of tall, slender trees with a mix of coniferous and deciduous species. The trees are dusted with snow, their branches heavy with the white substance. The forest floor is blanketed in a thick layer of snow, untouched except for the path the skier is following. The lighting in the image is soft and diffused, indicative of an overcast sky. The colors are muted, with the red of the skier's jacket standing out against the otherwise monochromatic palette of the snowy forest. The image is rich in detail, with the texture of the snow and the bark of the trees being particularly noticeable. The composition of the image is balanced, with the skier positioned slightly off-center, following the rule of thirds. The path the skier is on leads the viewer's eye deeper into the forest, creating a sense of depth and perspective. The background is blurred, drawing focus to the skier and the immediate surroundings, but still providing a sense of the vastness of the forest. The image is a testament to the beauty of nature, captured in a moment of quiet solitude. The skier's journey through the snowy forest is a journey of exploration and connection with the natural world. The image is a masterclass in landscape photography, with its attention to detail, use of color, and composition. It is a snapshot of a moment, frozen in time, that speaks to the viewer's sense of adventure and the allure of the great outdoors. This image is a stunning example of professional photography, with its attention to detail, use of color, and composition. It is a snapshot of a moment, frozen in time, that speaks to the viewer's sense of adventure and the allure of the great outdoors. The image is a testament to the beauty of nature, captured in a moment of quiet solitude. The skier's journey through the snowy forest is a journey of exploration and connection with the natural world. The image is a masterclass in landscape photography, with its attention to detail, use of color, and composition. It is a snapshot of a moment, frozen in time, that speaks to the viewer's sense of adventure and the allure of the great outdoors. This image is a stunning example of professional photography, with its attention to detail, use of color, and composition. It is a snapshot of a moment, frozen in time, that speaks to the viewer's sense of adventure and the allure of the great outdoors. The image is a testament to the beauty of nature, captured in a moment of quiet solitude. The skier's journey through the snowy forest is a journey of exploration and connection with the natural world. The image is a masterclass in landscape photography, with its attention to detail, use of color, and composition. It is a snapshot of a moment, frozen in time, that speaks to the viewer's sense of adventure and the allure of the great outdoors."

Note: The LLaVA output exhibited significant repetition in the latter portion, repeating descriptive phrases multiple times—a common characteristic requiring downstream filtering, further supporting our claim of how increasing LlaVA output length does not lead to meaningful increase in image generation performance.

Stage 3: GPT-4.1 Prompt Optimization The lengthy LLaVA description was condensed into a focused 77-word prompt optimized for diffusion model generation:

> *"Lone figure in vibrant red jacket and black pants, walking away with ski poles on a snow-covered forest path, surrounded by tall coniferous and deciduous trees dusted with snow, balanced off-center composition, soft diffused lighting, muted monochromatic palette with striking red accent, professional landscape photography, serene winter atmosphere, detailed snow texture and tree bark, blurred background for depth, sense of solitude and adventure, rule of thirds, natural beauty, exploration, connection with nature, high detail."*

Stage 4: Image Generation The optimized prompt was fed to SDXL, producing the output shown in Figure 7. The generated image successfully captures the core semantic elements: winter forest setting, lone figure in red jacket, snow-covered environment, and professional photography aesthetic.

B.2 Analysis

This case study demonstrates Copy2Prompt's ability to preserve semantic fidelity while avoiding direct pixel reuse. The generated image maintains the original's compositional structure and pho-

Figure 7: Generated image: SDXL output from Copy2Prompt pipeline

tographic style almost perfectly. However, the color pallete is slightly off as the subject is slightly wearing a different color clothing the trees have different colors. We determine that this is a succesful generation, as this is compositionally faithful to the original image, while producing a distinctly different visual output suitable for copyright-compliant applications. Furthermore, it can be fed back into similar image-to-image models with custom prompts to edit the image to be even closer to the source image. For example, a sample prompt could be "most tree leaves are yellow-green, and the person is wearing a black jacket with a red camping backpack."