

Self-Improvement of Image Synthesis via Natural Language Feedback

Anonymous ACL submission

Abstract

In recent years, substantial strides have been made in the field of text-to-image generation, where models seek to convert textual descriptions into visual representations. Despite notable advancements, challenges persist in generating high-quality images that accurately reflect user-provided prompts. The primary hurdle lies in the simplicity of user prompts, often lacking the necessary details for comprehensive model understanding and accurate visualization. To overcome this, we propose an innovative solution utilizing multimodal large language models (MLLMs). Our enhanced pipeline integrates MLLMs to augment user input text, coupled with a robust evaluation process assessing image quality and alignment with original prompts. This approach not only improves prompt comprehensibility for enhanced image generation but also ensures alignment with user preferences. Preliminary experiments on the MS-COCO dataset showcase the potential of our pipeline, producing images meeting user requirements with greater details than traditional models and prompt refinement pipelines. This highlights the practical applicability of our approach in advancing text-to-image models.

1 Introduction

Over the past few years, there has been a rapid and notable advancement in the field of deep generative models. Specifically, models focused on converting text into images have emerged as a prominent area of research, drawing significant interest from both academic circles and the general public. These generative models have increasingly made an impact across various domains, encompassing specialized tasks like image manipulation and object recognition in computer vision, as well as broader applications such as interior design and the depiction of abstract concepts (Ledig et al., 2017; Liu and Tuzel, 2016; Brock et al., 2019; Liu et al., 2019; Prakash and Karam, 2021). However, despite the

considerable progress made in this field, text-to-image models often face substantial challenges in generating high-fidelity images that accurately correspond to user-provided prompts (Lee et al., 2023; Feng et al., 2022; Liu et al., 2022a,b). This intricate obstacle represents a significant hurdle to the continuous advancement of text-to-image generation.

To mitigate these challenges, numerous research endeavors have been initiated to address issues related to attribute binding, often involving modifications to the attention mechanism in diffusion models (Agarwal et al., 2023; Chefer et al., 2023). While these approaches exhibit promise, they often fall short in fully tackling the complexities inherent in text-to-image generation. They frequently struggle to overcome hurdles associated with producing high-quality images that faithfully capture the input prompts.

Moreover, the task of evaluating the correspondence between the generated image and the input text remains inadequately explored. This aspect holds paramount importance in text-to-image generation to ensure the fidelity of the generated images to the intended prompts. Various methods, including those leveraging models like CLIP (Radford et al., 2021) and BLIP (Li et al., 2023), have been proposed to address this challenge. However, these methods often lack precision in capturing compositional alignment, highlighting the necessity for further research and development in this field to enhance the accuracy and quality of text-to-image generation.

In this study, we introduce a complex computational pipeline that effectively combines MLLMs and text-to-image models to tackle the inherent challenges in text-to-image generation. Our main goal is to improve the clarity and accuracy of user input prompts by immediately applying MLLMs, ensuring a detailed understanding by the text-to-image model. We further enhance our pipeline's capabilities by incorporating MLLMs into the post-

generation phase. After each iteration, MLLMs are used for prompt adherence validation, triggering a continuous refinement process if the generated image deviates from the set criteria. This dynamic iterative validation mechanism, coupled with our innovative methodologies inspired by models like CLIP and BLIP, not only underscores the effectiveness of the collaborative text-to-image and MLLM models but also establishes our study as a trailblazing effort in surmounting the ongoing challenges in the field. By carefully addressing issues related to attribute binding and compositional alignment, our approach aims to significantly enhance the accuracy and quality of text-to-image generation. With these advancements, our contributions include efforts in tackling persistent challenges within the field of text-to-image generation, with a focus on refining user prompts to significantly improve the accuracy and quality of the generated images, thereby enhancing the user experience.

2 Methodology

In our pursuit of the goal of producing high-quality and detailed images from simple input prompts, we have devised a pipeline that integrates the capabilities of both the SD model (Rombach et al., 2022), LLaVA model (Chen et al., 2023), CogVLM model (Wang et al., 2023; Hong et al., 2023), and GPT-4V model (OpenAI, 2023). The synergy between these models is strategically harnessed to optimize the final output. Our pipeline is meticulously designed to not only enhance the richness and quality of the generated image, surpassing the simplicity of the input prompt but also to preserve the core components specified in the prompt. By strategically leveraging the strengths of both the SD model and the proposed models, we ensure that the resulting image is not only aesthetically superior but also faithfully aligned with the essential elements outlined in the initial prompt.

2.1 Overview of Pipelines

The implemented pipeline commences with the original prompt as its input, initiating an iterative process aimed at crafting a suitable image. Within this iterative loop, the prompt undergoes refinement, subsequently, by one of four models: LLaVA 7b, LLaVA 13b, CogVLM, or GPT-4V, in order to leverage their expansive language model capabilities (The Enhance function). The enhanced prompt is then forwarded to the SD model to generate a cor-

responding image (The Generate function). Subsequently, the generated image undergoes scrutiny by one of four above models, again. Each model refines the image according to its unique capabilities. Following refinement, the image is assessed alongside the original prompt to verify the congruence of the refined prompt and the image with the user's original input (The Verify function). Successful verification results in the output of the process. However, if the image fails verification, the pipeline iteratively refines the original prompt again to generate new images. To prevent excessive time consumption, a maximum iteration limit is set at 5. If the process fails to produce an improved image aligning with the original prompt after 5 attempts, the last generated image becomes the final output. This iterative approach balances refinement and verification, optimizing the generation of suitable images while efficiently managing computational resources. Our proposed pipeline is outlined in Algorithm 1 and Figure 1 illustrates our pipeline with examples.

Algorithm 1 Modified Pseudo-code of our pipeline

```
Input: descr: Description of the image;  
max_iter: Maximum iteration;  
agent: Multimodal Large Language Model;  
Output: Image generated by improved prompt  
i  $\leftarrow$  0  
while i  $<$  max_iter do  
    prompt  $\leftarrow$  agent.Enhance(descr)  
    image  $\leftarrow$  SD.Generate(prompt)  
    if agent.Verify(image, descr) then  
        break  
    end if  
    i  $\leftarrow$  i + 1  
    if i == max_iter then  
        break  
    end if  
end while  
return image
```

2.2 Prompt Enhancement Phase

The pivotal prompt enhancement phase within the proposed pipeline involves iterative refinement facilitated by a series of models, including LLaVA-7b, LLaVA-13b, CogVLM, and GPT-4V. Each model leverages its unique capabilities to refine the original prompt, augmenting and elaborating upon it to provide more intricate and detailed instructions for subsequent image generation. This

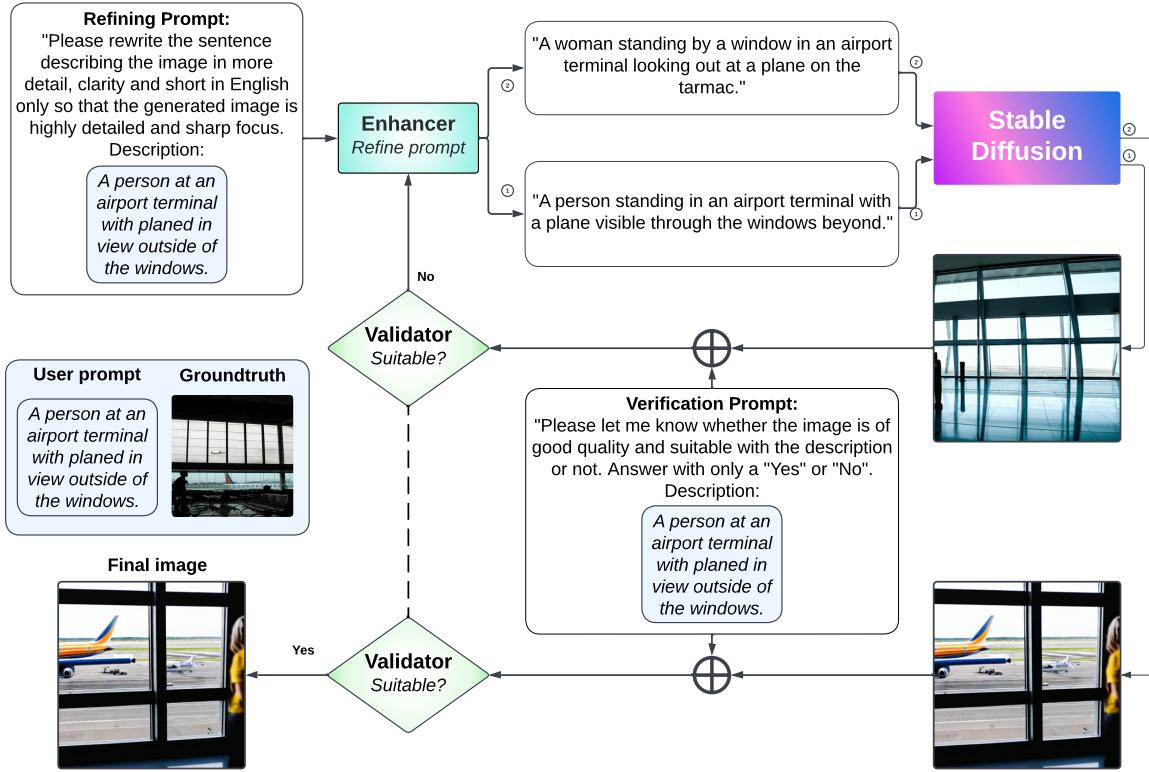


Figure 1: Overview of our pipeline, where **Enhancer** and **Validator** is one of our 4 models

163 iterative refinement process enables the models to
 164 progressively elevate the user’s input prompt, tailoring
 165 it to solicit a more nuanced and elaborate image
 166 output. The refined prompt assumes a pivotal role
 167 as the primary input for the SD model, ensuring
 168 that the image generation process is guided by so-
 169 phisticated instructions. This meticulous guidance,
 170 facilitated by the collective efforts of the models,
 171 ultimately contributes to the generation of high-
 172 quality, contextually relevant images.

173 2.3 Image Generation Phase

174 During the image generation phase, the refined
 175 prompt, enhanced through iterative improvements
 176 by LLaVA-7b, LLaVA-13b, CogVLM, and GPT-4V,
 177 serves as the input for the SD model. Leverag-
 178 ing its capabilities, the SD model employs diffusion
 179 processes to synthesize a high-quality image that
 180 aligns with the refined prompt.

181 The diffusion model, designed for stability, ensur-
 182 es the coherence and fidelity of the generated
 183 image. It operates as a critical component in the
 184 iterative loop, facilitating the translation of textual
 185 information into visual representations. The result-
 186 ing image is a manifestation of the collaborative

187 efforts between the various models’ linguistic en-
 188 hancements and the SD model’s image synthesis
 189 capabilities, encapsulating the details and nuances
 190 specified in the refined prompt.

191 2.4 Image Verification Phase

192 LLaVA-7b, LLaVA-13b, CogVLM, and GPT-4V
 193 models, play a crucial role in assessing the compati-
 194 bility between the generated image and the original
 195 input prompt. The purpose of this verification pro-
 196 cess is to ensure that the refined prompt, utilized
 197 in generating the image, maintains alignment with
 198 the user’s original prompt. The image verification
 199 serves as a safeguard against potential errors that
 200 may arise during the refinement of the prompt by
 201 the models or when the SD model encounters chal-
 202 lenges in creating an image that precisely aligns
 203 with the improved prompt.

204 The models act as discerning evaluators, exam-
 205 ining both the linguistic nuances in the refined
 206 prompt and the visual elements captured in the gen-
 207 erated image. If the image successfully aligns with
 208 the original input prompt according to the models’
 209 assessment, it is deemed verified and stands as the
 210 output of the process. By subjecting the refined

prompt and the generated image to the discerning scrutiny of the models in conjunction with the original input prompt, we create a fail-safe mechanism that detects and rectifies any potential errors in the image-generation process.

3 Experiments

3.1 Dataset

The Microsoft COCO, or Common Objects in Context, dataset (Lin et al., 2014) is an extensive collection of real-world images employed for the training and assessment of computer vision algorithms. What sets COCO apart from other datasets is its emphasis on comprehending objects within the context of their surroundings. Unlike datasets that isolate objects within controlled environments, COCO places them within the intricate and unstructured reality of our world. While this introduces added complexity for algorithms, it also renders the dataset more pertinent to real-world applications.

As the field of computer vision undergoes continuous evolution, the Microsoft COCO dataset is poised to assume a pivotal role in the development of novel and innovative applications. In the context of our research, we employ an inference method on the validation set from COCO version 2017. This subset consists of 5000 samples, each comprising two features: The image itself and a corresponding description of the image. This specific dataset configuration provides a rich and diverse set of examples, enabling the exploration and evaluation of various aspects of computer vision tasks.

3.2 Experiment Setup and Evaluation Metrics

To test our proposed pipelines, we feed all 5000 images and their descriptions from validation set of the MS-COCO dataset ((Lin et al., 2014)). For evaluation metrics, we compute Fréchet Inception Distance (FID, Heusel et al. (2017)) and Inception Score (IS, Salimans et al. (2016)) for the Text-to-Image models by comparing generated images with real images from dataset. FID and IS are computed with torch-fidelity.

3.2.1 Baselines

To evaluate the efficacy of our methodology, we conducted a series of experiments, encompassing the generation of images through the utilization of the SD model under various conditions. Specifically, the experiments included generating images based on the original textual descriptions, generating images utilizing the SD model with improved

descriptions refined by the LLaVA-7b model, generating images with improved descriptions refined by the LLaVA-13b model (both of these models are experienced with version 1.5 and version 1.6), generating images with improved descriptions refined by the CogVLM mode, and finally generating images with improved descriptions refined by the GPT-4V model, while incorporating an additional verification step.

The primary objective of these experiments was to systematically assess and compare the performance of the SD model when operating on distinct input prompts derived from our introduced pipeline. This comparative analysis allows us to discern variations in the generative capacity of the model based on the nature and refinement of the input prompts to produce better and more detailed images based on relatively simple user prompts.

3.2.2 Settings

In this study, we employed Stable Diffusion v1.5 (Rombach et al., 2022) in conjunction with models including: LLaVA (Liu et al., 2023) version 1.5 and version 1.6, CogVLM (Wang et al., 2023; Hong et al., 2023) , and GPT-4V (OpenAI, 2023). These models will transform the initial prompts into a more detailed and clear instructions to enhance the quality of the generated images.

Prompt Settings: At first, the initial prompt guiding our cooperating models is formulated as follows:

"Please rewrite the sentence describing the image in more detail, clarity and short in English only so that the generated image is highly detailed and sharp focus.

Description: <ORIGINAL DESCRIPTION>."

This prompt proves advantageous by explicitly instructing the models to enhance the description for SD image generation. The detailed and clear instructions mitigate issues arising from simple prompts, providing the models with the necessary information to refine and augment the prompt effectively.

Additionally, for experiments incorporating a verification step, we designed a specific prompt to guide the models in assessing the generated image's quality. The verification prompt is structured as follows:

"Please let me know whether the image is
 of good quality and suitable with the
 description or not. Answer with only a
 "Yes" or "No".
 Description: <ORIGINAL
 DESCRIPTION>."

The verification prompt addresses the crucial task of assessing the alignment between the generated image and the original prompt. By instructing our models to provide a binary response ("Yes" or "No"), this prompt streamlines the verification process, ensuring quick and decisive evaluations. The maximum iteration for verification is set at 5, signifying that if the pipeline fails to produce a satisfactory image after five attempts, the last generated image becomes the final output. This iterative approach balances refinement and verification, optimizing the generation of suitable images while efficiently managing computational resources.

GPT-4 Settings: In our pursuit of optimizing the GPT-4V model's performance across both prompt refinement and image verification phases, we carefully selected a default configuration for three pivotal parameters: Temperature, top p , and top k . This configuration, aimed at refining initial prompts for enhanced image generation and assessing alignment between generated images and original prompts, serves as a cornerstone for consistency and reliability throughout the text-to-image generation pipeline. Maintaining the temperature at 0.7, top p at 0.95, and top k at 800, our default parameter settings for the GPT-4V model establish a robust and dependable foundation. These settings prioritize coherence, diversity, and detail in the generated text, fostering improved performance in both prompt refinement and image verification tasks. By adhering to this selected configuration, we ensure a seamless and effective execution of the text-to-image generation process.

CogVLM Settings: To optimize the performance of the CogVLM model for each phase, we also carefully tuned its temperature, top p , and top k configurations. During the prompt improvement phase, aimed at refining the initial prompts for enhanced image generation, we set the temperature parameter to 0.6, the p parameter to 0.8, and k is 100. In contrast, during the image verification phase, where the focus shifts to assessing the alignment between generated images and original prompts, we adjusted the parameters to differ-

ent values. Specifically, we set the temperature to 0.5, slightly reducing the diversity of the generated text to ensure more consistent and focused evaluations. Furthermore, we decreased p to 0.5, limiting the probability mass for token generation and promoting more deterministic behavior. Finally, we lowered k to 10, restricting the number of tokens considered during sampling to prioritize simple and concise response.

3.3 Results

After finish running our experiments, the result brings back some positive results. The metrics include the Fréchet Inception Distance (FID) and Inception Score (IS), which serve as quantitative indicators of image quality and diversity. Table 1 shows the evaluation results of our pipeline compared with other ones.

Fréchet Inception Distance (FID) score: FID serves as a crucial metric in evaluating the fidelity of generated images compared to real ones. In this study, we observed notable variations in FID scores across different model iterations. A lower FID value, as observed without using any model to enhance or validating the prompt yields a score of 24.95, suggests relatively better image quality and closer resemblance to real images. However, as iterations progressed, we witnessed fluctuations in FID scores across various models.

It can also be observed that iterations involving prompt enhancements and validating prompts showcased improvements in FID scores, indicating enhanced fidelity and reduced perceptual distance from real images. Conversely, the increase in FID score for the LLaVA-7b v1.5 model iteration suggests potential challenges in maintaining image quality while introducing prompt enhancements. Or even some trending models like GPT-4V and CogVLM also yields quite a high FID comparing to the original description. These observations underscore the intricate balance required between prompt manipulation and image fidelity to ensure the generation of high-quality images. Using LLaVA-13b v1.5 for both prompt enhancement and validation results in a fairly descend FID score with respect to the original description.

Inception Score (IS): IS serves as a metric to assess the diversity and quality of generated images. Higher IS values signify greater image diversity and quality. Our analysis revealed interesting trends in IS scores across different model iterations. Models such as LLaVA-7b v1.6 and LLaVA-13b

Table 1: Evaluation results on the MS-COCO dataset

Prompt Enhancer	Validator	FID \downarrow	IS \uparrow	Iteration
None	None	24.95 ± 0.51	33.70 ± 0.47	1
LLaVA1.5-7b	None	30.00 ± 0.60	31.64 ± 0.47	1
LLaVA1.5-13b	None	25.59 ± 0.51	33.46 ± 0.48	1
LLaVA1.6-7b	None	28.85 ± 0.55	35.72 ± 0.53	1
LLaVA1.6-13b	None	29.56 ± 0.56	35.22 ± 0.52	1
GPT-4	None	29.78 ± 0.55	35.63 ± 0.52	1
CogVLM	None	29.71 ± 0.58	34.42 ± 0.53	1
LLaVA1.5-7b	LLaVA1.5-7b	26.10 ± 0.51	33.94 ± 0.52	1.23 ± 0.57
LLaVA1.5-13b	LLaVA1.5-13b	24.37 ± 0.46	34.61 ± 0.50	2.39 ± 1.64
LLaVA1.6-7b	LLaVA1.6-7b	28.07 ± 0.53	35.80 ± 0.51	1.01 ± 0.09
LLaVA1.6-13b	LLaVA1.6-13b	28.50 ± 0.54	35.78 ± 0.50	1.27 ± 0.64
GPT-4	GPT-4	29.78 ± 0.57	35.83 ± 0.52	2.32 ± 1.69
CogVLM	CogVLM	29.80 ± 0.58	34.53 ± 0.51	1.10 ± 0.53

399 v1.6 demonstrated consistent improvements in IS
400 scores over iterations, indicating enhanced image
401 diversity while maintaining quality.

402 Besides, the comparison between LLaVA-5b
403 v1.5 and LLaVA-7b v1.5 iterations highlighted the
404 impact of prompt enhancements on image diver-
405 sity. Despite similar FID scores, LLaVA-13b v1.5
406 exhibited a higher IS, suggesting superior image
407 diversity resulting from prompt modifications. Ad-
408 ditionally, the higher IS observed in GPT-4V iter-
409 ations compared to CogVLM underscores potential
410 differences in the diversity of generated images
411 between these models. GPT-4V proves superior
412 image generation diversity when achieving 35.83
413 IS score, indicating its efficiency when using it in
414 prompt improvement phase and verification phase.

415 To make it clearer, Figure 2 and Figure 3 illus-
416 trate some examples from the generated pipelines
417 after various improvements. The presented results
418 are based on multiple iterations, showcasing the
419 evolution of the pipelines over time. These im-
420 provements showcase the impact of algorithmic ad-
421 justments on the quality of generated outputs. For
422 example, the original image description serves as a
423 baseline, generating images related to the prompt,
424 yet with limited intricacies. As we progress through
425 the enhancements, such as using LLaVA-7b, the
426 generated images exhibit a noticeable refinement in
427 capturing the details of the two monkeys and their
428 surroundings. This improvement becomes even
429 more pronounced with the integration of LLaVA-
430 13b, showcasing a significant leap in the quality of
431 generated outputs.

3.4 Limitations of Our Experiment

432 Our experiment is subject to several limitations that
433 may impact the interpretation and generalizability
434 of our findings. Firstly, resource constraints have
435 imposed limitations on the scale and depth of our
436 experimentation. Limited computational resources
437 may have constrained the complexity and size of
438 the models employed, potentially affecting their
439 performance and the breadth of insights derived
440 from our analyses.

442 Furthermore, while prompt engineering repre-
443 sents a promising approach to guide model behav-
444 ior and output generation, its efficacy in our experi-
445 ment may have been compromised by suboptimal
446 prompt design and formulation. The intricacies
447 of prompt engineering, including the selection of
448 keywords, formatting, and contextual framing, ne-
449 cessitate careful consideration and refinement to
450 effectively steer model outputs towards desired ob-
451 jectives. Our experiment acknowledges the ongo-
452 ing challenges in optimizing prompt engineering
453 strategies, and future research endeavors should
454 prioritize the development of more robust and ef-
455 fective prompt design methodologies.

456 Additionally, our experiment did not comprehen-
457 sively control for multimodal outputs according to
458 predefined criteria or preferences. Multimodal gen-
459 eration, encompassing diverse modalities such as
460 text and image, introduces complexities in model
461 behavior and output interpretation. The lack of ex-
462 plicit control over multimodal outputs according
463 to specific preferences or constraints may have in-

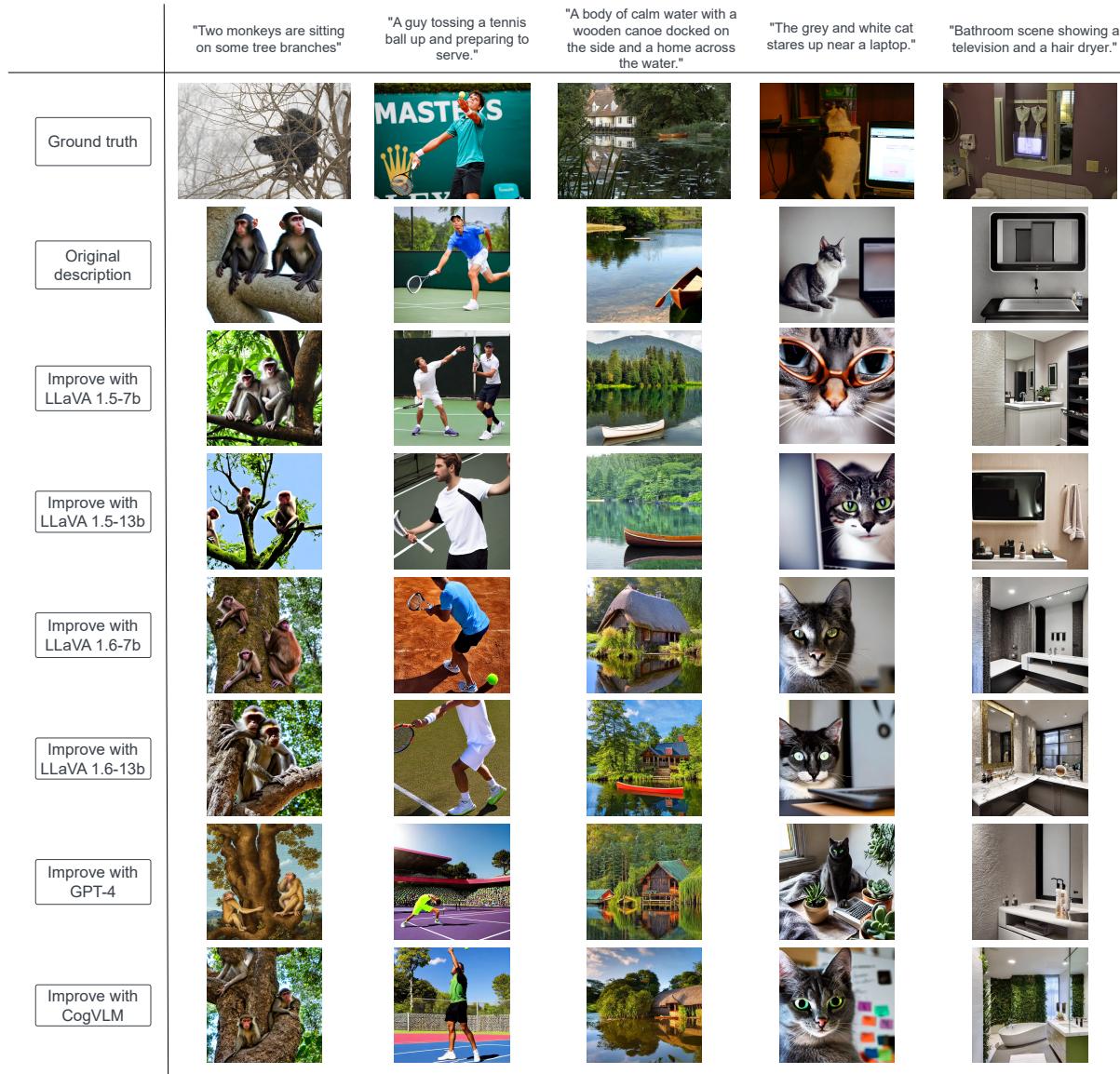


Figure 2: Comparing generated images giving different pipelines to improve prompting input, without verification phase.

464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
troduced variability and unpredictability in model performance, potentially limiting the reliability and reproducibility of our results.

477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
Therefore, future research endeavors should prioritize these considerations to enhance the robustness and validity of findings in the field of generative modeling and artificial intelligence. Incorporating verification from MLLMs into the fine-tuning process of Diffusion represents a promising avenue for future research endeavors in generative modeling.

493 3.5 Related Potential Risks

494
495
496
497
498
499
500
The generation of synthetic data, particularly in the context of image generation, may introduce bi-

ases in the training data or introduce new biases. These biases could potentially reinforce harmful stereotypes or contribute to unethical applications if not adequately addressed. The deployment of generative models in real-world applications, especially if they involve sensitive or confidential data, could pose security risks if malicious actors exploit vulnerabilities in the models to generate deceptive or malicious content. Moreover, the use of generative models, particularly in regulated domains such as healthcare or finance, may raise legal concerns regarding data privacy, intellectual property rights, and compliance with existing regulations. Finally, the deployment of generative models in real-world scenarios could lead to unforeseen consequences or

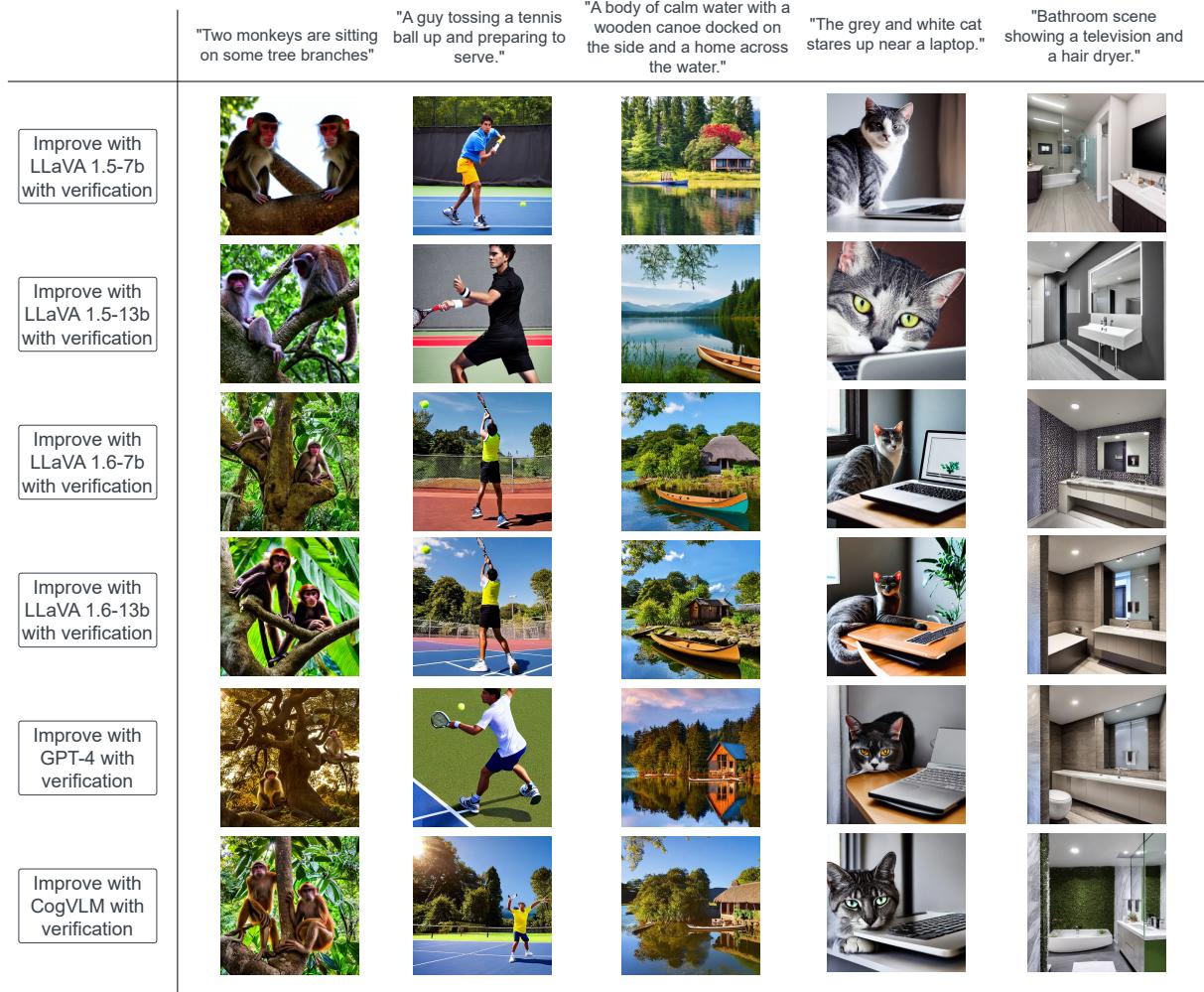


Figure 3: Comparing generated images giving different pipelines to improve prompting input, with additional verification phase.

493 unintended uses. These unintended consequences
494 may emerge due to the complex interplay between
495 the model’s behavior, societal dynamics, and tech-
496 nological advancements.

4 Conclusion

498 While text-to-image models have made impres-
499 sive strides, their reliance on simple, often vague
500 prompts often results in underwhelming outcomes.
501 The disconnect between user intent and generated
502 image arises from the model’s limited under-
503 standing of nuances and details. To bridge this gap, we
504 propose a revolutionary approach that leverages the
505 power of MLLMs.

506 Instead of relying solely on user input, our sys-
507 tem harnesses the MLLM’s prowess to enrich the
508 initial prompts. This enriched prompt equips the
509 image generation model with a deeper under-
510 standing of the user’s vision, allowing it to translate

511 intent into stunning visuals.

512 But our system does not stop at a single attempt.
513 The MLLMs then steps in again, analyzing the
514 generated image for alignment with the enhanced
515 prompt. If discrepancies exist, it provides feedback
516 for further refinement. This iterative loop of gen-
517 eration and evaluation ensures the final image not
518 only matches the user’s vision but also reflects the
519 intricate details woven into the enhanced prompt.
520 By empowering text-to-image with MLLM-driven
521 prompt enhancement, we unlock a future where
522 creativity knows no bounds. This is not just a tech-
523 nological advancement; it’s a bridge between imag-
524 ination and reality, a canvas where words transform
525 into a tapestry of meticulously crafted visuals.

References

- 526 Aishwarya Agarwal, Srikrishna Karanam, KJ Joseph,
527 Apoorv Saxena, Koustava Goswami, and Bal-
529 aji Vasan Srinivasan. 2023. A-star: Test-time at-
530 tention segregation and retention for text-to-image
531 synthesis. *IEEE International Conference on Com-
532 puter Vision*.
- 533 Andrew Brock, Jeff Donahue, and Karen Simonyan.
534 2019. Large scale GAN training for high fidelity
535 natural image synthesis. In *International Conference
536 on Learning Representations*.
- 537 Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf,
538 and Daniel Cohen-Or. 2023. Attend-and-excite:
539 Attention-based semantic guidance for text-to-image
540 diffusion models. *ACM Transactions on Graphics
541 (TOG)*, 42(4):1–10.
- 542 Wei-Ge Chen, Irina Spiridonova, Jianwei Yang, Jian-
543 feng Gao, and Chunyuan Li. 2023. Llava-interactive:
544 An all-in-one demo for image chat, segmenta-
545 tion, generation and editing. *arXiv preprint
546 arXiv:2311.00571*.
- 547 Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani,
548 Arjun Akula, Pradyumna Narayana, Sugato Basu,
549 Xin Eric Wang, and William Yang Wang. 2022.
550 Training-free structured diffusion guidance for com-
551 positional text-to-image synthesis. *The Eleventh
552 International Conference on Learning Representations*.
- 553 Martin Heusel, Hubert Ramsauer, Thomas Unterthiner,
554 Bernhard Nessler, and Sepp Hochreiter. 2017. Gans
555 trained by a two time-scale update rule converge to a
556 local nash equilibrium. *Advances in neural infor-
557 mation processing systems*, 30.
- 558 Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng
559 Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang,
560 Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cog-
561 agent: A visual language model for gui agents.
- 562 Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Ca-
563 ballero, Andrew Cunningham, Alejandro Acosta, An-
564 drew Aitken, Alykhan Tejani, Johannes Totz, Zehan
565 Wang, et al. 2017. Photo-realistic single image super-
566 resolution using a generative adversarial network. In
567 *Proceedings of the IEEE conference on computer
568 vision and pattern recognition*, pages 4681–4690.
- 569 Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins,
570 Yuqing Du, Craig Boutilier, P. Abbeel, Mohammad
571 Ghavamzadeh, and Shixiang Shane Gu. 2023. Align-
572 ing text-to-image models using human feedback.
573 *ArXiv*, abs/2302.12192.
- 574 Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi.
575 2023. Blip-2: Bootstrapping language-image pre-
576 training with frozen image encoders and large lan-
577 guage models. *International Conference on Machine
578 Learning*.
- 579 Tsung-Yi Lin, Michael Maire, Serge Belongie, James
580 Hays, Pietro Perona, Deva Ramanan, Piotr Dollár,
and C Lawrence Zitnick. 2014. Microsoft coco:
581 Common objects in context. In *Computer Vision-
582 ECCV 2014: 13th European Conference, Zurich,
583 Switzerland, September 6-12, 2014, Proceedings,
584 Part V 13*, pages 740–755. Springer.
- 585 Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae
586 Lee. 2023. Improved baselines with visual instruc-
587 tion tuning. *arXiv preprint arXiv:2310.03744*.
- 588 Lanlan Liu, Michael Muelly, Jia Deng, Tomas Pfis-
589 ter, and Li-Jia Li. 2019. Generative modeling for
590 small-data object detection. In *Proceedings of the
591 IEEE/CVF International Conference on Computer
592 Vision*, pages 6073–6081.
- 593 Ming-Yu Liu and Oncel Tuzel. 2016. Coupled gen-
594 erative adversarial networks. *Advances in neural
595 information processing systems*, 29.
- 596 Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and
597 Joshua B Tenenbaum. 2022a. Compositional visual
598 generation with composable diffusion models. In
599 *European Conference on Computer Vision*, pages
600 423–439. Springer.
- 601 Rosanne Liu, Dan Garrette, Chitwan Saharia, William
602 Chan, Adam Roberts, Sharan Narang, Irina Blok,
603 RJ Mical, Mohammad Norouzi, and Noah Constant.
604 2022b. Character-aware models improve visual text
605 rendering. *Annual Meeting of the Association for
606 Computational Linguistics*.
- 607 OpenAI. 2023. Gpt-4 technical report. Technical report.
- 608 Charan D Prakash and Lina J Karam. 2021. It gan do
609 better: Gan-based detection of objects on images
610 with varying quality. *IEEE Transactions on Image
611 Processing*, 30:9220–9230.
- 612 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya
613 Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sas-
614 try, Amanda Askell, Pamela Mishkin, Jack Clark,
615 et al. 2021. Learning transferable visual models from
616 natural language supervision. In *International confer-
617 ence on machine learning*, pages 8748–8763. PMLR.
- 618 Robin Rombach, Andreas Blattmann, Dominik Lorenz,
619 Patrick Esser, and Björn Ommer. 2022. High-
620 resolution image synthesis with latent diffusion mod-
621 els. In *Proceedings of the IEEE/CVF conference
622 on computer vision and pattern recognition*, pages
623 10684–10695.
- 624 Tim Salimans, Ian Goodfellow, Wojciech Zaremba,
625 Vicki Cheung, Alec Radford, and Xi Chen. 2016.
626 Improved techniques for training gans. *Advances in
627 neural information processing systems*, 29.
- 628 Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi
629 Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang,
630 Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi
631 Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023.
632 Cogvlm: Visual expert for pretrained language mod-
633 els.
- 634

635

A Appendix

636 A.1 Demonstrating our pipeline process by 637 using LLaVA 1.5-7b with verification

638 A.1.1 Example 1

- 639 **Image description:** "Two monkeys are sitting
640 on some tree branches."
- 641 **Loop 1:**
 - 642 **Refined prompt:** "2 monkeys perched
643 on tree limbs."
 - 644 **Generated image:**



- 644 **Accepted:** YES

646 A.1.2 Example 2

- 647 **Image description:** "A guy tossing a tennis
648 ball up and preparing to serve."
- 649 **Loop 1:**

- 650 **Refined prompt:** "The scene captures
651 an intense moment during a game of
652 tennis. The player, positioned on the left
653 side of the frame, is in action, tossing
654 the sports ball up with precision as he
655 gets ready to serve it over the net. He
656 holds his racquet at the ready, focused
657 intently on executing a powerful and ac-
658 curate shot. The image captures the dy-
659 namic nature of tennis, where each point
660 can be filled with excitement and antici-
661 pation."
- 662 **Generated image:**
- 663 **Accepted:** YES

664 A.1.3 Example 3

- 665 **Image description:** "A body of calm water
666 with a wooden canoe docked on the side and
667 a home across the water."



- 668 **Loop 1:**

- 669 **Refined prompt:** "The image features
670 a peaceful scene where a small wooden
671 boat is gently floating on top of a serene
672 lake or pond, resting near the shore. On
673 the other side of the water, a charming
674 home can be seen in the background,
675 creating a picturesque and idyllic atmo-
676 sphere.."
- 677 **Generated image:**



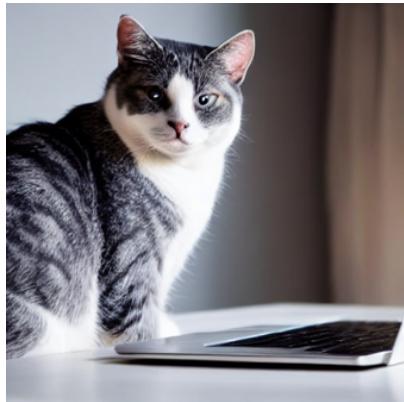
- 678 **Accepted:** YES

679 A.1.4 Example 4

- 680 **Image description:** "The grey and white cat
681 stares up near a laptop."

- 682 **Loop 1:**

- 683 **Refined prompt:** "A close-up of an in-
684 door grey and white cat sitting next to
685 a laptop computer, displaying alertness
686 and curiosity as it gazes at something
687 beyond the frame."
- 688 **Generated image:**



– Accepted: YES

A.1.5 Example 5

- **Image description:** "Bathroom scene showing a television and a hair dryer."

• Loop 1:

- **Refined prompt:** "Image of a bathroom with a TV mounted above the countertop near a sink and cup, accompanied by a hairdryer placed nearby."

– Generated image:



– Accepted: YES

A.1.6 Example 6

- **Image description:** "a couple of buckets in a white room"

• Loop 1:

- **Refined prompt:** "2 clear plastic buckets placed on a white countertop or table in a clean room."

– Generated image:



– Accepted: YES

A.1.7 Example 7

- **Image description:** "A person takes a picture in a hotel bathroom."

• Loop 1:

- **Refined prompt:** "A guest captures an image of themselves inside their hotel room's bathroom using a mirror reflection."

– Generated image:



– Accepted: YES

A.2 Demonstrating our pipeline process by using LLaVA 1.5-13b with verification

A.2.1 Example 1

- **Image description:** "Two monkeys are sitting on some tree branches."

• Loop 1:

- **Refined prompt:** "2 monkeys are perched on branches of a tree."

– Generated image:



– Accepted: YES

A.2.2 Example 2

- **Image description:** "A guy tossing a tennis ball up and preparing to serve."

• Loop 1:

- **Refined prompt:** "Man holding a tennis racket on a court, preparing to hit the tennis ball."

– Generated image:



– Accepted: YES

A.2.3 Example 3

- **Image description:** "A body of calm water with a wooden canoe docked on the side and a home across the water."

• Loop 1:

- **Refined prompt:** "A serene lake scene featuring a wooden canoe resting peacefully on the shore, surrounded by lush greenery. In the distance, a picturesque home adds to the tranquil atmosphere."

– Generated image:



– Accepted: YES

A.2.4 Example 4

- **Image description:** "The grey and white cat stares up near a laptop."

• Loop 1:

- **Refined prompt:** "A close-up of a grey and white cat sitting on the floor next to an open laptop computer. The cat appears to be looking up, possibly at something or someone above it. The laptop takes up a significant portion of the frame, extending from the left edge almost all the way to the right side of the image. The focus is sharp and clear, capturing the details of both the cat and the laptop."

– Generated image:



– Accepted: NO

• Loop 2:

- **Refined prompt:** "The cat is sitting next to a laptop computer."

748

728

729

730

731

732

733

734

735

736

749

750

751

752

753

754

755

756

757

758

759

760

761

762

763

764

765

737

738

739

740

741

742

743

744

745

746

747

– **Generated image:**

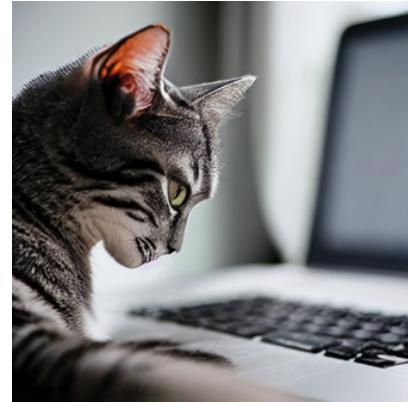


– **Accepted:** NO

• **Loop 3:**

– **Refined prompt:** "The cat, a mix of grey and white fur, intently gazes into the distance as it sits next to an open laptop computer."

– **Generated image:**



– **Accepted:** YES

A.2.5 Example 5

• **Image description:** "Bathroom scene showing a television and a hair dryer."

• **Loop 1:**

– **Refined prompt:** "A bathroom featuring a hairdryer placed near the sink and a TV mounted on the wall above it."

– **Generated image:**



– **Accepted:** NO

– **Accepted:** YES

• **Loop 4:**

– **Refined prompt:** "A gray and white cat looking up next to an open laptop. The cat appears focused on something above it, while the laptop is placed nearby, possibly indicating that someone was working or studying when they were interrupted by the curious feline."

– **Generated image:**

A.2.6 Example 6

• **Image description:** "a couple of buckets in a white room"

• **Loop 1:**

– **Refined prompt:** "2 buckets placed against wall in all-white room."

– **Generated image:**



805 – Accepted: NO

806 • Loop 2:

807 – Refined prompt: "2 white buckets in a
808 white room."

809 – Generated image:



810 – Accepted: YES

811 A.2.7 Example 7

812 • Image description: "A person takes a picture
813 in a hotel bathroom."

814 • Loop 1:

815 – Refined prompt: "The photo depicts
816 a person standing inside a well-lit bath-
817 room of a hotel room, holding up a cam-
818 era to take a picture. The bathroom fea-
819 tures a sink situated towards the left side
820 and a large mirror above it. The reflec-
821 tion in the mirror captures the essence
822 of the scene. The room's interior design
823 showcases a mix of comfort and func-
824 tionality, with a cozy bed located on the
825 right side of the image."

826 – Generated image:



827 – Accepted: NO

828 • Loop 2:

829 – Refined prompt: "The scene depicts an
830 individual standing inside a modern ho-
831 tel bathroom while taking a photo. The
832 bathroom features sleek white tiles on
833 the walls, floors, and countertops, which
834 create a clean and bright ambiance. The
835 person capturing the image is positioned
836 towards the right side of the frame, hold-
837 ing up their smartphone to take the per-
838 fect shot. A small handbag can be seen
839 placed near the bottom-left corner of the
840 image, likely belonging to the person tak-
841 ing the photo."

842 – Generated image:



843 – Accepted: YES

844 A.3 Demonstrating our pipeline process by
845 using LLaVA 1.6-7b with verification

846 A.3.1 Example 1

847 • Image description: "Two monkeys are sitting
848 on some tree branches."

849 • Loop 1:

850
851
852
853
854
855
856
857
858
859
860
861

– **Refined prompt:** "Amid the lush verdure of a dense tropical rainforest, two capricious monkeys are perched with poise on the gnarled, intertwining branches of an ancient fig tree. Their keen eyes, alight with mischief, scan the verdant canopy, while their nimble tails curl gracefully behind them. Sunlight filters through the dense foliage, dappling their sleek, russet fur with patterns of light and shadow, casting a sharp, vivid tableau of jungle life."



862
863

– **Generated image:**



864
865
866

– **Accepted:** YES

A.3.2 Example 2

• **Image description:** "A guy tossing a tennis ball up and preparing to serve."

867

• **Loop 1:**

868
869
870
871
872
873
874
875

– **Refined prompt:** "A focused athlete, with his feet firmly planted on the sun-kissed clay court, arches his back gracefully as he launches a neon-yellow tennis ball high into the clear, azure sky. His poised racket awaits the perfect moment to meet the ball in a powerful serve."

– **Generated image:**

876

– **Accepted:** YES

A.3.3 Example 3

877
878
879
880

• **Image description:** "A body of calm water with a wooden canoe docked on the side and a home across the water."

881

• **Loop 1:**

882
883
884
885
886
887
888

– **Refined prompt:** "A placid azure expanse is graced by a rustic wooden canoe, tethered gently to an aged, sun-bleached dock. Lush foliage envelops the scene, and a quaint cottage with a thatched roof sits serenely on the water's far edge, its reflection shimmering in the stillness."

889

– **Generated image:**



890

– **Accepted:** YES

A.3.4 Example 4

891
892
893

• **Image description:** "The grey and white cat stares up near a laptop."

894

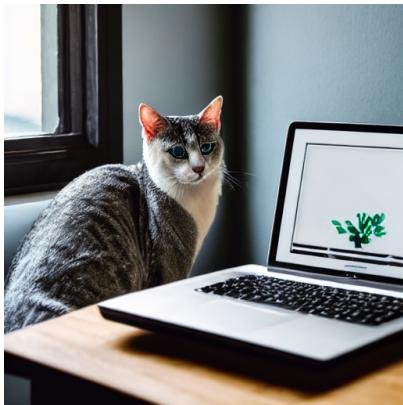
• **Loop 1:**

895
896
897

– **Refined prompt:** "A sleek grey and white feline, with piercing emerald eyes, fixates its gaze upward from beside a

898 sleek, modern laptop, which sits atop an
899 oak desk scattered with vibrant green pot-
900 ted plants and a steaming mug of coffee,
901 all bathed in the soft glow of morning
902 light filtering through a nearby window."

903 – **Generated image:**



904 – **Accepted: YES**

905 A.3.5 Example 5

906 • **Image description:** "Bathroom scene show-
907 ing a television and a hair dryer."

908 • **Loop 1:**

909 – **Refined prompt:** "The image captures a
910 pristine modern bathroom, its sleek mar-
911 ble tiles gleaming under the soft glow
912 of recessed lighting. Centered on a wall
913 of intricately patterned ceramic, a flat-
914 screen television displays a vibrant under-
915 water documentary, the colors vivid
916 against the monochromatic backdrop. To
917 the right, a state-of-the-art chrome hair
918 dryer is mounted neatly on the wall, its
919 cord coiled meticulously around the base,
920 awaiting use. Adjacent to these, a plush
921 white towel is draped with precision over
922 a polished stainless-steel towel warmer,
923 inviting a touch of cozy warmth to the
924 room. A vase with a single, dew-kissed
925 orchid beside the sink adds a splash of
926 nature's elegance to the scene."

927 – **Generated image:**



928 – **Accepted: YES**

929 A.4 Demonstrating our pipeline process by 930 using LLaVA 1.6-13b with verification

931 A.4.1 Example 1

932 • **Image description:** "Two monkeys are sitting
933 on some tree branches."

934 • **Loop 1:**

935 – **Refined prompt:** "Amidst the dense fo-
936 liage of a lush tropical rainforest, two ca-
937 puchin monkeys perch with intent grace
938 on the gnarled branches of an ancient
939 tree, their expressive faces alive with cu-
940 riosity as they survey their verdant do-
941 main. The afternoon sun filters through
942 the canopy, casting dappled light that
943 dances on their sleek fur, while a kalei-
944 doscope of vibrant flowers and ripe fruits
945 adds a burst of color to the scene."

946 – **Generated image:**



947 – **Accepted: YES**

948 A.4.2 Example 2

949 • **Image description:** "A guy tossing a tennis
950 ball up and preparing to serve."

951
952
953
954
955
956
957
958

• **Loop 1:**

- **Refined prompt:** "The image showcases a man in mid-air, having just thrown a tennis ball into the air and preparing to serve it with a tennis racket. The focus is on the man and his actions, capturing the anticipation and precision of the sport.",
- **Generated image:**



- **Accepted:** NO

• **Loop 2:**

- **Refined prompt:** "A focused athlete clad in a crisp white tennis outfit stands poised on the sun-drenched court, muscles tensed as he expertly flips a bright yellow tennis ball skyward with his left hand, while gripping his racquet firmly in his right, readying for a powerful serve."

- **Generated image:**



- **Accepted:** YES

A.4.3 Example 3

- **Image description:** "A body of calm water with a wooden canoe docked on the side and a home across the water."

• **Loop 1:**

- **Refined prompt:** "Glistening beneath the serene azure sky, a tranquil expanse of crystalline water lies undisturbed, save for the gentle brush of a soft zephyr. Along its mirror-like surface, a rustic wooden canoe rests gracefully, tethered to a quaint, weathered dock that whispers tales of many a dawn greeted. Across this placid basin, a charming cottage nestles amid a flourish of verdant foliage, its cozy silhouette reflecting in the water, a silent sentinel to the enduring peace."

- **Generated image:**



- **Accepted:** YES

A.4.4 Example 4

- **Image description:** "The grey and white cat stares up near a laptop."

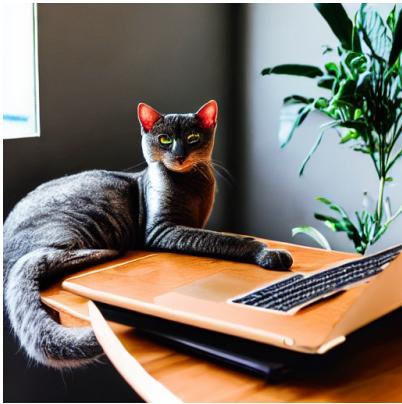
• **Loop 1:**

- **Refined prompt:** "A sleek, silver-furred cat with piercing emerald eyes gazes intently upward, perched elegantly beside a sleek, modern laptop that rests on a polished wooden desk, scattered with vibrant, green potted plants and a steaming mug of coffee, casting a warm glow under the soft, ambient light."

- **Generated image:**

974
975
976
977
978
979
980
981
982
983
984
985
986
987

988
989
990
991
992
993
994
995
996
997
998
999
1000
1001



– Accepted: YES

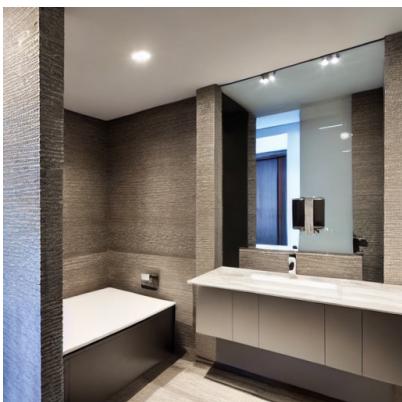
A.4.5 Example 5

- **Image description:** "Bathroom scene showing a television and a hair dryer."

• Loop 1:

– **Refined prompt:** "An immaculate, modern bathroom comes into view, boasting sleek, marble countertops that gleam under the soft, diffused lighting. Perched on a wall mount, a slim, high-definition television streams the morning news, its vibrant screen casting a gentle glow. Adjacent to the TV, a chrome hair dryer rests in its cradle, reflecting the light with its polished surface, ready for use. Delicate ferns in a small, ceramic pot add a touch of greenery beside the basin, enhancing the serene atmosphere. The entire scene is captured in sharp focus, highlighting the harmony of technology and comfort in this personal oasis."

– Generated image:



– Accepted: YES

A.5 Demonstrating our pipeline process by using GPT-4V with verification

A.5.1 Example 1

- **Image description:** "Two monkeys are sitting on some tree branches."

• Loop 1:

– **Refined prompt:** "Two capricious monkeys with chestnut-hued fur are perched attentively upon a labyrinth of sturdy, gnarled branches, nestled within an ancient, emerald-canopied tree, with the golden hues of the late afternoon sun peeking through the foliage, casting dappled light upon their playful antics."

– Generated image:



– Accepted: YES

A.5.2 Example 2

- **Image description:** "A guy tossing a tennis ball up and preparing to serve."

• Loop 1:

– **Refined prompt:** "A man, in a poised stance, expertly flips a tennis ball aloft, gearing up for a precise service motion."

– Generated image:

1025
1026

1027

1028
1029

1030

1031

1032
1033

1034
1035

1036
1037

1038
1039

1039

1040

1041

1042
1043

1044

1045
1046
1047

1048



– Accepted: NO

• Loop 2:

– **Refined prompt:** "A man, in a poised stance, is captured mid-motion as he expertly flips a tennis ball upward, gearing up for a precise and powerful service."

– Generated image:



– Accepted: NO

• Loop 3:

– **Refined prompt:** "A focused man, clad in a crisp white tennis outfit, expertly balances on the balls of his feet at the baseline. His arm is outstretched toward the clear blue sky, fingers poised to release a fluorescent yellow tennis ball. With his eyes locked on the ball, he coils his body, ready to unleash a powerful serve. The sun glints off his racket, positioned at the ready, while his shadow stretches across the freshly chalked court."

– Generated image:



– Accepted: YES

A.5.3 Example 3

• **Image description:** "A body of calm water with a wooden canoe docked on the side and a home across the water."

• Loop 1:

– **Refined prompt:** "A serene, glass-like water surface with a wooden canoe moored along the bank, and a house situated directly opposite on the far shore."

– Generated image:



– Accepted: NO

• Loop 2:

– **Refined prompt:** "A serene, glassy lake reflects the soft hues of twilight, with a rustic wooden canoe tethered to a weathered dock, and an inviting cottage nestled amongst verdant pines on the distant shore, exuding a sense of tranquil seclusion."

– Generated image:



– Accepted: YES

A.5.4 Example 4

- **Image description:** "The grey and white cat stares up near a laptop."

- **Loop 1:**

– **Refined prompt:** "With vivid intensity, a sleek grey and white cat, adorned with piercing emerald eyes, fixates its gaze upwards, perched gracefully beside an open silver laptop that gleams under the soft glow of a nearby desk lamp. The laptop, possibly awaiting its user's return, sits on a polished wooden table scattered with a few colorful sticky notes and a sleek, black pen resting parallel to the keyboard."

– **Generated image:**



– Accepted: YES

A.5.5 Example 5

- **Image description:** "Bathroom scene showing a television and a hair dryer."

- **Loop 1:**

– **Refined prompt:** "Detailed image of a pristine bathroom, featuring a modern wall-mounted television adjacent to a sleek hair dryer set against a glossy tiled backdrop."

– **Generated image:**



– Accepted: NO

- **Loop 2:**

– **Refined prompt:** "A meticulously detailed image of a modern bathroom, featuring a sleek, wall-mounted flat-screen television adjacent to a polished, contemporary hair dryer stationed on the marble countertop."

– **Generated image:**



– Accepted: NO

- **Loop 3:**

– **Refined prompt:** "A meticulously detailed and sharply focused image of a modern bathroom interior, featuring a sleek, wall-mounted flat-screen television adjacent to a gleaming hair dryer stationed on a marble countertop."

1113
1114
1115
1116
1117
1118

1119
1120
1121
1122
1123
1124
1125
1126
1127

1128
1129
1130
1131
1132
1133
1134
1135

1136

– **Generated image:**



1137

– **Accepted:** NO

1138

• **Loop 4:**

1139
1140
1141
1142
1143
1144

– **Refined prompt:** "Detailed image of a modern bathroom, featuring a sleek flat-screen television mounted on the wall and a professional hair dryer placed neatly on the marble countertop, both in sharp focus."

1145

– **Generated image:**



1146

– **Accepted:** NO

1147

• **Loop 5:**

1148
1149
1150
1151
1152
1153
1154
1155

– **Refined prompt:** "In the pristine bathroom, a sleek, wall-mounted flat-screen television faces the luxurious soaking tub, while a state-of-the-art, chrome-finished hair dryer rests beside a set of neatly folded, fluffy white towels on a gleaming marble countertop."

– **Generated image:**



– **Accepted:** YES

A.6 Demonstrating our pipeline process by using CogVLM with verification

A.6.1 Example 1

• **Image description:** "Two monkeys are sitting on some tree branches."

• **Loop 1:**

– **Refined prompt:** "In the dappled sunlight of a dense jungle, two capricious monkeys are perched with poise on the gnarled branches of an ancient mahogany tree, their expressive faces full of curiosity, their tails curling playfully. Amidst a lush tapestry of verdant leaves and vibrant, exotic flowers, the scene captures the essence of vibrant life, with a kaleidoscope of butterflies fluttering nearby and a vividly plumed parrot observing from a higher bough."

– **Generated image:**



– **Accepted:** YES

A.6.2 Example 2

• **Image description:** "A guy tossing a tennis ball up and preparing to serve."

1180 • **Loop 1:**

1181
1182
1183
1184
1185
1186
1187
1188
1189

– **Refined prompt:** "A focused athlete, dressed in a crisp white tennis outfit, skillfully balances on the balls of his feet at the baseline. With a sunlit court at his back, he tosses a vibrant yellow tennis ball high into the clear blue sky, his eyes locked on it as he coils his muscular arm back, gripping the racket with precision, poised to unleash a powerful serve."

1190 – **Generated image:**



1191 – **Accepted:** YES

1192 **A.6.3 Example 3**

1193
1194
1195

• **Image description:** "A body of calm water with a wooden canoe docked on the side and a home across the water."

1196 • **Loop 1:**

1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208

– **Refined prompt:** "Glistening under the tranquil azure sky, a serene expanse of crystal-clear water reflects the gentle embrace of dawn's light. Nestled at the water's edge, a rustic wooden canoe lies moored, its oars resting in silent anticipation. Across the still mirror-like surface, a quaint cottage with a thatched roof and smoke wisping from its stone chimney sits amidst a copse of whispering willows, its cozy presence offering a picture of idyllic solitude."

1209 – **Generated image:**



1210 – **Accepted:** YES

1211 **A.6.4 Example 4**

1212
1213

• **Image description:** "The grey and white cat stares up near a laptop."

1214 • **Loop 1:**

1215
1216
1217
1218
1219
1220
1221
1222
1223
1224

– **Refined prompt:** "A sleek, grey and white feline with piercing emerald eyes gazes intently upward, perched beside a sleek, modern laptop with a silver finish. Sunlight filters through a nearby window, casting a warm glow that dances across the cat's soft fur and highlights the meticulous craftsmanship of the laptop's design. A half-filled cup of steaming coffee sits just to the side, a silent testament to a paused moment of productivity."

1225 – **Generated image:**



1227 – **Accepted:** YES

1228 **A.6.5 Example 5**

1229
1230

• **Image description:** "Bathroom scene showing a television and a hair dryer."

1231 • **Loop 1:**

1232
1233
1234
1235
1236
1237
1238
1239

– **Refined prompt:** "A meticulously detailed image of a modern bathroom interior, featuring a sleek, wall-mounted flat-screen television opposite the vanity, and a professional-grade hair dryer stationed neatly on the marble countertop, all in crystal-clear, sharp focus."

– **Generated image:**



1240
1241

– **Accepted:** YES

• **Loop 2:**

1242
1243
1244
1245
1246
1247
1248

– **Refined prompt:** "A meticulously detailed image of a modern bathroom interior, featuring a sleek flat-screen television mounted on the wall and a professional hair dryer placed neatly on a marble countertop, all in sharp focus."

– **Generated image:**



1249
1250

– **Accepted:** NO

• **Loop 3:**

1251
1252
1253

– **Refined prompt:** "The image captures a pristine modern bathroom, where a sleek flat-screen television is mounted on the

cool gray marble wall opposite a luxurious soaking tub. Beside a clear glass shower enclosure, a high-end hair dryer with a polished chrome finish rests elegantly on a crisp white countertop dotted with fresh green succulents and fluffy folded towels, all bathed in the warm glow of recessed lighting."

– **Generated image:**



– **Accepted:** YES

A.7 Implementation:

Our implementation for this paper can be found here: <https://anonymous.4open.science/r/mllm-acl-2024-0963>

1254
1255
1256
1257
1258
1259
1260
1261
1262