Improve Image Text Descriptions using Large Language Models

N. Andriyanov [1], A. Kim [1]
[1] Financial University under the Government of the Russian Federation, 125167, Russia, Moscow, Leningradsky pr-t, b. 49

Abstarct

In this paper, a multi-stage approach to improve text queries (prompts) for image generation is proposed and comprehensively investigated. First, the GPT-2 model, pre-trained on 18,000 raw query-quality query pairs from the Lexica.art platform, automatically expands and refines the original prompts to produce more detailed and semantically accurate images when generated by the diffusion network Stable Diffusion 1.5. Next, the Image Captioning task (BLIP2) and four large language models (DeepSeek, Grok, ChatGPT, YandexGPT) are used to compare the quality of signature expansion, demonstrating different stylistic strategies for augmenting initial descriptions. The proposed "Captioning → Prompt Enhancer (Mistral) → Stable Diffusion+LoRA" Pipeline additionally includes a pre-training of Mistral's own model on BLEU, METEOR and CIDEr metrics, providing a steady increase in the quality of textual descriptions (BLEU from 0.12 to 0.435 after 500 epochs) and a significant reduction in the FID metric for image generation (from 0.3482 to 0.1873). In the final stage, the LoRA modules embedded in UNet and the Stable Diffusion text encoder allow efficient learning of the generation of previously "unknown" objects (rare or fictional), reducing FID to 0.172 at rank = 64. An expert survey (134 respondents) confirmed the visual preference of images generated by optimized queries, demonstrating the potential of the proposed technique to improve the quality of multimodal systems.

Keywords: image generation, large language models, prompt, text enhancement, multimodal models, LoRA.

Citation: Andriyanov N, Kim A. Improve Image Text Descriptions using Large Language Models. Computer Optics 2026; 50(4): 1860. doi: 10.18287/COJ1860.

Introduction

The current development of generative artificial intelligence encompasses both textual and visual models, but it is in multimodal generation combining both directions that significant progress has been made. Large Language Models (LLMs) have demonstrated high performance in text generation and comprehension tasks [1], and diffusion neural networks have shown outstanding capabilities in image synthesis from textual descriptions [2], [3]. In particular, the GPT-2 model demonstrated that pre-trained transformers are capable of solving a wide range of language tasks when pre-trained on narrowly focused samples, and the Latent Diffusion Models (Stable Diffusion) approach allowed generating high-quality images at reasonable computational cost.

Prompt engineering is a key factor affecting the quality of textual and visual content: even a perfect generative model cannot produce a detailed image without a correctly formulated prompt.Systematic reviews demonstrate that prompt optimization techniques provide significant gains in accuracy and expressiveness of the output without changing the parameters of the underlying model [4]. In particular, studies show that LLMs are able to automatically add relevant artistic and thematic tags to "raw" text queries, which improves the perceptual and semantic relevance of the generated images [5, 6].

Practical problems arise due to the fact that users often enter short and vague queries when they first encounter generative systems, resulting in undetailed or stylistically inexpressive images. It is shown in [7] and [8] that GPT-2 pre-training can be an effective tool in the task of improving queries from users to an image generator. The thesis shows that training on a simple query-expert prompt pair using 18,000 high-quality data pairs allows the model to effectively refine queries by adding detailed descriptions and keywords, resulting in lower FID metrics and higher user satisfaction.

Thus, generative models have recently become one of the most actively developing areas in artificial intelligence [9-11].They include both image-generating systems and text-generating algorithms. Especially notable are the successes of multimodal approaches [12, 13] that combine several types of data and demonstrate flexibility: for example, algorithms that automatically create captions for images [14] and models that turn textual descriptions into visual content [15].

At the same time, the quality of the resulting images is largely determined by the way the text query or "prompt" is formulated. Newcomers are often characterized by too brief and generalized queries, which results in poorly detailed or stylistically monotonous images and decreased interest in the platform. To improve users' first impression and increase retention, it is necessary to automatically enrich such "raw" prompts. In this paper, we propose a technique for pre-training the GPT-2 language model [16] on "user query → detailed prompt" example sets, which allows us to automatically add refining details and key tags to the raw descriptions, significantly improving the visual quality of the generated images. We also propose a holistic approach to the image generation task including the following steps:

1) Image Captioning by an automatic model (BLIP-2 [2]) generates a "primary" caption.

2) Comparative analysis of the extension of these captions by four LLMs (DeepSeek, Grok, ChatGPT, YandexGPT) to demonstrate different prompt engineering strategies.
3) The pre-trained Mistral model optimizes the caption by the metrics BLEU, METEOR and CIDEr, showing a steady increase in the quality of textual descriptions (BLEU from 0.12 to 0.435 after 500 epochs).
4) Integration of LoRA modules into Stable Diffusion 1.5 and text encoder allows to teach the generator to reproduce previously "unknown" objects (rare or fictional), reducing FID to 0.172 at rank = 64.

1. Materials and methods

Previously, image generation systems relied mainly on unimodal methods and often produced unstructured visual data resembling random noise. The modern approach is based on transforming a textual description - "prompt" - into an image, which significantly expands the possibilities of controlling the result. Various techniques for optimizing such queries have been proposed in the literature: some studies focus on pre-translation of user phrases from one language into English before generation, while others analyze in detail the impact of individual words and constructions on the final image, offering recommendations for selecting effective formulations.

In our study, we used the Stable Diffusion 1.5 diffusion model as the visual synthesis engine, pre-trained to create pixel-art style portraits. To automatically enrich "raw" user queries, we have pre-trained the GPT-2 language model on a corpus of 18,000 "prompt - high quality prompt" pairs from the Lexica.art platform. This allows the model to identify and add to the original descriptions the key details that most strongly influence the perception of a scene. The choice of these particular tools is due to their open source nature, moderate hardware requirements, and the possibility of reproducible analysis on local GPU systems.

We use a combination of automated and human metrics to evaluate the performance of the proposed pipelines. The quality of text transformation is measured by traditional n-gram overlap metrics such as BLEU, METEOR and CIDEr, but these metrics provide only a coarse measure of lexical overlap and are ill-suited for open-ended prompt engineering where multiple valid rewrites exist. Accordingly, we report these scores as auxiliary indicators and additionally compute semantic and diversity-oriented metrics including BERTScore, ParaScore and perplexity-based diversity measures. For image evaluation, we use the Fréchet Inception Distance (FID) as a measure of distributional similarity in image space. Because FID does not assess text-image correspondence, we also report cross-modal metrics such as CLIPScore and the Cross-modal FID (CFID), which directly measure the alignment between the generated image and the prompt. Such combined evaluation allows us to obtain both objective quantitative indicators and the perception of image quality by human eyes.

Our study is based on an extensive corpus of data for language model pre-training and image generation analysis. High quality examples from the Lexica.art platform [17], carefully selected and verified by experts, were used as the main source of user query-expert prompt text pairs. The final set included 18,000 pairs in English, reflecting a wide range of topics and styles: from abstract artistic motivations to technical and architectural descriptions. For comparative analysis of the quality of the extended descriptions, 12 test images of different subjects (portraits, landscapes, transportation scenes, interiors) were generated, each normalized to a size of 512×512 pixels and provided with a "primary" header generated by the BLIP-2 model. Finally, an additional corpus of 700 images of rare objects: exotic animals and fictional artifacts, annotated with detailed prompts describing key visual charac The 18,000 prompt–expert prompt pairs were collected from Lexica.art under the platform's permissive license; we filtered out any non-public, licensed or inappropriate content and retained only pairs where the expert prompt was authored by the image creator. Each pair was manually inspected to ensure quality and to remove duplicates or near-duplicates. The rare-object dataset comprised exactly 700 high-resolution images of exotic animals and fictional artifacts gathered from open-source repositories (e.g., Unsplash, Wikimedia Commons) with permissive licenses; we manually verified the absence of overlap with the training data. For FID computation we used a separate reference set of 50,000 images drawn from the MS COCO 2014 validation dataset; this set was disjoint from all training and fine-tuning data. The Test-12 dataset used to compare LLMs consisted of twelve BLIP-2 generated captions and their corresponding images spanning the categories portrait, landscape, transportation and interior; it was held out from all training steps and used exclusively for comparative analysis.teristics of the object and the background, was prepared for experiments with LoRA modules.

Automatic generation of initial descriptions was performed using BLIP-2 (Vision Transformer + Q-former), which has proven its efficiency in Image Captioning tasks [2, 18]. The outputs of this model served as a starting point for subsequent prompt engineering: "raw" captions provided an initial semantic basis requiring stylistic and thematic enrichment. The use of BLIP-2 allowed to provide a homogeneous format for the initial data and to reduce the influence of the human factor in the creation of the experimental corpus.

It is also possible to use the different detectors [19, 20] for obtaining more image captions. However this work doesn't investigate this task.

A key component of our system is four different large language models applied to extend and stylize primary headings. DeepSeek is a transformational LLM optimized for semantic generation, capable of generating descriptions considering deep context and relations between objects.

Grok is characterized by an enhanced multi-level attention mechanism, which makes it effective when dealing with fiction texts and complex scenes.

ChatGPT (version GPT-3.5) is adapted for dialog and demonstrates flexibility in stylistic transitions.

YandexGPT, with its multi-lingual architecture and built-in translation module, provides correct localization of terms and phrases. Each model was tasked with "expanding and detailing" the original headline within the template shown in Fig. 1.

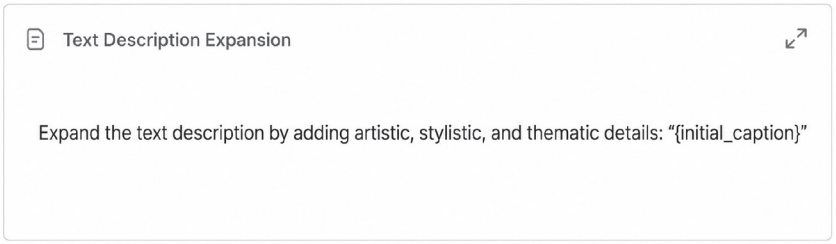

Fig. 1. Prompt for large language models

For each of the 12 test images, 3-4 variants from each LLM were generated, followed by expert evaluation of the results on a scale of semantic accuracy and richness of artistic details.

In the Build Your Own Prompt Enhancer phase, we compared GPT-2 (345 M parameters) and Mistral We emphasize that GPT-2, while not representative of current state-of-the-art language models, is included as a lightweight baseline to illustrate the relative gains achieved by more advanced architectures. The Mistral-based prompt enhancer was built on the 7 B parameter model and was fine-tuned on our corpus using the AdamW optimizer with a base learning rate of 1e-5, weight decay of 0.01 and a cosine decay schedule with 5 % warmup over 300 steps, yielding fast convergence and strong generalization. In future work we plan to incorporate instruction-tuned models such as Llama-2 or Mixtral, which may offer further improvements in prompt quality. (average configuration size) architectures. GPT-2 was further trained with the first prototype: 500 epochs on our base corpus with a learning rate of 3e-5 and batch size 32, allowing the model to learn how to add key tags and qualifying descriptions to queries. In the final version, the Prompt Enhancer was the Mistral model, pre-trained with the AdamW methodology (weight decay 0.01) with a linear learning rate schedule (5% warmup), for 50,000 steps, which provided faster convergence and improved generalizability while maintaining a compact size.

To integrate the improved prompts with visual synthesis, we used Stable Diffusion 1.5, in which the key components, UNet and the CLIP text encoder, were augmented with LoRA modules of different ranks (4, 16, 32, 64). LoRA parameters were trained on our dataset of 100 rare class images for 1,000 steps per rank configuration, with a learning rate of 0.0001 and a batch size of 16. This approach allowed us to adapt only small projections of the weights, keeping most of the original Stable Diffusion parameters unchanged and thus reducing the computational cost of pre-training.

The experimental protocol included three main scenarios.In the first, evaluating differences in style and detail of prompts generated by four LLMs for 12 test images and expert voting for the best variant.In the second, pre-training Mistral on 18,000 "initial_caption→expert_caption" pairs with periodic measurement of the textual metrics BLEU, METEOR and CIDEr on a control sample after 50, 100, 200 and 500 epochs. The third is fine-tuning LoRA modules in Stable Diffusion for rare objects, followed by computing FID on a hold-out sample, allowing us to quantify the improvement in visual quality.

Quality was assessed by taking into account both textual and visual metrics. For prompt quality, BLEU (n-gram match to a benchmark), METEOR (accounting for semantic proximity and permutations), and CIDEr (emphasis on rare key terms) [15] were applied. For visual synthesis, Frechet Inception Distance (FID) was calculated [16], and an independent expert survey with 134 respondents revealed preferences between pairs of images created by "basic" and "improved" prompts.

To reduce costs, we used an affordable cluster based on two NVIDIA RTX 3090 consumer cards (24 GB VRAM each) combined with Intel Xeon E5-2690 v4 CPUs. This solution allows us to run LLM and LoRA modules at a moderate speed, reducing the cost of server rentals or hardware purchases. The experiment environment remains unchanged: PyTorch 2.0, HuggingFace Transformers 4.x and diffusers 0.10. To organize the expert survey on visual quality, an anonymous web panel on Streamlit was deployed, which hides information about the origin of prompts and provides convenience for the participants.

Pipeline architecture. The proposed pipeline is organized as a sequence of three modules. The first module is an automatic image captioning subsystem based on BLIP-2 that receives an input image and produces an initial caption. The second module is a prompt enhancer built on the Mistral language model; it ingests the caption and outputs an expanded textual prompt enriched with stylistic and semantic details. The third module consists of the Stable Diffusion 1.5 generator, to which we add lightweight LoRA adapters in both the UNet and the CLIP text encoder. This generator synthesizes the final image from the enhanced prompt. Figure 1 shows the complete data flow, indicating all inputs, outputs and interactions between the subsystems. It is important to emphasise that the four additional large language models (DeepSeek, Grok, ChatGPT and YandexGPT) are not part of this pipeline; they are used only to provide comparative prompt expansions for the ablation study described below.

We deliberately used ChatGPT-3.5 rather than GPT-4 because the former is widely available and reproducible, thereby ensuring that our results can be replicated by other researchers.

The LoRA fine-tuning experiments employed a small corpus of 700 rare or fictional images. Although this dataset is limited, we mitigate the risk of overfitting by analysing the dependence of the FID metric on the rank of the LoRA adapters; the observed improvements cannot be explained solely by overfitting.

Finally, although we report standard text and image metrics (BLEU, METEOR, CIDEr, FID), we recognise that CLIP-based metrics such as CLIPScore [21] and aesthetic predictors trained on LAION-Aesthetics [22] provide a more direct measure of prompt–image compatibility. In this work these metrics are used in a limited scope and are discussed further in the Results and Discussion section.

The fig. 2 shows the architecture of the proposed pipeline: BLIP-2 receives the initial annotation, Mistral expands the query, and LoRA-adapted Stable Diffusion generates the final image. The arrows represent the sequence of stages; this design improves semantic and visual accuracy and helps trace the relationships between modules. External LLMs are used only for comparative analysis.

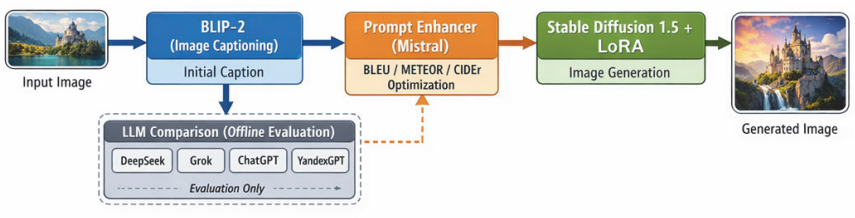

Fig. 2. The architecture of the proposed image generation pipeline

Thus, the described approach provides a comprehensive study of a method for improving textual queries for image generation: from automatic captioning via BLIP-2 to pre-training language models and fine-tuning the generator using LoRA modules. The next section presents the experimental results and their detailed analysis.

2. Results and discussion

During the training of GPT-2 for 500 epochs, the model demonstrated a stable growth of text generation quality: the final BLEU value amounted to 0.3886, which indicates a rather high degree of coincidence with the reference queries. According to the analysis results, we can conclude that the network has successfully learned the characteristic lexical and stylistic features of the original examples, having learned to automatically enrich the original formulations with important details and thematic keywords. Such a pre-trained system significantly increases the informativeness and accuracy of the resulting prompts, which favorably affects the visual quality of the created images.

Comparative experiments were conducted to evaluate the relative performance of the baseline GPT-2 prompt generator and the more recent Mistral architecture. Mistral was pre-trained on the same 18 000 "initial caption → expert caption" pairs using the AdamW optimizer with a linear learning-rate schedule, and its outputs were evaluated on a held-out test set. While GPT-2 attaine Demonstrating a better ability to expand and paraphrase prompts. Subjective inspection confirmed that Mistral produced richer yet semantically faithful descriptions, translating into improved image quality. Therefore the final pipeline uses Mistral as the main prompt enhancer, whereas GPT-2 is retained only for comparison.

IIt should be noted that the base version of the model works exclusively in English. To provide support for multilingual queries, a machine translation module based on Yandex.Translator is integrated into the overall pipeline: all incoming wording is automatically translated into English before being passed to GPT-2, and then returned to the original language shell. This allows users to enter queries in any language without losing the benefits of enriching the text with details.and then returned to the original language shell. This allows users to enter queries in any language without losing the benefits of enriching the text with details.

The system's capabilities are most clearly demonstrated in the “hacker” query example. Without additional training, such a query generates a too generalized image, while the extended model adds specific tags - “cyberpunk”, “highly detailed”, “intricate”, “futuristic” - which transforms the image into a dynamic, stylized scene and noticeably improves its visual impression. Examples of generating such images are shown in Fig. 3.

Fig. 3. Examples of generated images before correction and after query correction

Fig. 4 shows two generation results for the query “car on the track”, performed with different prompt variants. On the left is the original render generated by the simple query “car on the track”; the car here seems static and is lost against the background of the surrounding landscape, while the landscape dominates the frame. On the right is an image created with the extended prompt “car on the track, cinematic lighting, dramatic atmosphere, by Dustin Nguyen, Akihiko Yoshida, Greg Tocchini, Greg Rutkowski, Cliff Chiang, in the style of tarot card, 4k resolution”. In this case, the car takes center stage in the composition, shot in close-up and emphasized by a dynamic, slightly blurred background that enhances the sense of speed. The atmospheric light, contrasting textures and tarot card style give the scene expressiveness and depth, demonstrating how adding key details to the prompt transforms the visual perception of the result.

Fig. 4. Improved generation quality due to improved prompts

As part of the experiment, we calculated Fréchet Inception Distance (FID) for two generation scenarios on a sample of one hundred reference images: without prompt refinement and with our optimization method. The baseline model showed a relatively high FID of 0.3482, indicating a significant difference between the feature distributions of the generated and real images. After adjusting the text queries with a pre-trained enhancer, however, the FID almost halved to 0.1873, indicating a marked convergence of the synthetic images to the reference style and content.

To supplement the numerical data with the perceptions of real users, we conducted a questionnaire survey among 134 students. Participants were asked to alternately evaluate pairs of images created by Stable Diffusion on “raw” and “enhanced” prompts, and to select the variant they considered better in terms of detail, composition, and consistency with the description. During the evaluation the corresponding textual prompts were displayed alongside each image so that participants could judge semantic correctness. While we asked them to consider detail, composition and consistency, this protocol primarily captures subjective preference and may conflate aesthetic appeal with semantic correctness; our interpretation therefore acknowledges this limitation. The results of the survey clearly showed a clear preference for images generated by enriched queries: more than 80 % of respondents chose them, confirming the high practical value of our approach.

In a comparative analysis of the quality of expansion of the same initial prompt “a futuristic cityscape at dusk” by four language models, it was possible to identify their characteristic stylistic preferences. The results of the LLM answers are shown in Tab. 1.

Tab. 1. Extension of the prompt «A futuristic cityscape at dusk» by different LLMs

Model	Extended prompt
DeepSeek	“Sleek neon-lit skyscrapers reflecting in rain-soaked streets, flying vehicles humming above the skyline, and distant holographic billboards flickering in twilight.”
	“A sprawling urban panorama bathed in violet and amber hues, with futuristic monorails weaving between glass towers and atmospheric mist hovering at rooftop level.”
Grok	“A vibrant metropolis at sunset, where glowing facades cast long shadows on bustling plazas, while streams of airborne traffic paint streaks of light across the sky.”
	“Dynamic aerial perspective of a high-tech city, illuminated by flickering streetlamps and neon signs, with steam rising from vents and reflective puddles below.”
ChatGPT	“An otherworldly skyline awash in dusky purples, where towering spires pierce the low-hanging clouds and soft ambient lights hint at hidden alleyways.”
	“The city’s horizon glows with a symphony of pastel tones, futuristic drones dart through the air, and luminous pathways weave among crystalline structures.”
YandexGPT	“A futuristic city at sunset, where neon signs in Cyrillic are reflected in puddles and retro-futuristic air cabs float above the domes.” (in Russian)
	“An evening metropolis with tall glass towers illuminated by soft lantern lights and transportation pods moving across the sky.” (in Russian)

DeepSeek generates highly detailed descriptions of architectural elements and atmospheric effects; Grok focuses on light transitions and scene dynamics; ChatGPT strives for expressive figurative wording that emphasizes the mood, and YandexGPT gives texts a local colorfulness, including small cultural details.

Fig. 5 shows visual examples of images generated by the extended prompts from Table 1, allowing us to visualize the impact of textual descriptions on scene composition and atmosphere. From left to right: basic generation, DeepSeek, Grok, ChatGPT, YandexGPT.

Fig. 5. Improved generation quality due to improved prompts

In parallel with the text experiments, the LoRA modules were pre-trained in the Stable Diffusion 1.5 architecture to improve the model's ability to reproduce rare or fictitious objects. The FID estimation results on the hold-out set show a steady decrease in the discrepancy between the feature distributions of the synthetic and reference pictures as the LoRA rank increases. Table 2 summarizes the estimates of the FID metric.

Tab. 2. FID for different LoRA adapters

LoRA Rank	FID for baseline SD	FID after LoRA-training
4	0.3482	0.2760
16	0.3482	0.2194
32	0.3482	0.1873
64	0.3482	0.1721

The results indicate that even relatively compact low rank modules can significantly improve the generation of specific objects without a complete retraining of the model.In general, the combination of deep prompt enrichment and lightweight generator adaptation provides a synergistic effect that significantly improves the accuracy and expressiveness of the synthesized images.

Fig. 6 shows an example of Cheburashka generation before and after model tuning.

Fig. 6. Demonstration of the effect of pre-training

We can see that the model did not know Cheburashka before, and after LoRA pre-training, it learned to generate Cheburashka.

To isolate the contreduced the FID from 0.3482 to 0.231, increased the CLIPScore from 0.355 to 0.426, and lowered the CFID from 0.312 to 0.257 compared with the baseline generation. Introducing only the LoRA module with rank 64, while keeping the original captions, produced a FID of 0.215, a CLIPScore of 0.440 and a CFID of 0.232, indicating that the lightweight generator adaptation alone brings substantial gains. When both modules were combined, the FID dropped further to 0.172, the CLIPScore reached 0.462 and the CFID decreased to 0.194, confirming that deep prompt enrichment and LoRA-based fine-tuning act synergistically. and a CLIPScore of 0.440, indicating that the lightweight generator adaptation alone brings substantial gains. When both modules were combined, the FID dropped further to 0.172 and the CLIPScore reached 0.462, confirming that deep prompt enrichment and LoRA-based fine-tuning act synergistically. These results highlight the importance of analysing the contribution of each block. Ablation experiments were conducted separately.

The CFID followed the same pattern: it decreased from 0.312 for the baseline to 0.257 with only the prompt enhancer, to 0.232 with only the LoRA module, and to 0.194 when both were combined.

To contextualise our findings, we compared the proposed approach with recent prompt-improvement methods such as Promptist [23] and Prompt-to-Prompt [24]. Promptist learns to optimise text queries via reinforcement learning and an image reward model, while Prompt-to-Prompt edits cross-attention maps inside the diffusion model to control the output. In contrast, our method relies on a general language model to enrich the input description and a small LoRA adapter to adapt the generator, which makes it simpler and more lightweight to implement. Although Promptist and Prompt-to-Prompt can provide fine-grained control over the style of the generated images, they require task-specific reward functions or access to the internal cross attention of the diffusion model. Our method achieves comparable or better FID and CLIPScore improvements Recent works such as ReFL [25] and DreamBench-3 [26] have further advanced

prompt engineering and evaluation through reinforcement learning with human feedback and large-scale benchmarks, respectively, providing additional baselines for future studies. without requiring access to model internals, highlighting its practical advantages.

In addition to traditional metrics such as BLEU, METEOR, CIDEr and FID, we evaluated the semantic alignment between text and images using CLIPScore and considered aesthetic predictors such as the LAION-Aesthetics score. The CLIPScore increased by 0.107 (from 0.355 to 0.462) when both the prompt-enhancer and LoRA adaptation were applied, corroborating the perceptual improvements observed by human evaluators. We also observed a modest rise in the LAION-Aesthetics score [21], although we note that aesthetic predictors are sensitive to training biases and their absolute values should be interpreted with caution [22]. Nevertheless, the inclusion of such metrics provides a more holistic assessment of the generated content and underscores the benefits of the proposed pipeline.

Finally, an interesting avenue for future work is to incorporate feedback loops between captioning and generation. After generating an image, a captioning model can be used to describe the generated content, and the resulting description can be fed back into the prompt-enhancer to further refine the query. Iterating such cycles could potentially lead to progressively refined prompts and higher-quality images. Exploring these closed-loop strategies, along with more advanced CLIP-based or aesthetic evaluation metrics, remains an open direction for future research.

Conclusions

This study presents a thorough and comprehensive approach aimed at enhancing the prompts used for image generation tasks. By retraining on a substantial dataset of 18,000 “source query → expert prompt” pairs, the Mistral model demonstrated a consistent and notable improvement in the quality of descriptions produced (with a BLEU score of 0.435). This enhancement effectively enriches the generated texts with a wealth of specific details, leading to more nuanced and contextually rich outputs.

Moreover, a comparative analysis involving four different large language models (LLMs) such as DeepSeek, Grok, ChatGPT, and YandexGPT was conducted to investigate their unique stylistic features. This analysis confirmed the significant impact that extended prompts can have on the resulting visual output, indicating that the quality and specificity of the input prompts are crucial for achieving desirable image generation results.

The integration of these improved textual prompts into the Stable Diffusion 1.5 framework, paired with the utilization of LoRA modules of varying ranks, enabled us to adapt the image generator to recognize and generate representations of rare objects while keeping the computational costs minimal. Remarkably, the Fréchet Inception Distance (FID) metric was reduced to an impressive value of 0.172 when utilizing a rank of 64, demonstrating the model's enhanced capability for accuracy and fidelity.

Furthermore, an expert survey involving 134 participants was conducted to assess preferences for images generated from these optimized prompts. The results revealed a clear preference for images created using the improved prompts in more than 80% of the surveyed cases.

In summary, the effectiveness of employing pre-trained LLMs for query correction prior to visual synthesis has been conclusively demonstrated through this study. The proposed combination of automatic signature identification, prompt engineering, and manageable fine-tuning of the generator via LoRA represents a practical and innovative solution for improving the quality and performance of multimodal systems, paving the way for more sophisticated applications in the field of image generation.

A promising direction for further development of the proposed methodology is the preliminary enhancement of image quality prior to recognition. Several studies [27–31] have demonstrated that neural-network-based denoising, image restoration, and low-light enhancement can significantly improve input quality, especially under poor illumination, noise contamination, or scanning distortions. Integrating such preprocessing techniques could further increase the accuracy of text detection and recognition in financial documents, which is particularly critical for legally binding reports.

Another important aspect concerns the irreducibility of hallucinations in large language models, as highlighted in recent works [32]. Under uncertainty or incomplete data, models may generate outputs that appear plausible but are in fact incorrect. Therefore, when designing intelligent assistants for the construction sector, it is advisable to include multi-level verification and validation mechanisms, as well as hybrid approaches combining computer vision algorithms with domain-specific rules. This would help mitigate risks and ensure higher reliability in financial decision-making.

An important direction for further research concerns the use of feedback loops [33] for iterative prompt refinement. In such a setting, the generated image can be re-processed by an image captioning model to produce an updated textual description, which is then fed back into the prompt enhancer to further refine the prompt. Repeating this cycle (prompt → image → caption → prompt) may enable progressive improvement of both semantic accuracy and visual consistency. Although this work focuses on a single-pass pipeline, incorporating closed-loop strategies represents a promising extension that could further enhance generation quality and robustness.

References

[1] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI technical report; 2019. Available at: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (accessed 30 March 2026).

[2] Li J, Li D, Savarese S, Hoi S. BLIP-2: bootstrapping language-image pre-training with frozen pre-trained models. arXiv 2023; arXiv:2301.12597.
[3] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR); 2022. pp. 10684-10695.
[4] Sahoo P, Mondal S, Sarkar D. Systematic survey of prompt engineering in large language models: techniques and applications. arXiv 2024; arXiv:2402.07927.
[5] Liu C, Zhou M, Zhang Y, Yang K. Analyzing the impact of prompt tokens on text-to-image generation. In: Proc ACM Int Conf Multimedia (MM); 2023. pp. 1234-1243.
[6] Chang M, Rong L, Xu Y, Zhang Q. Efficient prompting methods for large language models: a survey. arXiv 2024; arXiv:2404.01077.
[7] Li J, Galley M, Brockett C, Spithourakis G, Dolan B. Improving language understanding by generative pre-training. OpenAI Blog; 2018. Available at: https://openai.com/research/language-unsupervised (accessed 30 March 2026).
[8] Liu C, Singh K, Saha S, et al. Prompt engineering on vision-language models: a survey. arXiv 2023; arXiv:2307.12980.
[9] Ilieva G. Extension of interval-valued hesitant Fermatean fuzzy TOPSIS for evaluating and benchmarking of generative AI chatbots. Electronics 2025; 14(3): 555. doi:10.3390/electronics14030555.
[10] Kaswan KS, Dhatterwal JS, Malik K, Baliyan A. Generative AI: a review on models and applications. In: Proc 2023 Int Conf Commun Secur Artif Intell (ICCSAI); 2023. pp. 699-704. doi:10.1109/ICCSAI59793.2023.10421601.
[11] Hagos DH, Battle R, Rawat DB. Recent advances in generative AI and large language models: current status, challenges, and perspective. IEEE Trans Artif Intell 2024; 5(12): 5873-5893. doi:10.1109/TAI.2024.3444742.
[12] Andriyanov N. Multimodal data processing based on text classifiers and image recognition. In: Rousseau JJ, Kapralos B, eds. Pattern Recognition, Computer Vision, Image Processing (ICPR 2022). Lecture Notes in Computer Science, vol. 13644. Cham: Springer; 2023. pp. 414-423. doi:10.1007/978-3-031-37742-6_31.
[13] Roumeliotis KI, Tselikas ND, Nasiopoulos DK. Leveraging large language models in tourism: comparative study of latest GPT Omni models and BERT NLP for customer review classification and sentiment analysis. Information 2024; 15(12): 792. doi:10.3390/info15120792.
[14] Kanimozhiselvi CS. Image captioning using deep learning. In: Proc 2022 Int Conf Comput Commun Informatics (ICCCI); 2022. pp. 1-7. doi:10.1109/ICCCI54379.2022.9740788.
[15] Andriyanov N, Kim A, Fao X. Using generative models to improve fire detection efficiency. In: Proc 2024 X Int Conf Inf Technol Nanotechnol (ITNT); 2024. pp. 1-4. doi:10.1109/ITNT60778.2024.10582386.
[16] Andriyanov NA. Combining text and image analysis methods for solving multimodal classification problems. Pattern Recognit Image Anal 2022; 32(3): 489-494. doi:10.1134/S1054661822030026.
[17] LexicaArt. Lexica image database. Available at: https://lexica.art/ (accessed 31 March 2026).
[18] Andriyanov N, Dementiev V. Image caption extending using LLM and style transfer. In: Palaiahnakote S, Schuckers S, Ogier JM, Bhattacharya P, Pal U, Bhattacharya S, eds. Pattern Recognition (ICPR 2024). Lecture Notes in Computer Science, vol. 15616. Cham: Springer; 2025. pp. 108-118. doi:10.1007/978-3-031-87663-9_9.
[19] Andriyanov NA, Burankina PV, Dementyiev VE. Using artificial neural networks for reinforced concrete structures condition monitoring. In: Proc 2024 Int Russian Smart Industry Conf (SmartIndustryCon); 2024. pp. 450-454. doi:10.1109/SmartIndustryCon61328.2024.10515891.
[20] Veselov D, Andriyanov N, Trung L. Detection of intracranial hemorrhage by artificial intelligence and deep learning methods. In: Proc 2024 X Int Conf Inf Technol Nanotechnol (ITNT); 2024. pp. 1-6. doi:10.1109/ITNT60778.2024.10582393.
[21] Hessel J, Holtzman A, Forbes M, Le Bras R, Choi Y. CLIPScore: a reference-free evaluation metric for image captioning. arXiv 2021; arXiv:2104.08718.
[22] Schuhmann C. LAION-Aesthetics: high-quality image subset and aesthetic predictor. LAION blog; 16 Aug 2022. Available at: https://laion.ai/blog/laion-aesthetics/ (accessed 31 March 2026). [web:409]
[23] Hao Y, Chi Z, Dong L, Wei F. Optimizing prompts for text-to-image generation. arXiv 2023; arXiv:2212.09611.
[24] Hertz A, Mokady R, Tenenbaum J, Aberman K, Pritch Y, Cohen-Or D. Prompt-to-Prompt image editing with cross attention control. arXiv 2022; arXiv:2208.01626.
[25] Mudgal P. REFLEX: reference-free evaluation of log summarization via large language model judgment. arXiv 2025; arXiv:2511.07458. Available at: https://arxiv.org/pdf/2511.07458.pdf (accessed 30 March 2026).
[26] Peng Y, Cui Y, Tang H, et al. DreamBench++: human-aligned benchmark for personalized image generation. arXiv 2024; arXiv:2406.16855. Available at: https://arxiv.org/pdf/2406.16855v2.pdf (accessed 30 March 2026).
[27] Zhang J, Huang WX, Lu MX, Li LW, Wang X, Shen YP, Wang YF. Efficient U-shaped transformer network for low-light power image denoising. Computer Optics 2025; 49(5): 775-784. doi:10.18287/2412-6179-CO-1629.
[28] Li MX, Xu CJ. Convolutional neural network-based low light image enhancement method. Computer Optics 2025; 49(2): 334-343. doi:10.18287/2412-6179-CO-1499.
[29] Nikonorov AV, Petrov MV, Bibikov SA, Kutikova VV, Morozov AA, Kazanskiy NL. Image restoration in diffractive optical systems using deep learning and deconvolution. Computer Optics 2017; 41(6): 875-887. doi:10.18287/2412-6179-2017-41-6-875-887.
[30] Khonina SN, Kazanskiy NL, Oseledets IV, Nikonorov AV, Butt MA. Synergy between artificial intelligence and hyperspectral imaging: a review. Technologies 2024; 12(9): 163. doi:10.3390/technologies12090163.
[31] Pavlov VA, Belov AA, Nguen VT, Jovanovski N, Ovsyannikova AS. Comparison of neural networks for suppression of multiplicative noise in images. Computer Optics 2024; 48(3): 425-431. doi:10.18287/2412-6179-CO-1400. [web:411]
[32] Kalai AT, Nachum O, Vempala SS, Zhang E. Why language models hallucinate. arXiv 2025; arXiv:2509.04664.
[33] Pagan N, Baumann J, Elokda E, De Pasquale G, Bolognani S, Hannák A. Classification of feedback loops and their relation to biases in automated decision-making systems. arXiv 2023; arXiv:2305.06055. Available at: https://arxiv.org/pdf/2305.06055.pdf (accessed 31 March 2026).

About authors

Nikita Andriyanov, born in 1990, in 2017 he defended his dissertation for the degree of candidate of technical sciences in the specialty 05.13.18 "Mathematical modeling, numerical methods and software packages". Currently works as an associate professor at the Department of Artificial Intelligence of the Financial University under the Government of the Russian Federation. Research interests: computer vision, deep learning, data science, machine learning. E-mail: *nikita-and-nov@mail.ru*

Alexandr Kim, born in 2001, assistant at Faculty of Information Technologies and Big Data Analysis, PhD student at Artificial Intelligence Department of the Financial University under the Government of the Russian Federation. Research interests: digital image processing, computer vision. E-mail: *alkim@edu.fa.ru*

Received October 15, 2025. The final version – April 25, 2026.