CaRVE: Critiquing and Refining Visual Elaborations for Figurative Language Illustrations

Manishit Kundu
CMInDS, IIT Bombay
manishit@minds.iitb.ac.in

Tejomay Kishor Padole
CSE, IIT Bombay
tejomaypadole@cse.iitb.ac.in

Sumit Shekhar
Adobe Systems, India
sushekha@adobe.com

Biplab Banerjee
CSRE, IIT Bombay
bbanerjee@iitb.ac.in

Pushpak Bhattacharyya
CSE, IIT Bombay
pb@cse.iitb.ac.in

Abstract

Illustrating figurative language remains challenging due to its non-literal semantics, and existing text-to-image frameworks rely heavily on proprietary models or human supervision to achieve adequate alignment. We introduce **CaRVE**, a lightweight and fully open-source critique-driven framework that employs VLM feedback to refine visual elaborations for figurative image generation. CaRVE bridges the semantic alignment gap even in sub-4B models by correcting visual and conceptual misalignments, reducing over-literalization, and improving robustness to complex figurative expressions. Using only open-source models, CaRVE achieves a **6.49%** improvement over prior baselines on intrinsic automatic evaluations and a **+0.37** average rank gain in human preference. We further release **MetaCaRVE**, an enhanced figurative image dataset constructed by refining HAIVMet using CaRVE[1].

1 Introduction

Illustrating figurative language poses unique challenges due to its non-literal and multi-layered semantics. Prior text-to-image frameworks for figurative illustration (Shahmohammadi et al., 2023; Chakrabarty et al., 2023; Zhang et al., 2024) address this by generating intermediate visual elaborations, typically via chain-of-thought prompting, to guide image synthesis. However, these elaborations frequently remain underspecified or overly literal, resulting in semantic drift and misaligned generations.

Prior works rely heavily on closed, proprietary models or human verification to minimise such errors, limiting scalability and accessibility. As a result, no lightweight and fully open solution currently exists, largely due to the significant gap in reasoning capability between proprietary large language models and smaller, open-source LLMs. We posit that this semantic gap can be effectively bridged in smaller models through a VLM-critique-driven feedback mechanism.

In this work, we introduce **Critiquing-and-Refining Visual Elaborations (CaRVE)**, a VLM-feedback-driven framework for figurative text-to-image generation. CaRVE employs VLM-based critique to iteratively refine visual elaborations for image generation, yielding significantly improved semantic alignment even with smaller sub-4B models. Our contributions are:

1. **CaRVE:** a novel, lightweight, open-source, critique-driven framework for illustrating figurative language that outperforms existing figurative text-to-image generation pipelines, achieving a 6.49% improvement on intrinsic automatic evaluations and a +0.37 rank gain in human preference using only open-source sub-4B models. This demonstrates its effectiveness in aligning generated images with the intended meaning of figurative text.

2. **MetaCaRVE:** An enhanced multimodal figurative dataset constructed by refining the HAIVMet dataset (Chakrabarty et al., 2023) using our proposed framework. The dataset comprises 1100 instances, each consisting of a linguistic metaphor paired with a synthetically generated image.

2 Methodology

Following prior work, we adopt a two-stage figurative text-to-image pipeline consisting of (i) generating a literal visual elaboration from figurative input and (ii) rendering the elaboration into an image using an off-the-shelf text-to-image (T2I) model. We extend this pipeline with a third stage, **Critiquing-and-Refining Visual Elaborations (CaRVE)**, which introduces image-conditioned

[1] https://github.com/manishitkundu/CaRVE_ACL_2026

Findings of the Association for Computational Linguistics: ACL 2026, pages 42760–42777

Figure 1: An example illustrating the role of CaRVE's visual feedback in enhancing semantic alignment.

feedback to improve the visual elaboration and consequently, image-text semantic alignment.

Stage I–II: Vanilla Baseline. Given a figurative input, an LLM first produces a grounded literal interpretation and a corresponding visual elaboration, which is then rendered into an image by a pretrained T2I model. Prompt details are provided in Appendix J.

Stage III: Critiquing and Refining (CaRVE). CaRVE employs a VLM as a semantic critic that evaluates the generated image in the context of the figurative input and its interpretation. The VLM identifies conceptual, visual, and communicative misalignments and produces targeted refinement suggestions.

The critique is structured around three key semiotic dimensions of visual metaphor expression (Bolognesi et al., 2018): **Conceptualisation**, **Expression**, and **Communication**. It evaluates whether the primary and secondary concepts and their associated attributes are correctly identified (*Conceptualisation*), how abstract and concrete elements are visually grounded in the generated image (*Expression*), and whether metaphorical elements are meaningfully integrated without over-literalization in a manner consistent with the grounded interpretation (*Communication*). Based on these analyses, the VLM produces a set of targeted suggestions, which are then passed to the LLM to generate a refined elaboration. This revised elaboration is subsequently fed to the T2I model to produce the final image.

3 Results and Analysis

We generate images from HAIVMet (Chakrabarty et al., 2023), GOME (Zhang et al., 2024), and CaRVE visual elaborations using two text-to-image models: PixArt-α (Chen et al., 2024) and Lumina (Qin et al., 2025). We additionally evaluate the same two-stage backbone without the critique pass, referred to as *Vanilla*. For context, we also include the original generations from HAIVMet and GOME using their respective models. All experiments use figurative texts from the HAIVMet-Data (also used in GOME-Data) for standardised comparison.

T2I Model	VE Model	TT ↑	TI ↑	Overall ↑
DALL-E	HAIVMet	0.722	0.386	0.554
SD 1.4	GOME	0.710	0.400	0.555
Pixart-α	Vanilla	0.681	0.384	0.533
	HAIVMet	0.722	0.384	0.553
	GOME	0.710	0.384	0.547
	CaRVE	**0.785**	**0.385**	**0.585**
Lumina	Vanilla	0.681	0.386	0.534
	HAIVMet	0.722	0.386	0.555
	GOME	0.710	0.386	0.548
	CaRVE	**0.788**	**0.388**	**0.588**

Table 1: **Intrinsic evaluation of models using our composite metric.** TT denotes the Text-to-Text component, TI denotes the Text-to-Image component, and Overall is their mean. All scores range from 0 to 1. Yellow highlights baseline methods as reported in their respective papers, while Green indicates our proposed methods. Vanilla indicates the two-stage pipeline with the same backbone without CaRVE. T2I indicates the Text-to-Image and VE indicates Visual Elaboration.

3.1 Intrinsic Automatic Evaluation

We introduce a composite metric for evaluating figurative text-to-image pipelines that separately scores the elaboration and image stages and aggregates them into a final score in the range [0,1].

(1) Figurative Text to Visual Elaboration (TT). TT measures elaboration quality as the mean of

three sub-metrics: **Flesch Readability** (Flesch, 1948), **Literality**, and **Faithfulness**, each normalised to [0,1]. Literality and Faithfulness are scored by an LLM (Qwen2.5-7B) acting as a semantic evaluator, while Flesch Readability is computed automatically. Additional evaluations using two other LLMs are reported in Appendix G.

(2) Visual Elaboration to Image (TI). TI evaluates image–elaboration alignment using **DA-Score** (Singh and Zheng, 2023), which decomposes elaborations into assertions and computes their BLIP similarity to the image. The final TI score is the mean similarity, normalised to [0,1].

(3) Overall. The final score is the mean of TT and TI, providing an end-to-end measure of image–text semantic alignment.

CaRVE substantially improves elaboration quality while relying solely on lightweight, fully open models. CaRVE improves the TT metric by **15.27%** over the Vanilla pipeline and by an average of **9.85%** over corresponding prior baselines, directly supporting our claim that critique-based refinement bridges the semantic gap in figurative text-to-image generation. Notably, while HAIVMet employs DALL·E 2 (Ramesh et al., 2021) (3.5B parameters) together with GPT-generated and human-verified elaborations, and GOME relies on GPT-4 (OpenAI, 2023) with a scenario visualisation mechanism, CaRVE operates entirely with small, fully open-source models (a 4B LLM, a 4B VLM, and a 0.6B T2I model), eliminating reliance on proprietary systems and human supervision while delivering competitive, and often superior, performance. This establishes CaRVE as a lightweight, reproducible, and scalable alternative for high-quality figurative image generation in low-resource settings. For further fine-grained analysis of the TT metric and illustrative examples of intermediate visual elaborations, see Section D.

Variations due to T2I models are limited. CaRVE achieves an average TI improvement of **0.36%** over the baselines, indicating minimal variation. This marginal gain is expected, as the underlying T2I backbone remains unchanged, the visual elaborations are already literal scene descriptions, and off-the-shelf T2I models are of sufficiently high quality. Any residual variation is likely attributable to linguistic differences and the degree of specification in the elaborations. We also note that the original GOME framework employs an attention-based alignment mechanism during image generation, contributing to its comparatively higher TI score in the reported configuration.

Small TI gains do not imply small semantic changes. CaRVE introduces subtle but meaningful modifications, such as adding, removing, or refining elements, that significantly affect the generated image while producing only minor numerical changes in the TI metric. This behavior arises because DA-Score averages alignment across multiple assertions, diluting the impact of any single change; as the number of assertions increases, the contribution of individual improvements becomes less pronounced. Consequently, even clear visual enhancements often yield only marginal gains in TI, especially when compared to improvements in TT.

CaRVE introduces significant change

Input: The ripples wimple on the rills, like sparkling little lasses.

Vanilla: A narrow forest stream flows through a clearing. Small glowing girls appear to dance across the rippling water. Their shimmering dresses catch the light as they skip and play, as if the stream itself has transformed into tiny, sparkling lasses.

CaRVE: A small stream flows slowly through a forest clearing. Water moves in gentle ripples, reflecting the soft light of early morning. The ripples are clear and bright.

The above example demonstrates a change induced by CaRVE. In the Vanilla case, figurative elements (e.g., "small glowing girls") are explicitly realized in the text and consequently, in the generated image, whereas CaRVE removes such abstractions in both text and image, producing a more semantically aligned depiction. Despite this clear semantic shift, both (elaboration, image) pairs remain internally consistent: the Vanilla output includes the figurative elements in both modalities, while the CaRVE output consistently omits them. As a result, although CaRVE is semantically closer to the intended metaphor, the elaboration–image alignment remains similar across both cases, resulting in comparable DA-Scores. This indicates that TI

Setting	Mean Rank ↓
Zero-shot	3.26
HAIVMet	2.75
GOME	2.18
CaRVE	**1.81**

Table 2: **Human ranking of outputs generated by PixArt-α in the zeroshot setting and with elaborations from HAIVMet, GOME, and CaRVE.** Results indicate that CaRVE outputs are most preferred by humans, achieving a +0.37 average rank improvement.

primarily reflects the T2I model's ability to align text and image, rather than true semantic grounding. This is further supported by Table 1, where larger T2I models achieve higher absolute scores, while the relative improvement of CaRVE over Vanilla remains consistent across model sizes. Thus, TI is best understood as a sanity check: it ensures that text–image alignment is preserved and that no degradation in visual coherence occurs due to critiques.

3.2 Human Evaluation

We conduct human evaluations to assess the quality of generated images and visual elaborations, focusing on semantic alignment with the input metaphor and overall perceptual quality. Across multiple comparative settings, annotators consistently prefer CaRVE over prior baselines, demonstrating the effectiveness of critique-based refinement beyond what is captured by automatic metrics.

3.2.1 Image Comparison with Prior Baselines

We evaluate four settings: Pixart-α in a zero-shot setting, and, along with elaborations generated using HAIVMet, GOME, and CaRVE. We randomly sampled 100 figurative sentences and split them into five equal sets of 20 instances. Each instance comprised a figurative sentence and its corresponding image outputs from all four models. Five human evaluators were recruited, and each was assigned two disjoint sets, ensuring that every set was evaluated independently by two different annotators. Annotation details in Appendix I.

For each instance, annotators were asked to rank the four images based on how effectively they conveyed the intended meaning, with ties allowed. On average, CaRVE received a higher rank than all baselines by **0.37**, reflecting consistent annotator preference. Furthermore, our model was ranked highest (including joint-firsts) in **64.43%** of all comparisons.

Given the subjective nature of metaphor interpretation, we measured inter-annotator agreement using Spearman's ρ and observed a mean correlation of 0.59, indicating moderate-to-strong consistency among human preferences and reinforcing the reliability of our human annotations. The correlation between automatic scores and human rankings was 0.48, suggesting a moderate alignment. Importantly, this corresponds to 81.4% of the human–human agreement level, which is reasonable given the inherent subjectivity of the task.

3.2.2 Visual Elaboration Comparison with Prior Baselines

We conduct a 5-point Likert-scale evaluation of visual elaborations generated by Vanilla, HAIVMet, GOME, and CaRVE based on faithfulness to the input metaphor and quality. Human annotators consistently prefer CaRVE, which attains the highest average score of **4.27** compared to **4.16** for GOME, which was a close second. This validates our automated critique-based framework and alleviates concerns regarding the absence of human supervision.

3.2.3 Effect of Critique-based Refinement

We further evaluate CaRVE against the Vanilla pipeline using the same LLM and T2I backbone. Human annotators are asked to select the better image between the two methods, with ties permitted. We observe a clear preference for CaRVE-generated images, indicating that critique-based refinement substantially improves semantic alignment. Ties account for only 35% of the instances, suggesting that in the majority of cases the critique pass is necessary to achieve optimal visual representations. See examples in Appendices D and E.

3.3 CaRVE as an Alternative to Model Scaling

We analyze the effect of model scale on the relative benefit of the critique-based feedback loop by varying the size of the underlying language model in the vanilla two-stage pipeline and evaluating performance with and without the critique pass.

As shown in Table 3, increasing model size improves the performance of the Vanilla pipeline while reducing the relative gains from critique-based refinement, as larger models already produce stronger initial elaborations with less room for improvement. Nevertheless, CaRVE consis-

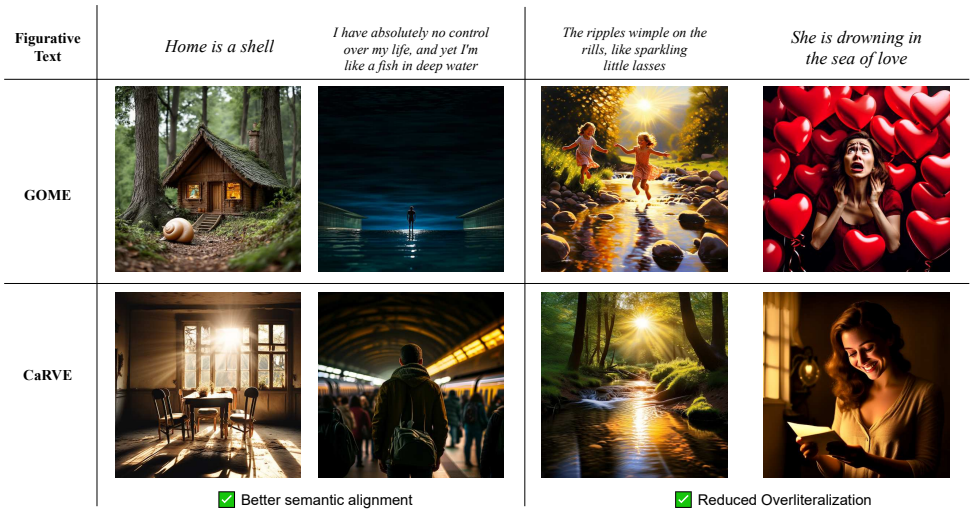

Figure 2: **Examples illustrating improved semantic alignment and reduced overliteralization in CaRVE outputs.** CaRVE goes beyond surface-level interpretation, capturing the underlying meaning with careful attention to conceptual representation.

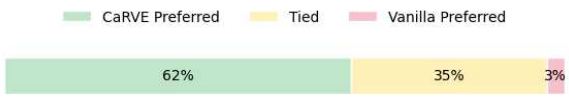

Figure 3: **Humans prefer CaRVE outputs over the Vanilla two-stage pipeline outputs.**

Model Size	Vanilla Score	CaRVE Score
4B	0.532	**0.585 (+9.75%)**
8B	0.548	0.591 (+7.85%)
14B	0.561	0.596 (+6.23%)

Table 3: Effect of model size on Vanilla and CaRVE performance. Percentages indicate relative gains over corresponding Vanilla configurations.

tently enhances performance across all scales, with the largest gains observed in low-resource settings. Notably, CaRVE with a 4B backbone (0.585) outperforms the Vanilla pipeline even at 14B (0.561). These results demonstrate that critique-based refinement can be more effective than brute-force model scaling, positioning CaRVE as a compute-efficient alternative, particularly in lightweight and deployment-constrained settings. A detailed analysis of CaRVE's computational overhead is provided in Appendix B.

4 Dataset Enhancement via CaRVE

We enhance 1100 instances common to GOME-Data and HAIVMet-Data by regenerating their visual elaborations and images using CaRVE. For dataset construction, we employ Qwen3-4B as the LLM, Gemma3-4B as the VLM, and PixArt-α as the T2I model.

The resulting dataset, denoted as **MetaCaRVE**, contains refined visual elaborations and corresponding images that exhibit higher literality, improved semantic faithfulness, and reduced overliteralization (Figure 2) compared to the original releases, as evidenced by our evaluations in Tables 1 and 2. By providing systematically refined supervision, MetaCaRVE offers a stronger benchmark for training and evaluating figurative text-to-image systems. We additionally release the Vanilla elaborations prior to CaRVE refinement, which may serve as useful negative examples for preference-based optimization and contrastive training settings. Additional examples in Appendix E.

5 Summary and Conclusion

We introduced CaRVE, a lightweight and fully open-source, critique-driven framework for figurative text-to-image generation, enabling sub-4B models to achieve strong semantic alignment without reliance on proprietary systems or human supervision. Through comprehensive automatic and human evaluations, we demonstrate that critique-based refinement consistently improves visual elaboration quality and overall alignment, allowing smaller models to compete with and even surpass substantially larger baselines. We further release MetaCaRVE, an enhanced figurative image dataset constructed using CaRVE.

Limitations

While our approach strengthens figurative language elaboration, there are several avenues that fall outside the scope of the current work and merit exploration in future studies.

1. **Use of Off-the-Shelf Text-to-Image Models.** We use pretrained text-to-image (T2I) models without any task-specific fine-tuning or adaptation. While this allows for modularity and fair comparison, it also means we do not optimize the image generation component of the pipeline. Integrating feedback-driven tuning for the T2I stage could potentially yield even stronger results, especially for metaphors requiring fine-grained visual composition.

2. **Single-Pass Refinement.** CaRVE applies a single critique-and-refine pass. Although this improves efficiency, more complex or layered figurative expressions may benefit from iterative refinement. Implementing such loops would likely improve output quality, but at the cost of increased inference time and compute, raising scalability concerns in real-world applications.

3. **Language and Resource Limitations.** Our experiments are conducted exclusively in English, a high-resource language with relatively simple morphology. The effectiveness of CaRVE in languages with richer morphological structures, or in low-resource settings where metaphorical constructs vary significantly, remains unexplored. Extending the method to such languages may require additional linguistic adaptation and culturally grounded training data.

Ethics Statement

Our work builds upon large pretrained models, including text-to-image (T2I) models, language models (LLMs), and vision-language models (VLMs), all of which may carry inherent biases learned from their training data. These biases can manifest in the generated elaborations and images, particularly in the depiction of social roles, gender, race, or cultural symbols, potentially leading to stereotypical or inappropriate representations in figurative contexts.

While our primary focus is on improving the semantic alignment and figurative depth of visual outputs, we acknowledge the risk of bias propagation through the pipeline. We ensure that no harmful or derogatory generations are included in our final qualitative or human evaluation results. Furthermore, no personally identifiable or sensitive data is used or produced in our pipeline. All datasets used are publicly available or synthetically generated without referencing real individuals.

All models used in this study are publicly released and used under their respective licenses.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Marianna Bolognesi, Romy van den Heerik, and Esther Berg. 2018. *Chapter 4. VisMet 1.0: An online corpus of visual metaphors*, pages 89–114.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7370–7388, Toronto, Canada. Association for Computational Linguistics.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James T. Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ricardo Kleinlein, Cristina Luna Jiménez, and Fernando Fernández Martínez. 2022. Language does more than describe: On the lack of figurative speech in text-to-image models. *CoRR*, abs/2210.10578.

Satyapriya Krishna, Jiaqi Ma, Dylan Slack, Asma Ghandeharioun, Sameer Singh, and Himabindu Lakkaraju. 2023. Post hoc explanations of language models can

improve language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Evelina Leivada, Elliot Murphy, and Gary Marcus. 2022. DALL-E 2 fails to reliably capture common syntactic processes. *CoRR*, abs/2210.12889.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Xinyue Li, Dongyang Liu, Xiangyang Zhu, Will Beddow, Erwann Millon, Wenhai Wang Victor Perez, Yu Qiao, Bo Zhang, Xiaohong Liu, Hongsheng Li, Chang Xu, and Peng Gao. 2025. Lumina-image 2.0: A unified and efficient image generative framework. *Preprint*, arXiv:2503.21758.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752.

Hassan Shahmohammadi, Adhiraj Ghosh, and Hendrik Lensch. 2023. ViPE: Visualise pretty-much everything. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5477–5494, Singapore. Association for Computational Linguistics.

Jaskirat Singh and Liang Zheng. 2023. Divide, evaluate, and refine: Evaluating and improving text-to-image alignment with iterative vqa feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Gemma Team. 2024. Gemma.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Linhao Zhang, Jintao Liu, Li Jin, Hao Wang, Kaiwen Wei, and Guangluan Xu. 2024. GOME: Grounding-based metaphor binding with conceptual elaboration for figurative language illustration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18500–18510, Miami, Florida, USA. Association for Computational Linguistics.

Zhiqiang Zhong, Kuangyu Zhou, and Davide Mottin. 2024. Harnessing large language models as post-hoc correctors. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 14559–14574. Association for Computational Linguistics.

A Related Work

Text-to-Image Generation: The advent of *diffusion-based models* marked a significant leap in the fidelity and controllability of text-to-image (T2I) synthesis. DALL·E 2 (Ramesh et al., 2021) introduced high-resolution generation with compositional reasoning, while STABLE DIFFUSION (Rombach et al., 2021) popularized open-source diffusion architectures with strong prompt controllability and scalability. Despite these advances, diffusion models often suffer from semantic misalignment, over-literal interpretations, and failures to faithfully render abstract or figurative prompts (Leivada et al., 2022; Kleinlein et al., 2022).

Figurative and Metaphorical Text-to-Image Synthesis: Recent efforts have addressed the unique challenges of generating images from *figurative* or *metaphorical text*, which require capturing non-literal, often symbolic meanings.

HAIVMET (Chakrabarty et al., 2023) introduced a benchmark of 1,540 metaphorical expressions annotated with visual elaborations through human–AI collaboration. VIPE (Shahmohammadi et al., 2023) proposed a lyric-based elaboration pipeline that uses LLMs to generate abstract textual prompts, improving image expressiveness and symbolic coherence.

More recently, GOME (Zhang et al., 2024) introduced a grounding-aware chain-of-thought (CoT) elaboration pipeline that explicitly aligns metaphorical attributes with visual entities. This reduces

over-literalization and enhances visual relevance for abstract prompts.

Critique and Self-Refinement Frameworks: Our method, CARVE, situates itself in the broader class of post-hoc critique and iterative refinement frameworks, where generative outputs are analyzed and improved through model-guided feedback.

In language generation, SELF-RAG (Asai et al., 2024) introduced a retrieve–generate–critique loop for improving factual correctness. Krishna et al. (2023) showed that post-hoc explanations of language model predictions can enhance downstream performance through self-rationalization. Similarly, Zhong et al. (2024) demonstrated how LLMs can be harnessed as post-hoc correctors to revise flawed generations using targeted critique.

While prior elaboration pipelines provide valuable abstractions for metaphorical prompts, they remain vulnerable to semantic drift, over-literalization, and incomplete grounding. These limitations motivate our approach which uses a vision–language model to identify representational gaps and guide an LLM in refining the visual elaboration accordingly.

To the best of our knowledge, **CaRVE** is the first critique-based figurative text-to-image generation framework.

B Practicality, Efficiency, and Computational Overhead Analysis

We analyze the runtime latency, GPU-hour cost, and memory footprint of CaRVE relative to the vanilla two-stage pipeline to quantify the practicality and deployability of our framework.

B.1 Latency

Table 4 reports the average per-instance inference latency. The vanilla two-stage pipeline requires 46.55 seconds per instance, while CaRVE incurs a latency of 97.75 seconds, corresponding to a 2.1× increase due to the additional critique pass. This behavior is expected, as the text-to-image generation stage (executed for 50 diffusion steps) constitutes the dominant contributor to inference latency; since CaRVE performs image generation twice, once in the initial stage and once after the critique pass, the resulting latency is approximately doubled relative to the vanilla pipeline.

Method	Latency (Seconds / Instance)
Vanilla	46.55
CaRVE	97.75

Table 4: **CaRVE takes roughly double the amount of inference time compared to the Vanilla pipeline.**

B.2 GPU Cost

We also measure computational cost in GPU-hours with the amount of time required to generate 1100 instances using the Vanilla method and using CaRVE. Over 1100 instances, the vanilla pipeline consumes 14.22 GPU-hours, while CaRVE consumes 29.86 GPU-hours, reflecting a bounded and linear overhead.

Method	Total GPU-hours
Vanilla	14.22
CaRVE	29.85

Table 5: **Total GPU-hours over 1100 instances.**

B.3 GPU Memory Footprint

Table 6 summarizes the peak GPU memory usage of both pipelines. The vanilla pipeline requires 23 GB of VRAM, consisting of 13 GB for the language model and 10 GB for the text-to-image model. CaRVE introduces an additive 14 GB footprint for the critique VLM, resulting in a total of 37 GB VRAM.

Method	Peak VRAM Usage
Vanilla	23 GB
CaRVE	37 GB

Table 6: Peak VRAM Usage while generating 1100 instances with a batch size of 1.

B.4 Practicality

Although CaRVE incurs a 2.1× latency overhead and an additive +14 GB VRAM footprint, it remains fully deployable and scalable. All components of CaRVE are open-source and rely exclusively on sub-4B models, without any proprietary APIs or human-in-the-loop verification. As a result, the computational cost scales linearly with dataset size, enabling predictable and cost-efficient large-scale deployment while delivering substantial performance improvements over the vanilla pipeline as demonstrated in Table 1 and Figure 3.

# Critique Passes	Performance Score ↑	Latency (s/inst) ↓
0 (Vanilla)	0.522	46.55
1 (CaRVE)	0.585	97.75
2	0.587	148.90
3	0.550	199.80

Table 7: Effect of increasing the number of critique passes on performance and latency.

C Ablations

C.1 Number of Critique Passes

We analyze the effect of increasing the number of critique passes on performance and computational cost. In addition to the default single critique pass used in CaRVE, we evaluate variants employing two and three successive critique passes. Each additional pass introduces an extra round of critique generation followed by a subsequent text-to-image generation stage, thereby increasing both latency and GPU cost.

As shown in Table 7, the introduction of a single critique pass yields a substantial improvement over the vanilla pipeline. However, additional critique passes provide only marginal gains while incurring near-linear increases in inference latency. This diminishing return indicates that a single critique pass offers the best trade-off between performance improvement and computational overhead, motivating our design choice in CaRVE.

We also observe that increasing the number of critique–generation passes leads to a degradation in output quality. This degradation arises primarily from errors compounding across successive iterations, as well as from the tendency of later passes to over-adjust already satisfactory visual elements, thereby introducing unnecessary details or unintended artifacts.

C.2 VLM Critic Size

We analyze the effect of VLM critic capacity on the performance of the critique-based feedback loop. Our primary setup employs a 4B-scale critic; increasing the critic size can yield more precise and semantically grounded feedback, leading to improved refinements. However, this comes at the cost of additional computational overhead. In particular, configurations combining a lightweight base model with a significantly larger critic (e.g., 4B LLM + 12B VLM) may approach or exceed the resource requirements of larger standalone models (e.g., 14B Vanilla). These observations highlight a trade-off between critic capacity and efficiency, suggesting that the most favorable regime lies in pairing smaller base models and critics to achieve the best performance–efficiency trade-off.

D Improvements due to CaRVE

D.1 Sub-metric Analysis for Automatic Evaluation

Our text-to-text automatic evaluation relies on three complementary sub-metrics: *Flesch Readability*, *Literality*, and *Faithfulness*. We analyze the improvements induced by CaRVE across these sub-metrics to identify which aspects of the generated visual elaborations are explicitly enhanced by the critique-based feedback mechanism.

As shown in Table 8, CaRVE consistently outperforms the Vanilla pipeline across all three sub-metrics, yielding absolute improvements of +0.09 in Flesch Readability, +0.10 in Literality, and +0.12 in Faithfulness. These gains indicate that the critique-based feedback loop enhances multiple complementary aspects of visual elaboration quality. The improvement in readability reflects increased linguistic clarity and concreteness, resulting in more visually executable descriptions, while the gains in literality and faithfulness demonstrate reduced over-literalization and improved preservation of the intended metaphorical meaning. Collectively, these results suggest that CaRVE functions as a general semantic alignment mechanism rather than merely improving surface-level fluency.

D.2 Qualitative analysis of Visual Elaboration improvement

We present qualitative examples to illustrate how CaRVE improves the semantic grounding and expressiveness of visual elaborations. While the Vanilla pipeline often exhibits overly literal or surface-level interpretations of figurative inputs, CaRVE produces outputs that better capture the intended meaning, tone, and abstraction. Across the following examples, we highlight three recurring patterns: reduction of overliteralization, improved semantic alignment with figurative intent, and more

Method	Readability ↑	Literality ↑	Faithfulness ↑
Vanilla	0.53	0.75	0.76
CaRVE	**0.62**	**0.85**	**0.88**

Table 8: Sub-metric comparison between visual elaborations generated using the Vanilla pipeline and CaRVE.

coherent tonal interpretation. These cases complement our quantitative results by demonstrating the nature of improvements that are not fully captured by automatic metrics.

Example: Reducing Overliteralization

Input: The ripples wimple on the rills, like sparkling little lasses.

Vanilla: A narrow forest stream flows through a clearing. Small glowing girls appear to dance across the rippling water. Their shimmering dresses catch the light as they skip and play, as if the stream itself has transformed into tiny, sparkling lasses.

CaRVE: A small stream flows slowly through a forest clearing. Water moves in gentle ripples, reflecting the soft light of early morning. The ripples are clear and bright.

Example: Semantic Alignment

Input: Home is a shell.

Vanilla: A large seashell sits on a sandy beach. Inside the shell is a small house with windows and a door carved into its surface, as waves roll in behind it.

CaRVE: A quiet house with empty chairs around a table, sunlight streaming through dusty windows. The walls are bare, and the air feels still and cold, with no laughter or voices to fill the space.

Example: Tonal shift

Input: Love is like Jenga.

Vanilla: A couple sits at a table playing a game of Jenga. Wooden blocks are stacked in the center, and one person reaches forward to pull out a piece while smiling. The scene looks casual and playful.

CaRVE: A couple sits across from each other at a kitchen table, staring at a broken vase between them. One clenches their fists while the other avoids eye contact. The room is quiet, lit by a low candle and the faint hum of a refrigerator, with a dripping faucet adding to the tense mood.

E Examples from the MetaCaRVE Dataset

We present additional qualitative examples from MetaCaRVE alongside corresponding samples from HAIVMet-Data and GOME-Data. Figure 4 illustrates that CaRVE consistently produces higher-quality, more semantically aligned visual elaborations and images that better capture the intended metaphorical meaning.

F Dataset Generation Details

All enhanced elaborations in MetaCaRVE are generated using a three-stage pipeline composed exclusively of open-source sub-4B models. Given a metaphorical caption, an initial literal visual elaboration is produced using Qwen-3-4B (fp16). A structured critique identifying over-literalization, ambiguity, and missing visual details is then generated using the same model. The final refined elaboration is obtained by conditioning on this critique. Image generation is performed using PixArt-α with 50 diffusion steps at 1024×1024 resolution and CFG 7.5. Text decoding uses temperature 0.7, top-p 0.9, and a maximum length of 256 tokens. The complete pipeline is released to ensure full reproducibility.

Metaphorical Input	HAIVMet-Data	GOME-Data	MetaCaRVE
Here comes the fool with his foggy brain.			
The ripples wimple on the rills, like sparkling little lasses			
Language is a road map of a culture.			
Monty swam in a sea of diamonds.			

Figure 4: Examples from the MetaCaRVE dataset.

Evaluator	Human A	Human B	Qwen 2.5 7B	LLaMA 3.1 8B	Gemma 2 9B
Human A	-	0.59	0.50	0.47	0.49
Human B	0.59	-	0.48	0.47	0.47
Qwen 2.5 7B	0.50	0.48	-	0.82	0.80
LLaMA 3.1 8B	0.47	0.47	0.82	-	0.86
Gemma 2 9B	0.49	0.47	0.80	0.86	-

Table 9: Pairwise Spearman rank correlations between human annotators and automatic evaluators.

G Extended Automatic Evaluations using multiple LLMs

G.1 Models Used

To ensure that our automatic evaluation is not biased toward any particular model, we analyze the agreement between multiple large language models used as evaluators. All evaluator LLMs are substantially larger than the models employed within our framework (all < 4B parameters), and therefore serve as strong external judges. Specifically, we use the following LLMs for evaluation: (i) Qwen 2.5 7B, (ii) LLaMA 3.1 8B, and (iii) Gemma 2 9B.

G.2 Results

We evaluate the performance of Vanilla and CaRVE using three different LLM evaluators to mitigate potential bias toward any single model. As shown in Table 10, the resulting scores exhibit limited variation across evaluators, which can be attributed to the coarse-grained nature of the literality and faithfulness scales and the shared training distributions of modern instruction-tuned LLMs.

Evaluator LLM	Vanilla ↑	CaRVE ↑
Qwen 2.5 7B	0.681	0.785
LLaMA 3.1 8B	0.673	0.764
Gemma 2×9B	0.675	0.775

Table 10: Text-to-text (TT) evaluation across multiple LLM evaluators.

G.3 Inter-LLM Correlation

As shown in Table 9, the evaluator models exhibit strong mutual agreement, with pairwise Spearman correlations in the 0.8s. This high level of consistency indicates that the automatic evaluation is stable across different LLM architectures and parameter scales, and that the observed improvements attributed to CaRVE are not artifacts of any single evaluator. Importantly, since all evaluator models are substantially larger than the sub-4B models used within our framework, these results further validate that the reported gains are consistently recognised by stronger external judges, supporting the robustness and generality of our evaluation protocol. In Table 1, we report scores obtained using Qwen 2.5 7B due to its comparatively higher correlation with human evaluators.

H Model Details

This section details the exact model names and versions used for our experiments:

1. LLM for CaRVE: Qwen/Qwen3-4B (Yang et al., 2025)
2. VLM for CaRVE: google/gemma-3-4b-it (Team et al., 2025)
3. Pixart-α: PixArt-alpha/PixArt-XL-2-1024-MS (Chen et al., 2024)
4. Lumina: Alpha-VLLM/Lumina-Image-2.0 (Qin et al., 2025)
5. LLM as a Judge: Qwen/Qwen2.5-7B-Instruct (Qwen et al., 2025); google/gemma-2-9b (Team, 2024); meta-llama/Llama-3.1-8B (Grattafiori et al., 2024)

I Annotation Details

I.1 Annotator Details

We employed five annotators between the ages of 25 and 30, all of whom were proficient in English and had prior experience in linguistic annotation. All annotators were working in the field of computational linguistics. They were compensated fairly with a competitive stipend for their time and contributions.

I.2 Human Evaluation Guidelines

General Instructions

1. You are required to annotate a total of **40 instances**.
2. Each instance should take approximately **2 minutes**, so the full evaluation will take around **80–90 minutes**.
3. Each instance consists of one text description (Column B) and four images (Columns C–F), each generated using a different method.
4. Your task is to **rank the images** in order of how well they express the meaning of the text, and enter your ranking in **Column G**.

Task Instructions

Step 1: Understand the Text

- The text in Column B contains figurative language.
- Take a moment to fully grasp the meaning of the text before evaluating the images.
- Do not proceed to Step 2 until you feel confident in your understanding of the text.

Step 2: Evaluate the Images

- Review the corresponding images in Columns C–F.
- Analyze how each image attempts to visually convey the intended meaning of the text.

While ranking the images, please keep in mind:

A. **Do not judge based solely on visual appeal.**

B. **Focus on how well the image expresses the core meaning of the text.**

B1. Look for meaningful representation of the text's concepts. For example, for "He is a lion," simply placing a man and a lion in the image is not enough. The image should convey bravery.

B2. Avoid rewarding over-literalization. For example, "He has a heart of gold" is about kindness. A literal golden heart misses the point unless done creatively and meaningfully.

B3. Not all metaphor components must be visualized. For "He is a lion," there's no need to show a lion if bravery is well communicated.

C. **Use your judgment.** These examples are meant to guide your thinking, but your final ranking should reflect your interpretation.

D. **If you have questions:**

- You may reach out to us for technical clarifications only.
- We cannot help interpret the text or images.

Step 3: Record Your Preference

- Once you've finalized your ranking, enter it in Column G.
- Image labels:
  - Image 1 → Column C
  - Image 2 → Column D
  - Image 3 → Column E
  - Image 4 → Column F

How to write your ranking (no ties): If your preference is Image 4 > Image 3 > Image 1 > Image 2, write: `4312` in Column G.
How to indicate ties:

- Use parentheses `()` to group tied images.
- Example 1: If Image 1 = Image 4 > Image 2 = Image 3 → `(14)(23)`
- Example 2: If all images are equally good/bad → `(1234)`

Please try to avoid excessive use of ties. Only use them if you genuinely find the images indistinguishable in quality.

If anything remains unclear, feel free to reach out for clarification. Thank you for your participation!

J Prompt details

This section outlines the prompts employed across different inference setups. There are mainly four prompts:

1. Prompt for grounded interpretation: Generates grounded, literal interpretations of figurative text using few-shot examples. (Figure 5)

2. Prompt for Visual Elaboration generation: Generated visual elaboration from grounded meaning and original figurative text using few-shot examples. (Figure 6)

3. Prompt for VLM Chain-of-Thought: Generates targeted feedback after analysing the image, the figurative text and the grounded interpretation is a step-by-step fashion. (Figure 7)

4. Prompt for LLM as a judge: Generates Literality and Faithfulness scores for our composite metric. (Figure 8)

For brevity, we omit some few-shot examples used across prompts.

Few-shot prompt for grounded interpretation generation

You are a language model that specialises in understanding figurative expressions. Given a sentence containing a figurative phenomenon, your task is to identify and explain its core grounded meaning in simple, literal terms. The grounded meaning should capture what the sentence intends to convey without figurative language. Respond only with the explanation, without adding extra commentary or repeating the sentence.

Example 1:
Input Sentence:
"Her words were a soothing balm on his wounded heart."

Output:
She spoke kindly or comfortingly, which made him feel better emotionally.

Example 2:
Input Sentence:
"His mind was a steel trap."

Output:
He was very quick and sharp at remembering or understanding things.

Example 3:
Input Sentence:
"The city was a jungle at night."

Output:
The city was chaotic, dangerous, or unpredictable at night.

Example 4:
Input Sentence:
"She's the sunshine of my life."

Output:
She makes me very happy and brings joy to my life.

Example 5:
Input Sentence:
"Time is a thief that steals our moments."

Output:
Time passes quickly and causes people to lose opportunities or experiences.

Now, given the following:

Input Sentence:
"text"

Output:

Figure 5: Few-shot prompt for grounded interpretation generation

Few-shot prompt for Visual Elaboration generation

You are a language model that specialises in creating visual descriptions that match a given meaning. Given a short explanation of a concept or emotion, describe a realistic, concrete visual scene that would align with and convey that meaning. Use simple, literal language without metaphors or abstract expressions.

Example 1:
Input Sentence:
"Her words were a soothing balm on his wounded heart."
Meaning:
She spoke kindly or comfortingly, which made him feel better emotionally.

Output:
A young woman sits beside a man on a park bench. She gently touches his shoulder, smiling softly as he wipes away a tear. They are surrounded by quiet trees and warm sunlight.

Example 2:
Input Sentence:
"His mind was a steel trap."
Meaning:
He was very quick and sharp at remembering or understanding things.

Output:
A boy in a classroom eagerly raises his hand as soon as the teacher finishes asking a question. His notebook is neatly filled, and his eyes are focused and bright.

Example 3:
Input Sentence:
"The city was a jungle at night."
Meaning:
The city was chaotic, dangerous, or unpredictable at night.

Output:
A busy street filled with honking cars, people rushing in all directions, flashing neon signs, and dark alleys where shadows move unpredictably.

Now, given the following:
Focus your scene generation primarily on the meaning provided, not the figurative text. The figurative text is only there for reference and as a sanity check. Your task is to bring the meaning to life through a concrete visual scene.

Input Sentence: text
Meaning: meaning
Output:

Figure 6: Few-shot prompt for Visual Elaboration generation

Chain-of-Thought Prompt for Figurative Evaluation

You are a vision-language expert tasked with evaluating how well an image conveys the meaning of a figurative sentence. Your job is to identify where the image aligns with or diverges from the intended figurative meaning, and to suggest refinements that could help the image better express the core idea.

The evaluation will be guided by three key semiotic dimensions:

1. Conceptualisation — identifying and mapping the key concepts involved in the figurative expression.
2. Expression — assessing how abstract and concrete elements are visually represented.
3. Communication — evaluating how effectively the image conveys the intended figurative meaning to a viewer.

Follow this step-by-step chain-of-thought reasoning process:

1. Conceptualisation
- What is the **primary concept** (the subject of the figurative sentence)?
- What is the **secondary concept** (what it is being compared to)?
- What is the **attribute** that links the two?
- Does the image clearly reflect this figurative mapping?

2. Expression
- Are both **concrete and abstract elements** from the figurative meaning visually represented?
- Is the figurative idea expressed in a **creative and meaningful way**?
- Is there **symbolic or artistic abstraction**, or is the representation overly literal?

3. Communication
- Does the image successfully convey the intended emotional or conceptual message?
- Are there any elements that are distractingly literal, irrelevant, or misleading?
- Would a viewer unfamiliar with the figurative input still grasp the core figurative idea?

Output
Based on your analysis, provide actionable suggestions for improvement as a JSON object with three fields: ‘modify‘, ‘add‘, and ‘remove‘. Each field should contain natural language suggestions to improve the image’s alignment with the figurative input.

Format:
“‘json

"modify": ["..."],
"add": ["..."],
"remove": ["..."]

Only include suggestions that are grounded in the analysis above. Be specific and concise.

Figure 7: Chain-of-Thought Prompt for Figurative Evaluation

LLM as a Judge prompt

You will be given an input text and a visual elaboration generated from that input text. The visual elaboration is intended for use in a text-to-image model to produce an image. Your task is to evaluate how well the visual elaboration captures the *intended meaning* of the figurative text. Carefully look at the text and think about what it really means.

Your evaluation has two parts:

1. **Faithfulness** (0 or 1):
- 1 = The visual elaboration clearly and accurately conveys the intended meaning of the figurative text
- 0 = The elaboration fails to capture the text's meaning, is too figurative, or directly paraphrases the metaphor.

2. **Literality** (1 to 3):
- 3 = Fully literal and visualizable; uses no metaphorical or symbolic phrases.
- 2 = Mostly literal with minor metaphorical hints or paraphrasing.
- 1 = Largely non-literal; uses figurative language or reuses metaphorical phrases from the input.

Return your evaluation in JSON format using the following keys:
- ‘"Faithfulness"‘ (0 or 1)
- ‘"Literality"‘ (1, 2, or 3)
- ‘"Justification"‘ (a short explanation of your reasoning)

Do not include any explanation outside the JSON.

Example:

Metaphor: Her voice poured like warm honey through the noise.
Visual Elaboration: An opera singer stands in front of a fully packed theatre. The back of the crowd is standing up and appear to be talking to each other. The front of the crowd is hyptonized by her singing and look directly at the singer. As her voice, symbolized by staff notes, reaches farther parts of the audience, they also get mesmerized.
Output:

"Faithfulness": 1,
"Literality": 3,
"Justification": "It aligns with the metaphor's meaning by showing the calming effect of her voice. The scene is literal and does not use metaphorical phrasing."

Now, it is your turn to evaluate. Output in the specified JSON format ONLY. No additional text or explanation is needed. Only the JSON.

Figurative Text: text
Visual Elaboration: ve
Output:

Figure 8: LLM as a Judge prompt