Let's ViCE! Mimicking Human Cognitive Behavior in Image Generation Evaluation

Federico Betti
federico.betti@unitn.it
University of Trento
Trento, Italy

Jacopo Staiano
jacopo.staiano@unitn.it
University of Trento
Trento, Italy

Lorenzo Baraldi
lorenzo.baraldi@phd.unipi.it
University of Pisa
Pisa, Italy

Lorenzo Baraldi
lorenzo.baraldi@unimore.it
University of Modena and Reggio Emilia
Modena, Italy

Rita Cucchiara
rita.cucchiara@unimore.it
University of Modena and Reggio Emilia
Modena, Italy

Nicu Sebe
nicu.sebe@unitn.it
University of Trento
Trento, Italy

ABSTRACT

Research in Image Generation has recently made significant progress, particularly boosted by the introduction of Vision-Language models which are able to produce high-quality visual content based on textual inputs. Despite ongoing advancements in terms of generation quality and realism, no methodical frameworks have been defined yet to quantitatively measure the quality of the generated content and the adherence with the prompted requests: so far, only human-based evaluations have been adopted for quality satisfaction and for comparing different generative methods. We introduce a novel automated method for *Visual Concept Evaluation* (ViCE), i.e. to assess consistency between a generated/edited image and the corresponding prompt/instructions, with a process inspired by the human cognitive behaviour. ViCE combines the strengths of Large Language Models (LLMs) and Visual Question Answering (VQA) into a unified pipeline, aiming to replicate the human cognitive process in quality assessment. This method outlines visual concepts, formulates image-specific verification questions, utilizes the Q&A system to investigate the image, and scores the combined outcome. Although this brave new hypothesis of mimicking humans in the image evaluation process is in its preliminary assessment stage, results are promising and open the door to a new form of automatic evaluation which could have significant impact as the image generation or the image target editing tasks become more and more sophisticated.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**.

KEYWORDS

image generation, automatic evaluation

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612706

ACM Reference Format:
Federico Betti, Jacopo Staiano, Lorenzo Baraldi, Lorenzo Baraldi, Rita Cucchiara, and Nicu Sebe. 2023. Let's ViCE! Mimicking Human Cognitive Behavior in Image Generation Evaluation. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3581783.3612706

1 BRAVE IDEA INTRODUCTION

Quantitatively assessing the results of image generation models is a complex task. The challenge is clear when we consider simpler generative models such as Variational Autoencoders (VAEs), and it escalates when we delve into unsupervised or self-supervised models like Generative Adversarial Networks (GANs) or Diffusion Models [7, 31]. Often, the evaluation of proposed models is based on some measurements in the latent space or on specific features extracted from the generated images [3, 11]; sometimes it is associated with checking the sharpness and diversity of results in comparison to a reference test dataset [12]. Thus, whether the model generates images from a single text prompt or modifies an existing image based on a textual input (a process commonly referred to as Image Target Editing), accurately gauging their effectiveness is an ongoing challenge.

In fact, thus far, the only universally agreed upon evaluation methodology is the ultimate human judgment.

In recent years, new models and commercial products have been introduced that offer unprecedented perceptual quality and realistic representation. Systems for prompt-based generation, such as those based on Diffusion Models [2, 13, 31], and multimodal models that allows for partial modification of the input image [5, 36], at first glance, seem to deliver satisfactory results. However, appearances can be misleading.

Given that the research community largely agrees that metrics are necessary for benchmarking these generative process, the key question here is: how can we evaluate the effectiveness of a generative process triggered by text or multimodal input without ground truth? This includes other underlying inquiries like: does the output image meet perceptual and semantic expectations? Does the output image meet the constraints of the textual prompt? Does the image accurately reflect changes requested during a generative editing process in a multimodal setting? Until now, no reference-less quantitative scoring framework has been proposed.

Figure 1: Different types of Image Generation tasks. *Top*: in a multimodal Image Targeted Editing setup, given an input image paired with a textual instruction, the generative system is called to modify the former according to the latter. *Bottom*: in a cross-modal image generation setup, the generative system is called to produce an image based on the textual description provide as the sole input.

Let's consider a situation where we ask a generative system to partially modify the content of a given image: we could request to only change the color of the motorbike to green (see Figure 1). Here, a concrete challenge lies in determining whether the system can fulfill these kind of requests without introducing unintended alterations, and still effectively implement the desired modifications and maintain a high level of perceptual quality. In other words, *how can we evaluate whether the system has precisely executed the requested changes, neither exceeding nor falling short, and has produced an aesthetically pleasing output?*

Right now, human input is crucial for this process. This involves either having humans annotate data beforehand or getting humans to assess the result after the generative process. Such human feedback can also be harnessed to continuously improve the models, e.g. through reinforcement learning methods [24].

Despite the complex task of perfectly imitating human judgment, we can aim to emulate the strategy employed by humans to make judgments, mainly by asking and answering questions. This provides a viable approach to modelling human judgement within an AI system. This is the essence of what we term as human-aligned Visual Concept Evaluation (ViCE).

To sum up, our contributions include:

- a novel interpretation method for images based on question answering, that reflects the human cognitive process;
- a universal evaluation protocol applicable to all image generation tasks, including Image Targeted Editing (ITE);
- an AI system, which leverages Large Language Models (LLMs) for dynamic question generation, which circumvents reliance on a static question pool;
- a semantic complement to perceptual quality metrics, contributing additional depth to evaluations, rather than attempting to replace existing metrics;
- enhanced alignment with human evaluations, bolstering the trustworthiness and authenticity of AI-generated assessments.

Embarking in this direction is indeed bold, as it seeks to bridge the cognitive gap between artificial intelligence systems and humans.

2 RELATED WORKS

Text-to-Image generation and editing. Over the past few years, many approaches have emerged in the realm of Generative AI, aiming to enhance the efficacy of image-generation tasks. Notably, advancements in the application of Generative Adversarial Network (GAN) [10, 29, 32, 35] and Diffusion Model [2, 7, 13, 15, 25, 31] have significantly elevated the current state-of-the-art with regard to the Text-to-Image paradigm, which involves the generation of an image from a given textual description (or prompt). For example, in [26] each step of the diffusion process is conditioned on the textual prompt input by the user, resulting in an output aimed at representing the starting textual concept.

Given the significant advances in this field, more recent efforts have enlarged the scope of the Text-to-Image paradigm to encompass human-written instructions for image editing [5, 14, 36]. In this particular task, the objective is to manipulate the semantics of an image using a textual prompt, while simultaneously avoiding any undesired alterations to the image itself.

Metrics for Automatic evaluation of image generation and editing. Despite the significant efforts by the research community to enhance the qualitative outcomes of image editing and generation, only a limited number of techniques have been proposed to effectively evaluate the produced results of both methods.

As emphasized in [23], current automatic metrics exhibit limited performance in evaluating Text-to-Image generation when compared to human evaluations. Metrics such as Fréchet Inception Distance (FID) [12] and Inception Score (IS) [28] primarily focus on assessing image fidelity, disregarding the alignment between the generated image and the associated text. Conversely, CLIPScore [11] aims to measure the cosine similarity between the image and text tokens that are tokenized using CLIP image and text encoders. However, there are instances [9, 21] where generative models employ this metric to optimize image generation during training, leading to potential biases and unfair measurements at evaluation time.

To address this challenge, [23] propose a solution involving human evaluation as the primary method of evaluating Text-to-Image models. Further, a recently proposed automatic metric, LLMScore [19], despite combining global and local descriptions using a Large Language Model (LLM) into an object-centric visual description, presents some limitations.

A significant drawback is that the generated captions often contain additional details that are not sourced from the image captioners, but instead fabricated by the LLM. Moreover, the final caption does not sufficiently incorporate the requirements and inputs from the original prompt, differing significantly from a human-like evaluation process. Eventually, LLMScore compares this description

with the textual prompt used during the generation process and utilizes an LLM to compute the final score. Our proposed metric strives to overcome these issues and to more effectively replicate human visual reasoning – by incorporating a pipeline that specifically evaluates the extent to which the generated image fits the textual requests.

Our work shares some commonalities with QuestEval [30], which implemented a similar strategy for text summarization tasks. In QuestEval, concepts crucial to the content were identified by means of question generation and question answering, and the summarized output was then evaluated based on the presence/absence of the same question/answer pairs.

3 VISUAL CONCEPT EVALUATION

The process of Visual Concept Evaluation, as we define it, aims to replicate human behavior during the assessment of a generated image. When a human is asked to rate, on a scale of 1 to 10, how well an image generation task has been executed, his/her brain unconsciously starts considering the "visual concepts" they expect to see within the generated image. Visual concepts go beyond basic elements, such as shapes and colors, and include complex aspects such as specific objects and their contextual interaction within a scene.

These concepts are dictated by the initial text that forms the basis for the image generation in the case of traditional image generation tasks. However, for multimodal inputs, these concepts hinge on both the text and the input image. Additionally, evaluators utilize their implicit knowledge to infer other intuitive aspects. For instance, if the prompt is "a cat on the stairs", the evaluator expects to see a cat, of which 1 to 4 legs might be visible, with the paws placed on the steps and a tail. All this information is easily deduced by a human brain and corresponds to the thought process a person goes through when assessing an image.

Hence, it is crucial to encapsulate not only the explicit instructions derived from the prompts, but also the implicit assumptions and expectations that humans naturally make. This brings forth the intricacy of the challenge - it's about recognizing and integrating these nuanced aspects of human cognition into the evaluation framework. Such broader understanding forms the foundation of the Visual Concept Evaluation process and is the key to aligning AI systems closer to human-like image assessment capabilities. To represent the creation of the visual concepts, which we denote as v_i, we can use the following formulas.

Visual concepts are generated from the text T, and, in case of ITE task, the input image I_{input}.

The visual concepts can be represented as:

$$\begin{aligned} V_T &= f(T) = v_1, v_2, ..., v_n \\ V_{TI_{input}} &= g(T, I_{input}) = v_1, v_2, ..., v_m \end{aligned} \tag{1}$$

where f translates the text into visual concepts and g translates the text and image into visual concepts, and v_i are the individual visual concepts. For the sake of simplicity and clarity in the following discussion, we will refer to them as V. Humans, likewise, as soon as they receive a prompt to inspect are able to directly generate the visual concepts.

Once visual concepts are formulated, the human being immediately proceeds to examine whether these visual concepts are manifested in the image and how they interact with each other. This exploration is not a simple casual observation, but involves an unconscious questioning process in which the mind raises a multitude of implicit questions and then attempts to answer them using its inherent ability to understand visual content. The same idea drives the ViCE process.

In ViCE, the genesis of the process is marked by the generation of a group of "blind questions." These questions are derived from the use of previously formulated visual concepts. They are called blind because they are not based, unlike the refinement questions, on information that has been seen and processed from the generated image. This can be expressed as:

$$Q_0 = q(V) \tag{2}$$

Here, Q_0 denotes the initial set of blind questions, which are generated by applying function q over the visual concepts, V.

Next, the image to be evaluated, I, is examined and, through reasoning, an effort is made to answer the questions and determine the presence of expected elements. This key step requires a comprehensive understanding and interpretation of the image.

$$A_0 = a(I, Q_0) \tag{3}$$

In the equation above, A_0 are the initial answers. The function a encapsulates the human-like capacity of the model to interpret and reason about the image to furnish responses to the blind questions. Hence, ViCE reflects the way a human mind functions when comprehending and evaluating visual content.

After obtaining the initial set of answers from the blind questions and having a clear understanding of the presence of the required visual concepts, the model (or human evaluator) has to make a decision D if to request additional information to make some aspects clearer, or to close the process and make the final evaluation. If more information is needed, the model creates a new set of questions, known as "refinement questions".

Such iterative process can be conducted indefinitely:

$$\begin{aligned} Q_i &= q'(V, Q_0, A_0, ..., Q_{i-1}, A_{i-1}) \\ A_i &= a(I, Q_i) \end{aligned} \tag{4}$$

Finally, the evaluation score is computed, using all the questions, answers, and initial visual concepts:

$$E = h(T, Q_0, A_0, Q_1, A_1, ..., Q_i, A_i) \tag{5}$$

This recurrent process mirrors the human strategy for assessing a generated image, with each phase performing a crucial function in the overall evaluation.

4 IMPLEMENTATION

In our implementation, we establish a pipeline that integrates various models, each having a specific function corresponding to the steps delineated in the previously mentioned equations 1, 2, 3, 5. The objective is to construct an autonomous system capable of evaluating synthetic generated images.

We have integrated a Large Language Model, specifically the GPT-3.5-turbo [22], for the reasoning process. We refer to this agent as the "Reasoning Model". This choice was driven by our intent

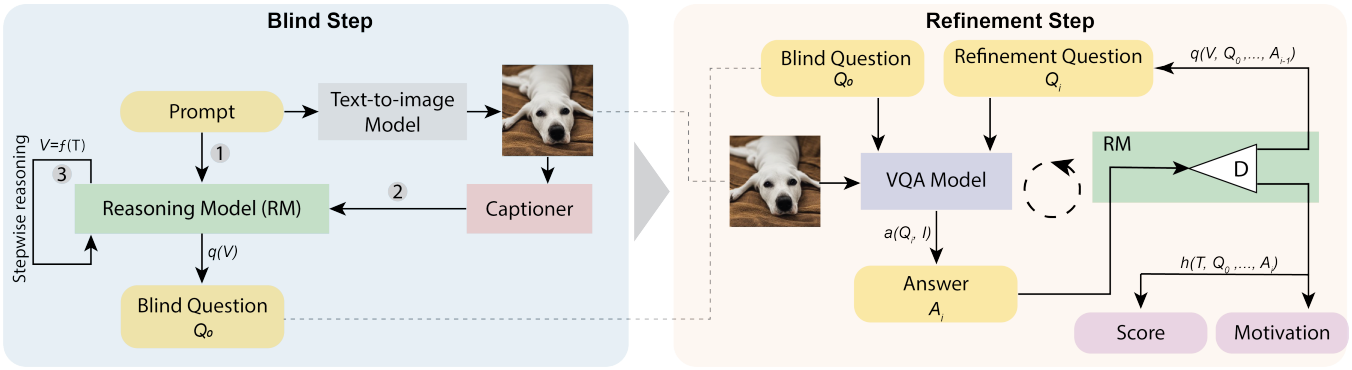


Figure 2: Visual Concept Evaluation Pipeline

to simulate human-like reasoning, which is inherently stepwise, a characteristic LLMs readily adapt to [34]. Stepwise reasoning consists in first asking the model what it expects to find in an image generated with that prompt and on what criteria it should evaluate the effectiveness of the generation. Only thereafter the actual questions are generated. As part of our future work, we plan to conduct a comparative study using different-sized LLMs and different stepwise reasoning approaches to verify and measure any impact on the results. To aid the model's reasoning process, we supplemented it with an image caption. Our analysis indicated that particularly for images that deviated from the expected generation, providing an image description significantly improved the model's capability to pose pertinent questions in the subsequent stages.

The question-generation phase unfolds in several steps. Initially, we set the model to generate a fixed number of questions (N=15 in our experiments) based on the image prompt and the expected visual concepts. Emulating the human evaluation process, the Reasoning Model may seek additional information to refine its understanding of the image. Therefore, as illustrated in Fig 2, the model is queried after the initial response phase about whether it requires further information. This triggers a refinement cycle featuring an iterative exchange of questions and answers until the model is satisfied with its comprehension of the image.

The questions span across semantic and qualitative aspects of the image, examining the presence or absence of objects described in the prompt, their interrelations, and qualitative characteristics. It is noteworthy how our Reasoning Model, in its pursuit to mirror human cognition, transcends the mere 'words' in the prompt. It comprehends the necessity to validate whether the objects are in the correct semantic relationship. For instance, in response to the prompt 'a vase of flowers', the model not only confirms the presence of the vase and the flowers but also verifies that the flowers are indeed in the vase, the vase is positioned on a surface, and the setting is congruent.

The responsibility of visual image analysis and generation of answers is vested in the Visual Question Answering (VQA) model. For our implementation, we utilized the BLIP2 model [17], built from the `salesforce-lavis` library [16]. As with the LLM, we intend to investigate the influence of the VQA model on the final output in our future endeavors. The capabilities of the VQA model are crucial as they enable the Reasoning Model to construct a detailed image schema that informs the subsequent question cycles and, ultimately, the final evaluation.

5 EXPERIMENTS

Our experimental setup focused on evaluating images that were generated from textual prompts. The key objective was to determine the extent of alignment between the evaluation scores procured from the Visual Concept Evaluation (ViCE) model and those rendered by human evaluators.

In the beginning, we used the Stable Diffusion 2 model to generate images, utilizing prompts extracted equally from a variety of datasets [6, 8, 18, 27] for a total number of 1000 images. The task given to the external evaluators was to assess the level of consistency between the prompt and the generated image by scoring on a scale from 0 to 10.

We compared the evaluation scores from our ViCE model with automated metrics such as CLIPScore [11] and BLIP-ITC/ITM [17], along with other model-based evaluation techniques like LLMScore. CLIPScore and BLIP-ITC measure the distance between the embedding of the generated image and the embedding of the prompt. BLIP ITM has an additional network submodule that outputs a probability of matching.

5.1 Comparison with Human Evaluation

We conducted a comparative study between the scores derived from human evaluations and the calculated metrics. This comparison was accomplished using two correlation coefficients: Spearman's rank correlation coefficient and Pearson's correlation coefficient; additionally, we also used the Bland-Altman plot to illustrate the agreement between human and model-derived scores. More in detail, we employed:

- Spearman's Rank Correlation Coefficient: This non-parametric measure assesses the strength and direction of the relationship between two ranked variables. As it is less sensitive to outliers and does not assume a linear relationship, it is ideal for comparing ordinal variables.
- Pearson's Correlation Coefficient: This measure evaluates the linear correlation between two continuous variables.

Model	Pearson	Spearman
CLIPscore	0.19467	0.17452
BLIP ITM	0.19404	0.18752
BLIP ITC	0.26943	0.25421
LLMScore	0.29264	**0.34065**
ViCE_5	0.25221	0.24981
ViCE_blind	0.27547	0.28325
ViCE	**0.33249**	0.32762

Table 1: Comparison of Evaluation Models. All metrics report p-value lower than 0.05, indicating statistically significant correlations. ViCE_5 applies the same pipeline with 5 questions and without refinements questions; ViCE_blind only uses the blind questions, without refinement.

- Bland-Altman Plot: This graphical method measures the agreement between two different ways of measuring a variable (in our case, human and model-derived evaluations). The plot showcases the difference between the two measurements against their average.

5.2 Results

The results presented in Table 1 reflect the evaluation carried out across several datasets, thus providing an overall score that accounts for different domains across these datasets. Notably, both LLMScore and ViCE significantly surpass all other automated metrics. An interesting observation is that while LLMScore performs better in terms of Spearman correlation, ViCE excels in Pearson correlation.

This outcome warrants a brief exploration. Spearman correlation evaluates the monotonic relationship between the two datasets, while Pearson correlation assesses the linear relationship. Therefore, ViCE's superiority in Pearson correlation might suggest a better linear relationship with the human scores.

Moving forward, our goal is to further refine ViCE by introducing an initial caption similar to the strategy employed by LLMScore. We envision that incorporating local and global descriptors, drawing from the methodology of GRIT [33], could improve the effectiveness of ViCE.

Additionally, in Table 1, we include the results from the ViCE model with only 5 initial questions ('ViCE_5') and without the refinement questions ('ViCE_blind'). Our hypothesis, which is supported by these results, suggests that reducing the number of questions or completely removing the refinement process prevents the model from effectively reason and use the visual feedback, two elements that even humans leverage during evaluation.

In Figure 3 we report the Bland-Altman graphs [4], an established method to visualize the differences between two measurement techniques. In our setup, it provides a visual representation of the agreement between a standard reference measure (i.e. the human evaluation) and an automated metric of interest (i.e. one amongst CLIPScore, LLM_score, and our proposed ViCE), while simultaneously exposing any potential biases in the assessment.

It can be observed how CLIPScore's data fluctuates in a narrow range. Conversely, LLM_score tends to assign higher scores than human assessments, a fact that indicates a potential overestimation of image quality. On the other hand, ViCE shows a balanced distribution, indicating a closer alignment with human evaluations and suggesting it can indeed offer a more reliable method for automatic evaluation.

6 EXTENSION TO ITE

Expanding on our prior discussions, we suggest that the Visual Concept Evaluation (ViCE) approach is not confined to image generation but can be extended to Image Targeted Editing (ITE). In the ITE task, the input comprises both an image and a descriptive prompt, with the latter containing instructions for the desired semantic changes to be applied to the image. Such modifications, rather than being stylistic adjustments, involve content alterations that touch upon only a section of the original image.

In the (recent) past, these models required an explicit mask to pinpoint the image section to be modified [1, 20]. However, the currently available large Vision-Language models are now capable of autonomously identifying the region for modification [5, 14, 36].

Still, the evaluation of such a task requires human evaluators to identify which parts of the image should remain untouched and which should be altered, and then to evaluate the precision of the implemented changes.

In this context, visual concepts can be divided into three distinct sets:

(1) V_{remain}: Visual concepts that should be kept from the original image;
(2) V_{remove}: Visual concepts that should no longer be present;
(3) V_{add}: Visual concepts that should be added to the output image.

Thus, the set of visual concepts V to be checked for in the edited image compounds to:

$$V = V_{\text{remain}} - V_{\text{remove}} + V_{\text{add}} \tag{6}$$

Through the reasoning process, questions related to the visual concepts that will be modified can be formulated, and the responses can subsequently be used to evaluate the effectiveness of the modification.

Visual concepts belonging to V_{remain} are expected to stay constant, and any change in the responses associated with these concepts would suggest that the portion of the image meant to be preserved has been altered. An illustrative example of this scenario can be found in Figure 4.

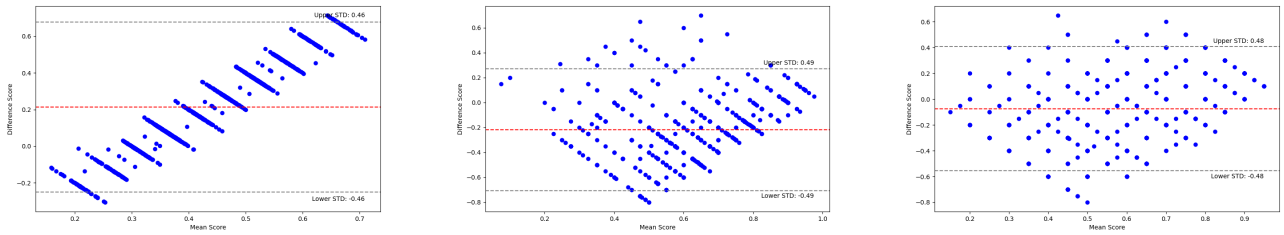
7 CONCLUSIONS

This work marks a initial step towards mirroring human reasoning when it comes to synthetic image evaluation. We have devised an approach that acknowledges both explicit and implicit facets of human cognition, creating a close alignment with human judgment.

This bold venture aims to narrow the cognitive gap between AI and humans, thereby advancing towards a more nuanced and reliable image evaluation methodology.

8 ACKNOWLEDGMENTS

This work was supported by the MUR PNRR project FAIR - Future AI Research (PE00000013) funded by the NextGenerationEU, the PRIN project CREATIVE (Prot. 2020ZSL9F9), and the Horizon Europe project "European Lighthouse on Safe and Secure AI (ELSA)" (HORIZON-CL4-2021-HUMAN-01-03), co-funded by the European



(a) CLIPScore (b) LLM_score (c) ViCE

Figure 3: Bland-Altman plots for different automated metrics.

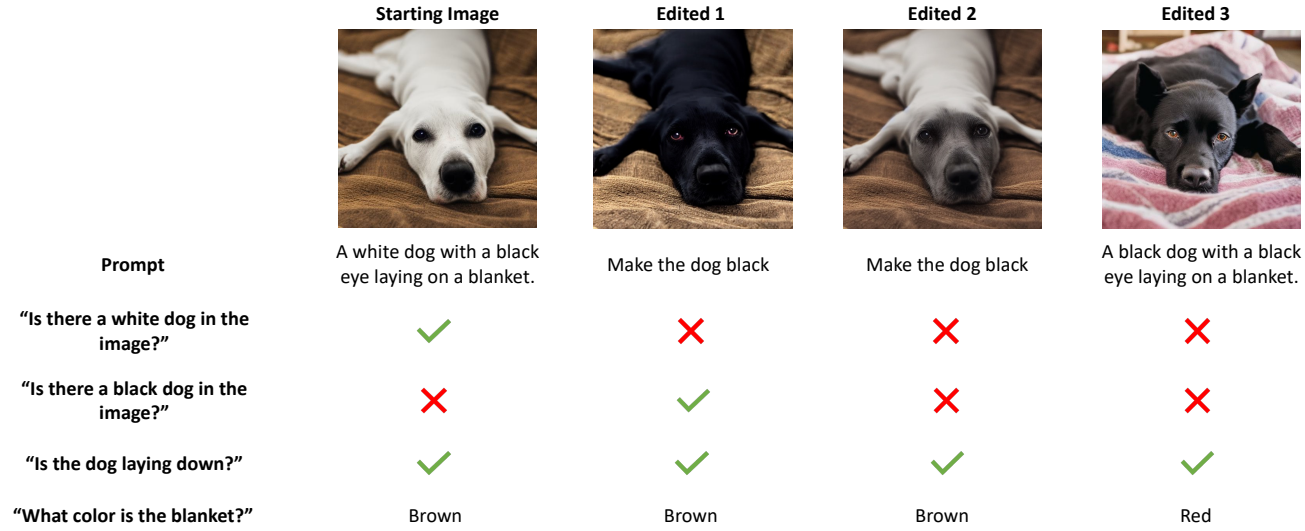


Figure 4: ViCE applied to the ITE task, whereby an LLM generates context-specific queries to assess the quality of the edit. The variation in the generated responses offers insights about the effectiveness of the edit operations.

Union (GA 101070617). The work of JS has been partially funded by Ipazia S.p.A..

REFERENCES

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended Diffusion for Text-driven Editing of Natural Images. https://doi.org/10.48550/arXiv.2111.14818 arXiv:2111.14818 [cs].

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. 2022. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. *arXiv preprint arXiv:2211.01324* (2022).

[3] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics, Ann Arbor, Michigan, 65–72. https://aclanthology.org/W05-0909

[4] J Martin Bland and DouglasG Altman. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet* 327, 8476 (1986), 307–310.

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*.

[6] Jaemin Cho, Abhay Zala, and Mohit Bansal. 2022. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053* (2022).

[7] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. *NeurIPS* (2021).

[8] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2022. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. *arXiv preprint arXiv:2212.05032* (2022).

[9] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)* (2022).

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. *NeurIPS* (2014).

[11] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*.

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2018. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. https://doi.org/10.48550/arXiv.1706.08500 arXiv:1706.08500 [cs, stat].

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *NeurIPS* (2020).

[14] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *CVPR*.

[15] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. 2016. Improved variational inference with inverse autoregressive flow. *NeurIPS* (2016).

[16] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C. H. Hoi. 2022. LAVIS: A Library for Language-Vision Intelligence. arXiv:2209.09019 [cs.CV]

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*.
[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.
[19] Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. 2023. LLMScore: Unveiling the Power of Large Language Models in Text-to-Image Synthesis Evaluation. *arXiv preprint arXiv:2305.11116* (2023).
[20] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. https://doi.org/10.48550/arXiv.2108.01073 arXiv:2108.01073 [cs].
[21] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *ICML*.
[22] OpenAI. 2023. GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774 arXiv:2303.08774 [cs].
[23] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. 2023. Toward verifiable and reproducible human evaluation for text-to-image generation. In *CVPR*.
[24] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.
[25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125* (2022).
[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.
[27] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *NeurIPS* (2022).
[28] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *NeurIPS* (2016).
[29] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. 2023. StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis. *arXiv preprint arXiv:2301.09515* (2023).
[30] Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021. QuestEval: Summarization Asks for Fact-based Evaluation. https://doi.org/10.48550/arXiv.2103.12693 arXiv:2103.12693 [cs].
[31] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*.
[32] Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. 2023. GALIP: Generative Adversarial CLIPs for Text-to-Image Synthesis. *arXiv preprint arXiv:2301.12959* (2023).
[33] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. 2022. GRiT: A Generative Region-to-text Transformer for Object Understanding. https://doi.org/10.48550/arXiv.2212.00280 arXiv:2212.00280 [cs].
[34] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. https://doi.org/10.48550/arXiv.2305.10601 arXiv:2305.10601 [cs].
[35] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*.
[36] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, Caiming Xiong, and Ran Xu. 2023. HIVE: Harnessing Human Feedback for Instructional Visual Editing. https://doi.org/10.48550/arXiv.2303.09618 arXiv:2303.09618 [cs].