Transferring Visual Attributes from Natural Language to Verified Image Generation

Rodrigo Valerio[1], Joao Bordalo[1], Michal Yarom[2], Yonattan Bitton[2,3], Idan Szpektor[2], Joao Magalhaes[1]
[1]Universidade NOVA de Lisboa
[2]Google Research
[3]The Hebrew University of Jerusalem
{r.valerio, j.bordalo}@campus.fct.unl.pt

Abstract

Text to image generation methods (T2I) are widely popular in generating art and other creative artifacts. While visual hallucinations can be a positive factor in scenarios where creativity is appreciated, such artifacts are poorly suited for cases where the generated image needs to be grounded in complex natural language without explicit visual elements. In this paper, we propose to strengthen the consistency property of T2I methods in the presence of *natural complex language*, which often breaks the limits of T2I methods by including non-visual information, and textual elements that require knowledge for accurate generation (see Figure 1). To address these phenomena, we propose a **N**atural **L**anguage to **V**erified **I**mage generation approach (**NL2VI**) that converts a natural prompt into a *visual prompt*, which is more suitable for image generation. A T2I model then generates an image for the visual prompt, which is then verified with VQA algorithms. Experimentally, aligning natural prompts with image generation can improve the consistency of the generated images by up to 11% over the state of the art. Moreover, improvements can generalize to challenging domains like cooking and DIY tasks, where the correctness of the generated image is crucial to illustrate actions.

1 Introduction

Text-to-image generation (T2I) methods [Ramesh et al., 2022, Rombach et al., 2022, Saharia et al., 2022, Chang et al., 2023] are able to map textual prompts to latent image representations in order to represent objects, actions, scenes or emotions mentioned in the prompt. Yet, these models still often produce inconsistencies between the prompt and the image [Park et al., 2021, Leivada et al., 2022], as well as hallucinations that escape visual common sense knowledge, e.g. left side of Figure 1. Visual and common sense inconsistencies are further exacerbated when the input is a natural text, instead of direct explicit drawing instructions. This

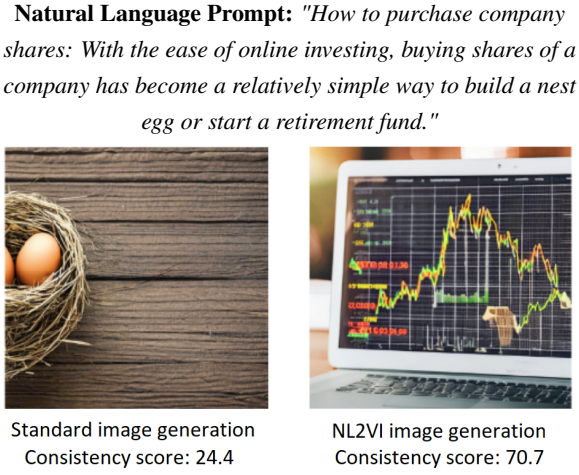


Figure 1: NL2VI can transfer the visual aspects of natural language into generated images and measure its consistency score.

is because natural texts often include information which is not depictable in images, such as emotions and domain knowledge inference e.g., "*How to buy company shares*". This creates a reliability problem when deploying T2I methods in scenarios where correctness is key, e.g. illustration of actions or data augmentation [Trabucco et al., 2023].

Previous attempts to verify the consistency of generated images have many limitations. Initial approaches [Cho et al., 2022, Gokhale et al., 2022] relied on heuristics extracted from image metadata together with an object detector. Recently, [Hu et al., 2023] generates questions with a large language model paired with Visual Question Answering (VQA) for benchmarking T2I models. However, it is important to note that the effectiveness of TIFA [Hu et al., 2023] is limited to simple and descriptive prompts.

In this paper, we move beyond synthetic and caption-based prompts and aim to *align image generation with natural prompts, while trying to guarantee the visual consistency of the generated image*. Concretely, we propose a **N**atural **L**anguage

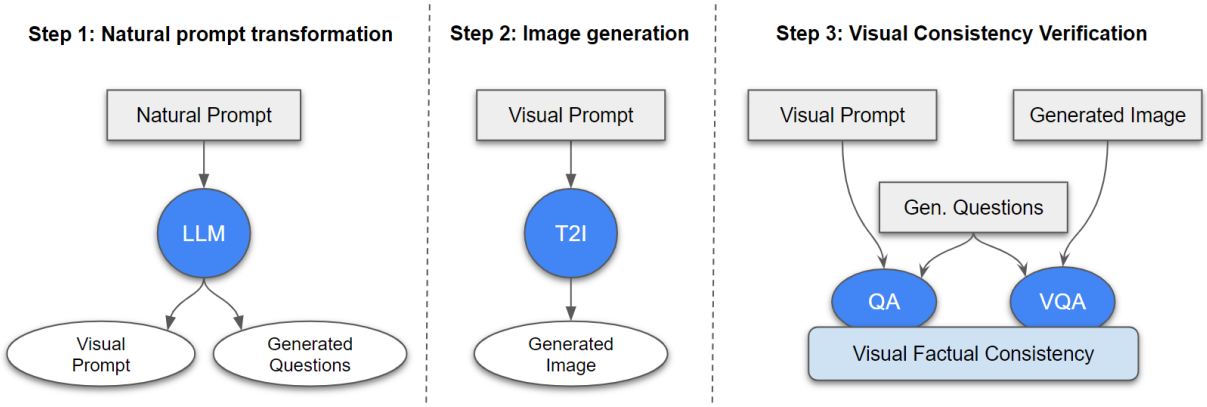


Figure 2: Natural language to verified image generation (NL2VI).

to **V**erified **I**mage generation approach (NL2VI). NL2VI first converts natural text into visually plausible text, which we refer to as a *visual prompt*. This conversion is achieved by employing a few-shot LLM that rewrites the natural text, removing non-visual aspects and providing details for textual information that would require common or domain knowledge inference to accurately image generation. Following that, we use a T2I method to generate candidate images for the visual prompt. Finally, the candidate images are ranked using a consistency verification module, and the top ranked image is offered as output.

We show that our method has significant gains on natural prompts over previous methods. Our testing demonstrates that the visual prompt depicts the visual elements in the natural prompt correctly, with an AUC of over 94%. Moreover, our method can predict the consistency of an image with an accuracy 7.8% and 11.0% higher on the recipes and DIY domains compared to the state of the art.

2 Aligning Natural Prompts with Image Generation

To improve the alignment between *natural prompts* and generated images, we propose the NL2VI approach to transform the natural prompt into a *visual prompt*, in which all the visual elements present in the natural prompt are well identified and complete, while non-visual elements are removed. We hypothesize that visual prompts would narrow the gap that T2I models need to bridge between the natural text input and the resulting image, and reduce visual hallucinations or generation of implausible artifacts.

Concretely, we propose the process depicted in Figure 2. It includes three phases, which rely on recent advances of LLMs and VQA. In the first phase, a large language model, e.g. PaLM [Chowdhery et al., 2022] or GPT-3.5, is instructed to distill a visual prompt from an input natural prompt, as well as explicitly indicate the main visual aspects that need to be verified in a generated image as a list of question/answer pairs. In the second phase, a conditioned T2I model is asked to follow the generated textual instructions to generate the image. Finally, in the third phase, a Visual Question Answering (VQA) model provides the answers to the textual questions generated in the first phase based on the generated images. VQA answers are then compared to the expected textual answers with to detect inconsistencies between the natural prompt and the generated image.

When all these aspects are taken into account, we obtain a method that offers guarantees of generating an image that is aligned with a natural prompt, even when that prompt provides no visual clues. In the following Sections, we detail each of these phases.

2.1 Visual Prompt and Question Generation for Visual Consistency

Modern LLMs have been trained on very large corpora and across a wide range of tasks. Leveraging this rich training data, LLMs are able to detect which elements of a textual passage are visually transferable into an image. Figure 1 illustrates how *"purchasing company shares"* is transferred into a *"computer screen with market shares"*. We build on this property of LLMs to translate a natural prompt into a visual prompt, and explore in-context few-shot learning to generate a visual prompt, see Table 1. Additionally, we aim to verify that a generated image, conditioned on the generated visual prompt,

follows the prompt instruction accurately. To this end, we also instruct the LLM with in-context few-shot examples to generate a set of question/answer pairs that will be used to verify the alignment between the text and the generated image, see Table 1. We have two types of questions: binary questions, which serve to verify the presence of objects from the prompt in the image, and open-ended questions, which are more general. The distribution of questions is further detailed in Table 6.

Finally, we note that since this phase is a generative process, it could also be prone to hallucinations and inconsistencies. To address potential inconsistencies in the generative process, we utilize a commonly accepted method of employing generated question/answer pairs [Honovich et al., 2021, Dagan et al., 2006, Honovich et al., 2022]. These question-answer pairs are used to validate the consistency of the generated images by evaluating if the VQA answers, based on the visual prompt and the generated images, are in accordance with the QA answers based solely on the visual prompt. We discuss this in further detail in Sections 2.3 and 4.4.

2.2 Text to Image Generation

In this phase, a T2I model is conditioned on the visual prompt, with more visual details and with fewer ambiguous descriptions, step 2 of Figure 2. All visual prompts were shorter than the input limit of the T2I methods we tested, hence no truncation happened.

2.3 NL2VI Consistency Verification

The final phase in NL2VI is to verify the consistency of the generated image with the visual prompt. The rationale is to check the consistency of the visual prompt with respect to the natural prompt, then the generated questions and lastly, the generated image. Inspired by Honovich et al. [2022] and Hu et al. [2023], we leverage the questions generated in phase 1 to probe both the image and the prompt and assess the consistency of the answers. To filter the questions based on the prompt, we utilize a Question Answering (QA) model [Khashabi et al., 2020] in conjunction with a Natural Language Inference (NLI) model [Nie et al., 2020]. Although we investigated various methods to verify the consistency between the natural prompt and the visual prompt, the fact is that the performance of the LLM has an AUC of over 90% which makes it sufficiently robust to consider that they are correct in general.

```
As an AI Image Verification Specialist, your primary responsibility
is to create a text2img prompt and examine its accuracy assuming an
associated image. Your task involves two main steps:

Construct a text2img prompt: You will be provided with a description.
It's crucial that your text2img prompt incorporates all visual
aspects mentioned in the description. The text2img prompt must be a
detailed visual description that accurately represents the visual
attributes of the description. Non-visual attributes should not be
included.

Formulate a series of questions: Your questions must be related to
the visible components within the image. Questions must be simple,
unambiguous, and answerable based on the observable content in the
image. Questions must be about elements that exist in the image and
are clearly visible. Questions must be about a single element.

Follow the examples below and complete:

Description: "Mango-Black Bean Salsa. This fiery, flavorful salsa
is excellent with tortilla chips, spooned over roasted meat or
fish, or as a topping for quesadillas. Made with mango, avocado,
no-salt-added black beans, red onion, jalapeño pepper, cilantro,
lime, salt."

text2img prompt: A bowl of mango and black bean salsa with tortilla
chips. The salsa also has onions, jalapeño peppers and cilantro.
  Q: is there a bowl of food? A: yes
  Q: is there salsa? A: yes
  Q: are there black beans in the salsa? A: yes
  Q: Are there mangos in the salsa? A: yes
  Q: are there tortilla chips? A: yes
  Q: is there cilantro? A: yes
......
(in context learning examples)
......
(unseen example)
Description: "Garlic Parmesan Pasta. The hardest part is chopping
the parsley.  Made with: parsley, garlic, butter, chicken broth,
milk, parmesan cheese, salt, ground pepper.

text2img prompt:  A bowl of garlic parmesan pasta with parmesan
cheese and parsley.
Questions:
  Q: what is in the bowl? A: pasta
  Q: is there a bowl of food? A: yes
  Q: is there cheese? A: yes
  Q: is there cheese on the pasta? A: yes
  Q: is there parsley? A: yes
```

Table 1: In-context instructions used to transform a natural prompt into a visual prompt and a set of visual consistency evaluation questions.

We are then left with the task of checking the consistency of the generated images with the questions generated by the LLM. For that purpose, we use VQA models, fine-tuned on the VQAv2 [Goyal et al., 2017] dataset, to answer the questions based on the generated image. Note that, although VQAv2 is a closed domain answer dataset, we use generative models, which may cause open-answers, in particular when there is a clear mismatch between the questions and the input text. Hence, to compare the answers, we tested several text matching algorithms, including string equality, BERTScore [Zhang et al., 2019a] and NLI [Dagan et al., 2006].

3 Experimental Methodology

In this Section, we present the NL2VI implementation details and describe the experimental setup used to evaluate NL2VI.

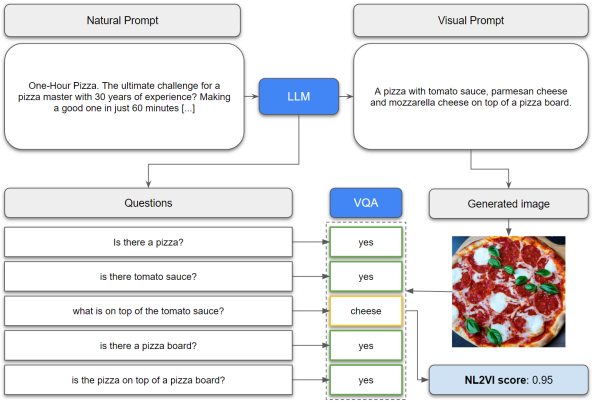


Figure 3: The NL2VI method implements an in-context few-shot learning approach: an LLM is responsible for generating the visual prompt and the verification questions for the QA and VQA methods.

3.1 Implementation

The implementation of NL2VI (see Fig. 3) relies on many pre-trained models that are used during the different steps of the pipeline. First, the natural prompt is transformed into a visual prompt and the consistency verification questions are generated, all through in-context learning. We experimented with two large language models: `gpt-3.5-turbo` from OpenAI and PaLM 540B [Chowdhery et al., 2022]. For reproducibility purposes, we release the visual prompts and generated questions, see Appendix A for details. Second, for the image generation step, we use Stable Diffusion 2.1, conditioned on the computed prompts, to generate the target image. While SD 2.1 is the current state-of-the-art, in the future, NL2VI can generalize to any other image generation method. Third, the generated questions are answered by both QA and VQA methods based on the visual prompt and on the image together with the visual prompt, respectively. As **QA** filtering model, we tested the widely popular QANLU [Namazifar et al., 2021] and the UnifiedQA model [Khashabi et al., 2020] followed by the NLI model RoBERTa-NLI [Nie et al., 2020]. For **VQA**, we experimented with BLIP [Li et al., 2022b], GIT [Wang et al., 2022a], OFA [Wang et al., 2022b], PaLI [Chen et al., 2022] and mPLUG [Li et al., 2022a].

3.2 Baselines

NL2VI is a general approach that supports any T2I base method, e.g. DALL-E2 [Ramesh et al., 2022], Imagen [Saharia et al., 2022]. In the following experiments, we used Stable Diffusion 2.1, results with more methods can be found in the supplementary material. Moreover, the image verification step can be done with other methods, such as with CLIPScore [Hessel et al., 2021] or with the TIFA model [Hu et al., 2023].

Models	LLM	VQA	Recipes	Wikihow
CLIPScore	n/a	n/a	57.4	53.8
TIFA	GPT-3.5	mPLUG	72.5	64.9
NL2VI	PaLM	OFA	78.4	73.6
NL2VI	GPT-3.5	PaLI	80.3	76.0

Table 2: Human evaluation of the different visual factual consistency methods.

3.3 NL2VI Public Dataset[1]

To study the generalization of T2I methods, we benchmark their performance in settings with natural prompts in the Recipes and WikiHow domains. The **NL2VI natural prompts and questions** dataset comprises 3000 curated natural prompts, where the correct illustration of an action and correct composition of overlapping objects is both challenging and critical. It was designed to allow for the realistic, yet controllable, research of image generation methods conditioned on natural prompts. See Annex A for details.

4 Results and Discussion

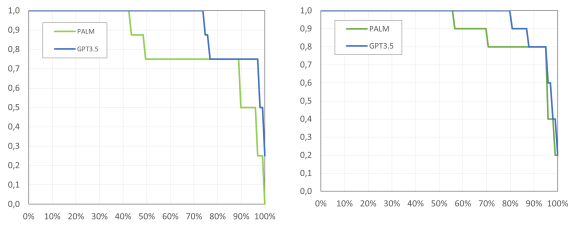
Next, we present and discuss the NL2VI method's experimental results, highlighting the key takeaways.

4.1 Verified Image Generation Results

In this section, we compare the performance of NL2VI against other verified image generation approaches. We consider several LLMs and VQA methods, as well as the CLIPScore [Hessel et al., 2021] metric and the recent TIFA model [Hu et al., 2023].

Table 2 summarizes the overall experimental results. We observed that CLIPScore has a very low variance across all generated images, with a mean value of $\sim$32% . We also noticed that this metric is not correlated with the visual consistency. This might be explained by two factors: first, CLIPScore is already used by some T2I methods as the metric to be optimized, and second, CLIPScore does not capture fine-grained information between the image and the prompt, as discussed

[1]Available after publication.



(a) Recipes domain. (b) WikiHow domain.

Figure 4: Visual prompt consistency with respect to the natural prompt.

by related works [Yuksekgonul et al., 2023]. TIFA performs better than CLIPScore, achieving good results with explicit prompts, but fails when presented with more challenging natural language prompts. NL2VI is clearly superior to the other methods, specially with the combination of GPT-3.5 and PaLI. NL2VI was able to solve common inconsistencies present in other methods, w.r.t the lack of visual common sense knowledge.

4.2 Visual Prompt Consistency

In this Section, we analyse the consistency of the visual prompt with respect to its alignment with the natural prompt. The objective is to understand how well the visual elements of a natural prompt are unambiguously captured in the visual prompt. A human annotation task was set up for this purpose. Figure 4 presents the precision curve over the annotated corpus, and Table 2 presents summarized metrics of the curves. In the Recipes domain, both LLMs perform exceedingly well, with GPT-3.5 achieving an average precision of 92.8% and PaLM 82.5%. This performance is even better in the WikiHow domain, with 90.2% and 94.9% average precision for PaLM and GPT-3.5, respectively. In terms of precision at 1 (prompts that are fully correct), precision decreases in both models, with GPT-3.5 being able to generate visual prompts that could capture the visual elements of natural prompts in 74.0% of the cases in the Recipes domain and 80.0% of the cases in the WikiHow domain. This is a highly positive result, given that natural language prompts can depict a wide range of situations and concepts, which may not have an obvious visual representation. Appendix A.3 illustrates some challenging examples.

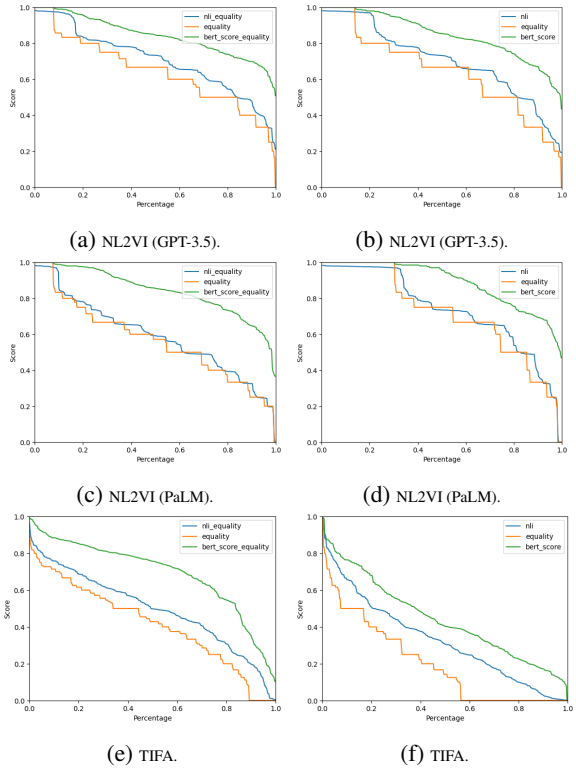
4.3 Answer Verification Methods

When comparing the VQA generative answers with the QA extractive answers, there are several discrepancies that need to be bridged due to the differences in the methods. Computing the correspondence between answers with string matching algorithms would be rather limiting, which is why we investigated NLI [Dagan et al., 2006] and BERTScore [Zhang et al., 2019a], as alternatives. Figure 5 provides a visualization of the score distribution for each method. TIFA images which are based on natural prompts have, on average, lower scores than our images generated from visual prompts. Therefore, based on NL2VI, our images are more likely to be consistent with the prompt.

Table 4 and Table 5 illustrate the comparison between answer matching algorithms, as judged by human annotators. The equality strategy is used

Models	Recipes		Wikihow	
	AUC	P@1	AUC	P@1
PaLM	82.5	42.0	90.2	56.0
GPT-3.5	92.8	74.0	94.9	80.0

Table 3: Alignment between natural language prompts and visual generation prompts.



(a) NL2VI (GPT-3.5). (b) NL2VI (GPT-3.5).

(c) NL2VI (PaLM). (d) NL2VI (PaLM).

(e) TIFA. (f) TIFA.

Figure 5: Image consistency on the Recipes (top row) and WikiHow (bottom row), according to the QA and VQA answers verification with string matching, NLI and BERTScore.

	Equality	NLI	BERT-Score
TIFA			
BLIP	64.7	68.9	72.0
GIT	62.4	67.3	**72.5**
OFA	61.8	65.4	72.2
mPLUG	64.3	67.4	71.1
PaLI	61.9	66.5	68.3
NL2VI w/ GPT-3.5			
BLIP	73.7	76.2	79.8
GIT	72.1	74.7	79.2
OFA	70.3	73.0	79.9
mPLUG	74.3	75.9	79.9
PaLI	75.5	78.3	**80.3**
NL2VI w/ PaLM			
BLIP	76.5	77.3	78.2
GIT	72.8	74.1	**78.4**
OFA	70.1	70.9	**78.4**
mPLUG	75.4	76.3	77.9
PaLI	75.7	76.5	77.7

Table 4: Ablation study on the Recipes domain: analysis of VQA and answer validation methods.

	Equality	NLI	BERT-Score
TIFA			
BLIP	56.1	61.6	64.2
GIT	56.1	62.7	64.8
OFA	56.3	62.7	64.2
mPLUG	58.4	64.3	64.9
PaLI	58.3	**65.1**	64.8
NL2VI w/ GPT-3.5			
BLIP	73.1	75.8	73.8
GIT	71.0	73.9	73.9
OFA	71.6	73.6	73.6
mPLUG	73.5	**76.1**	73.8
PaLI	74.0	76.0	74.1
NL2VI w/ PaLM			
BLIP	73.0	73.5	69.8
GIT	70.8	71.3	69.7
OFA	72.5	**73.6**	69.8
mPLUG	72.3	72.7	69.3
PaLI	72.7	72.6	70.0

Table 5: Ablation study on the WikiHow domain: analysis of VQA and answer validation methods.

for simple answers, like *yes* and *no*. NLI is used for longer answers, where entailment properties need to be checked. BERTScore achieved the best performance on the Recipes domain, while NLI was the best in the WikiHow domain. The first fact that stands out is that, independently of the method we use to validate the images, we observed that visual prompts (GPT-3.5 and PaLM) generate images that are better aligned with the natural prompt than images generated with the original natural prompt. This is due to the linguistic ambiguities that may exist in the natural prompt and need to be solved by T2I algorithms. With visual prompts, these linguistic ambiguities were solved by an LLM, which is more capable of handling linguistic idiosyncrasies. Furthermore, the visual prompt is more concise than the natural prompt and is therefore unaffected by the token limit of some T2I algorithms.

4.4 Impact of QA and VQA performance

In this Section, we report the results of an ablation study concerning the importance of the QA and VQA methods in the image consistency verification process. Table 6 presents the statistics of the image verification questions that need to be answered by the QA and the VQA methods. The key fact to note in this table is that the manual annotations of the generated questions show that 89.6% and 87.4% of the questions in the Recipes and WikiHow domains, respectively, are valid, thus confirming the overall quality of the generated questions. Moreover, we can see that many questions are open-ended, which forces the use of generative VQA methods. Table 7 presents the results of two QA methods. QA methods were restricted to span-extraction methods and

	Generated Questions	Verified Questions
Recipes		
Binary	392	386
Open-Ended	233	174
Wikihow		
Binary	403	388
Open-Ended	222	158

Table 6: Distribution of generated questions for the Recipes and WikiHow datasets before and after filtering (UnifiedQA).

	% of valid questions		Precision		Recall	
Question filtering	**Recipes**	**Wikihow**	**Recipes**	**Wikihow**	**Recipes**	**Wikihow**
QANLU	50.4	50.8	22.2	33.3	28.6	28.6
Unified-QA	89.6	87.8	90.9	84.2	71.4	76.2

Table 7: Evaluation of the question filtering stage.

Figure 6: Images generated for "A bowl of ice-cream with a spoon". There is no information about the colour or type of ice-cream, and the model must fill in the missing properties. The same can be observed for the bowl itself. Another aspect is quantity, with the image on the left having two balls of ice-cream, which is information not present in the prompt, and the other images showing a single ball. We can also see how the spoons appear deformed.

multiple-choice approaches, hence, the reason for using the QANLU [Namazifar et al., 2021] and the UnifiedQA [Khashabi et al., 2020] algorithms. UnifiedQA was superior both in terms of precision and recall, and was the model chosen for our experiments.

With respect to VQA methods, we ran an extensive ablation study as presented on Table 4 for the Recipes domain and Table 5 for the WikiHow domain. We cross-examined five VQA generative algorithms with natural prompts (TIFA) and visual prompts (PaLM and GPT-3.5) and three answer comparison methods: equality, NLI and BERTScore. From these results, we can observe that mPLUG was always the best, or the second best performing method, in all settings. PaLI was the best, or second best, in four of the six experimental settings. A key insight that we get from these results is that the performance of the VQA method is *directly connected* to the performance of the visual consistency verification of the generated image.

4.5 Hallucinations, Open-World Assumption and Visual Common Sense

Prompts fed to image generation models contain *limited information*. It is not reasonable to assume that all the information that will be present in the final image was originally present in the prompt. Some common missing aspects in prompts are the image background and object colours and texture. Examples can be seen in appendix in Figure 6. The lack of visual descriptions in the prompts, is filled in by model hallucinations, resulting in an image with more information than the original prompt. Hence, **hallucinations are needed** and unavoidable in the context of image generation, creating an **open-world setting** where multiple valid images can be generated from the original prompt. Verifying image consistency in an open-world setting is a challenging task. In the present work, instead of verifying the visual consistency in an open-world setting, we chose to verify consistency based on what is present in the visual prompt, thus adopting a conditional closed-world assumption during verification. LLMs are responsible to transform the open-world setting into a closed-world setting. A key concern is guaranteeing that these hallucinations are aligned with **visual common sense**, as some hallucinations are plausible while others invalidate the consistency of an image. In the Recipes domain, most generated images correctly depict the prompt, even when they are complex, such as in Figure 6 in the Appendix. However, T2I algorithms sometimes lack common sense [White and Cotterell, 2022], especially when the meaning of the words is not clear, as in Figure 6. This issue falls out-of-scope of the present work, as we focus

on verifying whether the elements present in the prompt show up in the final image.

5 Related work

Assessing the consistency of language or image generation methods is still an unsolved problem, despite having been addressed in the NLP field under different formulations, i.e., entailment [Li et al., 2018, Falke et al., 2019], counterfactual information [Zhang et al., 2019b], and question-answer approaches [Honovich et al., 2021, Gupta et al., 2022]. Early work was done in the related tasks of face image hallucination [Wang et al., 2014] and image forensics to detect deepfakes and image tampering [Nowroozi et al., 2021]. However, in the image generation domain, only after the publication of DALL-E [Ramesh et al., 2021] and Stable Diffusion [Rombach et al., 2022] models, has the community started to take the first steps towards assessing the visual consistency of T2I algorithms [Rassin et al., 2022, Leivada et al., 2022, White and Cotterell, 2022, Gokhale et al., 2022, Petsiuk et al., 2022, Park et al., 2021, Russo, 2022]. The hallucinations and errors of T2I methods were recently discussed by Rassin et al. [2022], Leivada et al. [2022], White and Cotterell [2022], Petsiuk et al. [2022], shedding some light on the lack of visual consistency between the generated image and the prompt. An inspiring step is taken by Russo [2022], discussing possible methods to evaluate the artistic value of image generation methods. However, better methods for evaluating visual consistency are still lacking. For example, traditional image quality metrics are not fine-grained enough, e.g., the inception score [Salimans et al., 2016] and FID [Heusel et al., 2017] are intended to measure the realism of the generated images and fail to catch inconsistencies [Park et al., 2021]. End-to-end evaluation of the similarity of image-text embeddings obtained with a dual-encoder multimodal model like CLIP [Radford et al., 2021] fails to encode compositional information, which is crucial for visual consistency evaluation [Yuksekgonul et al., 2022]. There have been attempts to improve such models, like CLIP-R Park et al. [2021] and CLIPScore [Hessel et al., 2021], but results are still suboptimal.

Another line of work proposed several synthetic benchmarks and metrics to measure the visual consistency properties of T2I methods. Early approaches relied upon object detection models and heuristics. Gokhale et al. [2022] proposed the VISOR metric, which evaluates the spatial relationship between objects detected in the scene. Similarly, Cho et al. [2022] proposed the PaintSkills dataset to assess object presence, properties and the spatial relationships described in the prompt. Drawbench is another benchmark introduced by Imagen [Saharia et al., 2022]. A more generic approach was proposed by Hu et al. [2023], which follows the literature of natural language factual consistency with QA [Honovich et al., 2022, Durmus et al., 2020]. In this method, several questions are generated from the prompt and verified in the image with a VQA method. Concurrent to our work, Yarom et al. [2023] introduces the SeeTRUE benchmark for meta-evaluation of image-text alignment. While all these methods can measure T2I average correctness with synthetic or visually descriptive prompts, they are not designed to correctly generate images conditioned on natural prompts.

6 Conclusions

Contributions. Generating images from natural language that is non-visual is, in many cases, an impossible task for image generation methods, like Stable Diffusion. In this context, the key contributions of this paper are twofold.

First, the NL2VI method to transfer the visual attributes of a natural language into a visual prompt that will generate a verified image. The visual prompt is correctly aligned with the natural prompt in over 90% of the cases and is also an enabler of a verification process based on VQA methods. The image verification step is the final safeguard to ensure that the image was correctly generated, according to the initial text.

Second, a *public dataset* with natural prompts, visual prompts and verification questions to benchmark image generation methods in the presence of natural language. The curated dataset aggregates a series of natural language prompts that are from the instructions domain with descriptions that are not always visual.

Limitations. As discussed in Section 4.5, this work assumes a closed-world setting when verifying the consistency of an image. Although our method improves consistency, image generation works in an open-world setting, and by following a prompt-based verification, we will be missing some important factors of consistency. To this end, the verification is limited by the information present in

the prompt, which limits verification. Furthermore, when generating visual prompts, some elements may be added, which were not present, originally. This can lead to some differences between the expected result and NL2VI's output. There is still important work to be done to be able to support a more broad verification of generated images.

Broader Impacts. By improving the consistency of an image generation process, we are trying to more faithfully depict what is described in the original prompt. The most obvious adverse impact, is the malicious use of generative image generation for disinformation or deceiving. We are against the applications of generative AI non-ethical uses and argue for a responsible and accountable use of these algorithms.

Acknowledgments

This work was partially funded by a Google Cloud gift and the NOVA LINCS research center, Ref. UIDP/04516/2020.

References

H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. Yang, K. Murphy, W. T. Freeman, M. Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

J. Cho, A. Zala, and M. Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*, 2022.

A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel. Palm: Scaling language modeling with pathways, 2022.

I. Dagan, O. Glickman, and B. Magnini. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment: First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, pages 177–190. Springer, 2006.

E. Durmus, H. He, and M. Diab. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.454. URL https://aclanthology.org/2020.acl-main.454.

T. Falke, L. F. R. Ribeiro, P. A. Utama, I. Dagan, and I. Gurevych. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1213. URL https://aclanthology.org/P19-1213.

T. Gokhale, H. Palangi, B. Nushi, V. Vineet, E. Horvitz, E. Kamar, C. Baral, and Y. Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*, 2022.

Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

P. Gupta, C.-S. Wu, W. Liu, and C. Xiong. DialFact: A benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.263. URL https://aclanthology.org/2022.acl-long.263.

J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

O. Honovich, L. Choshen, R. Aharoni, E. Neeman, I. Szpektor, and O. Abend. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*,

pages 7856–7870, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.619. URL https://aclanthology.org/2021.emnlp-main.619.

O. Honovich, R. Aharoni, J. Herzig, H. Taitelbaum, D. Kukliansy, V. Cohen, T. Scialom, I. Szpektor, A. Hassidim, and Y. Matias. True: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991*, 2022.

Y. Hu, B. Liu, J. Kasai, Y. Wang, M. Ostendorf, R. Krishna, and N. A. Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering, 2023.

D. Khashabi, S. Min, T. Khot, A. Sabharwal, O. Tafjord, P. Clark, and H. Hajishirzi. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.171. URL https://aclanthology.org/2020.findings-emnlp.171.

E. Leivada, E. Murphy, and G. Marcus. Dall-e 2 fails to reliably capture common syntactic processes. *arXiv preprint arXiv:2210.12889*, 2022.

C. Li, H. Xu, J. Tian, W. Wang, M. Yan, B. Bi, J. Ye, H. Chen, G. Xu, Z. Cao, J. Zhang, S. Huang, F. Huang, J. Zhou, and L. Si. mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7241–7259, Abu Dhabi, United Arab Emirates, Dec. 2022a. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.488.

H. Li, J. Zhu, J. Zhang, and C. Zong. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1430–1441, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1121.

J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022b.

M. Namazifar, A. Papangelis, G. Tur, and D. Hakkani-Tür. Language model is all you need: Natural language understanding as question answering. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7803–7807. IEEE, 2021.

Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

E. Nowroozi, A. Dehghantanha, R. M. Parizi, and K.-K. R. Choo. A survey of machine learning techniques in adversarial image forensics. *Computers & Security*, 100:102092, 2021.

D. H. Park, S. Azadi, X. Liu, T. Darrell, and A. Rohrbach. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

V. Petsiuk, A. E. Siemenn, S. Surbehera, Z. Chin, K. Tyser, G. Hunter, A. Raghavan, Y. Hicke, B. A. Plummer, O. Kerret, et al. Human evaluation of text-to-image models on a multi-task benchmark. *arXiv preprint arXiv:2211.12112*, 2022.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

R. Rassin, S. Ravfogel, and Y. Goldberg. Dalle-2 is seeing double: flaws in word-to-concept mapping in text2image models. *arXiv preprint arXiv:2210.10606*, 2022.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

I. Russo. Creative text-to-image generation: Suggestions for a benchmark. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 145–154, 2022.

C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

B. Trabucco, K. Doherty, M. Gurinas, and R. Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023.

J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022a.

N. Wang, D. Tao, X. Gao, X. Li, and J. Li. A comprehensive survey to face hallucination. *International journal of computer vision*, 106:9–30, 2014.

P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022b.

J. C. White and R. Cotterell. Schr\"{o} dinger's bat: Diffusion models sometimes generate polysemous words in superposition. *arXiv preprint arXiv:2211.13095*, 2022.

M. Yarom, Y. Bitton, S. Changpinyo, R. Aharoni, J. Herzig, O. Lang, E. Ofek, and I. Szpektor. What you see is what you read? improving text-image alignment evaluation, 2023.

M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou. When and why vision-language models behave like bag-of-words models, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou. When and why vision-language models behave like bags-of-words, and what to do about it?, 2023.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019a.

Y. Zhang, J. Baldridge, and L. He. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota, June 2019b. Association for Computational Linguistics. doi: 10.18653/v1/N19-1131. URL https://aclanthology.org/N19-1131.

A NL2VI Public Dataset

To study the generalization of T2I methods, we benchmark their performance in settings with natural prompts in the Recipes and WikiHow domains. This dataset was designed to allow for the realistic, yet controllable, research of image generation methods conditioned in natural prompts.

A.1 Statistics

The **NL2VI natural prompts and questions** dataset comprises 3000 curated natural prompts from the instructions domain, where the correct illustration of an action and correct composition of overlapping objects is critical.

A.2 Manual Annotation

In order to collect human data to use as a baseline, we created an annotation tool. As seen in Figure 7, labellers are shown 4 images, generated for a single prompt, and asked to rate them from 1 to 5. A score of 1 means the image was *Not Consistent* with the prompt, a score of 3 means it was *Somewhat Consistent* with the prompt, and a score of 3 means it was fully *Consistent* with the prompt. These annotations were collected for the all the methods we tried.

A.3 Verified Images Examples

Figure 8 presents examples of the NL2VI method. On the Recipes domain, the images from the natural prompt miss important visual ingredients, like parsley. The visual prompt improves the consistency and consists only of relevant visual information. On WikiHow, the natural prompt is penalized by being larger than the maximum sequence length allowed by Stable Diffusion. Moreover, it consists mostly of non-visual information, which degrades image generation. The visual prompt is able to overcome these problems.

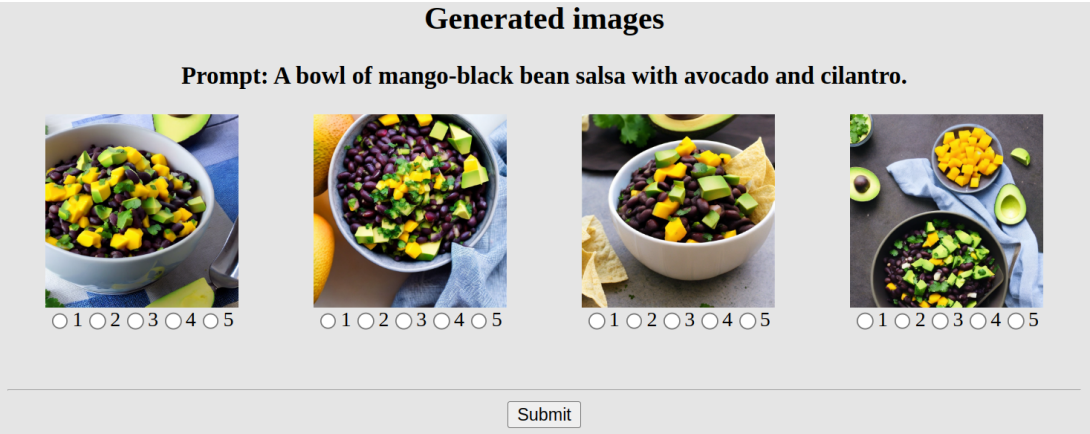


Figure 7: Image consistency annotation tool.

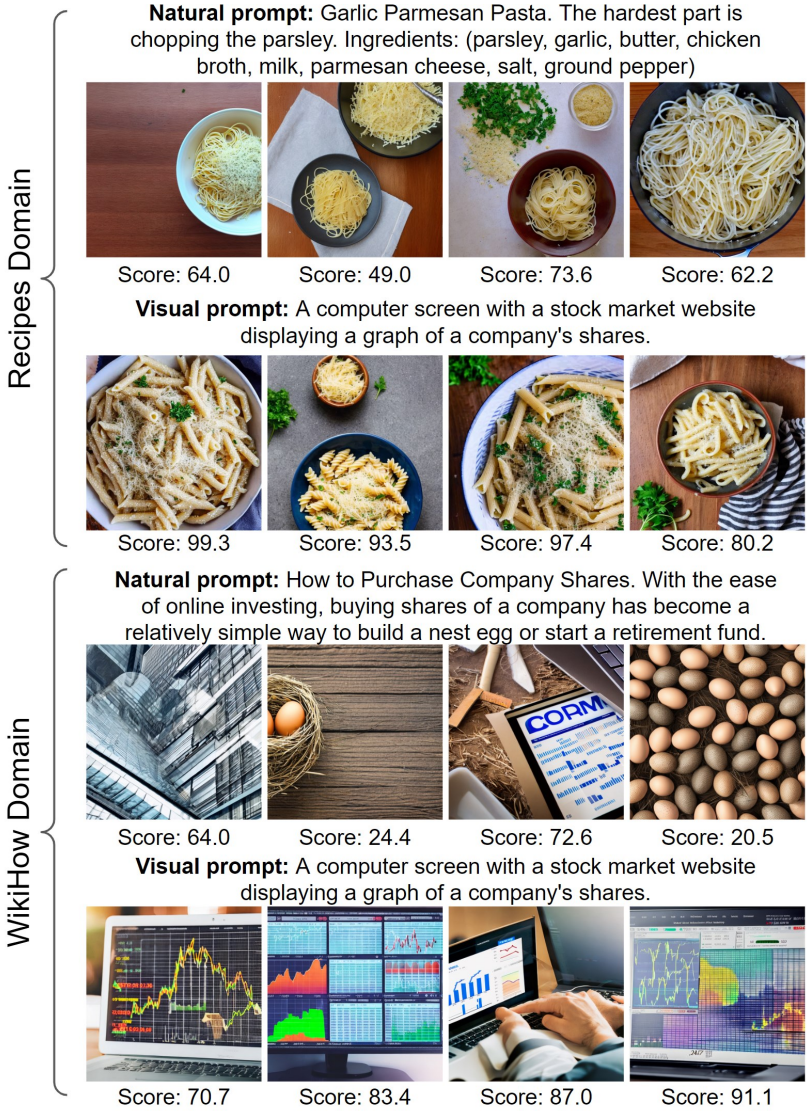


Figure 8: Comparison of natural and visual prompts for image generation across different domains.