ORES: Open-vocabulary Responsible Visual Synthesis

Minheng Ni[1]*, Chenfei Wu[1]*, Xiaodong Wang[1], Shengming Yin[1], Lijuan Wang[2], Zicheng Liu[2], Nan Duan[1]†

[1]Microsoft Research Asia
[2]Microsoft Azure AI
{t-mni, chewu, v-xiaodwang, v-sheyin, lijuanw, zliu, nanduan}@microsoft.com

Abstract

Avoiding synthesizing specific visual concepts is an essential challenge in responsible visual synthesis. However, the visual concept that needs to be avoided for responsible visual synthesis tends to be diverse, depending on the region, context, and usage scenarios. In this work, we formalize a new task, Open-vocabulary Responsible Visual Synthesis (ORES), where the synthesis model is able to avoid forbidden visual concepts while allowing users to input any desired content. To address this problem, we present a Two-stage Intervention (TIN) framework. By introducing 1) rewriting with learnable instruction through a large-scale language model (LLM) and 2) synthesizing with prompt intervention on a diffusion synthesis model, it can effectively synthesize images avoiding any concepts but following the user's query as much as possible. To evaluate on ORES, we provide a publicly available dataset, baseline models, and benchmark. Experimental results demonstrate the effectiveness of our method in reducing risks of image generation. Our work highlights the potential of LLMs in responsible visual synthesis. Our code and dataset is public available in https://github.com/kodenii/ORES.

Introduction

With the development of large-scale model training, visual synthesis models are capable of generating increasingly realistic images (Ramesh et al. 2021; Rombach et al. 2022; Saharia et al. 2022). Due to the growing risk of misuse of synthesized images, responsible AI has become increasingly important (Arrieta et al. 2020; Wearn, Freeman, and Jacoby 2019; Smith et al. 2022), especially to avoid some visual features, such as, nudity, sexual discrimination, and racism, during synthesis. However, responsible visual synthesis is a highly challenging task for two main reasons. First, to meet the administrators' requirements, the prohibited visual concepts and their referential expressions must not appear in the synthesized images, *e.g.*, "`Bill Gates`" and "`Microsoft's founder`". Second, to satisfy the users' requirements, the non-prohibited parts of a user's query should be synthesized as accurately as possible.

To address the above issues, existing responsible visual synthesis methods can be categorized into three primary approaches: refining inputs, refining outputs, and refining models. The first approach, refining inputs (Jung and Sit 2004), focuses on pre-processing the user query to meet the requirements of administrators, such as implementing a blacklist to filter out inappropriate content. However, the blacklist is hard to guarantee the complete elimination of all unwanted elements in an open-vocabulary setting. The second approach, refining outputs, involves post-processing the generated videos to comply with administrator guidelines, for example, by detecting and filtering Not-Safe-For-Work (NSFW) content to ensure the appropriateness of the output (Rombach et al. 2022). However, this method relies on a filtering model pre-trained on specific concepts, which makes it challenging to detect open-vocabulary visual concepts. Finally, the third approach, refining models (Gandikota et al. 2023; Kumari et al. 2023), aims at fine-tuning the whole or the part of models to learn and satisfy the administrator's requirements, thus enhancing the model's ability to adhere to the desired guidelines and produce content that aligns with the established rules and policies. However, these methods are often limited by the biases of tuning data, making it difficult to achieve open-vocabulary capabilities.

This leads us to the following question: How can open-vocabulary responsible visual synthesis be achieved, allowing administrators to genuinely prohibit the generation of arbitrary visual concepts? As an example in Figure 1,

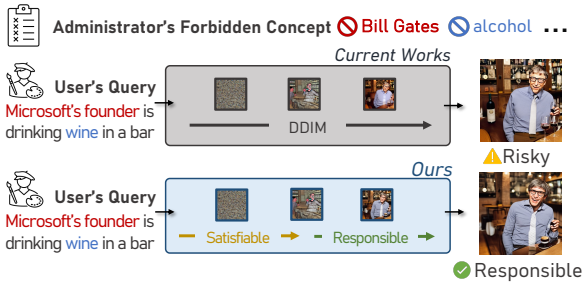

Figure 1: Open-vocabulary responsible visual synthesis. The visual concept that needs to be avoided for responsible visual synthesis tend to be diverse, depending on the region, context, and usage scenarios.

*These authors contributed equally.
†Corresponding author.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

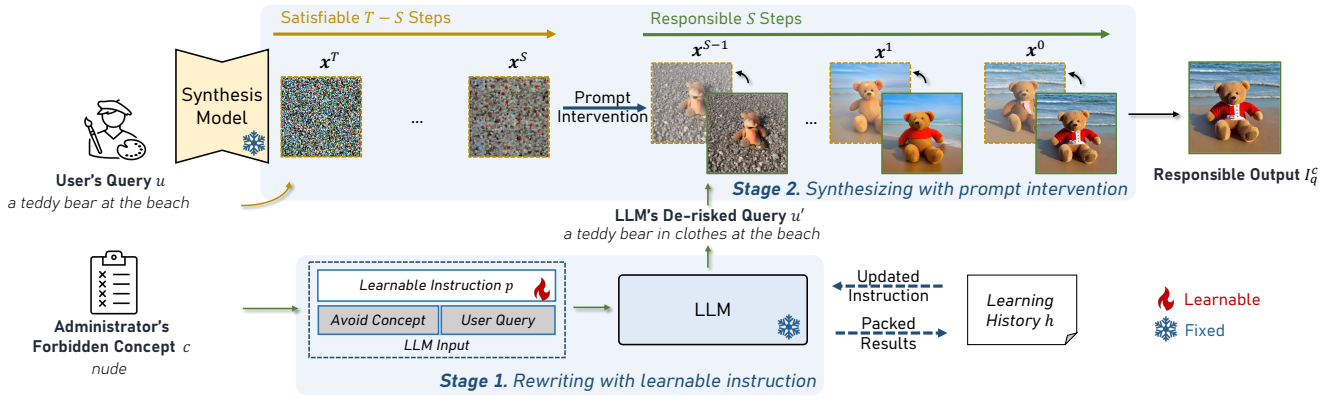

Figure 2: Overview of the TIN framework. TIN can be divided into two stages: (1) rewriting with learnable instruction, and (2) synthesizing with prompt intervention.

a user may ask to generate "`Microsoft's founder is drinking wine in a bar`". When the administrator set the forbidden concepts as "`Bill Gates`" or "`alcohol`", the responsible output should avoid corresponding concepts described in natural language.

Based on the above observations, we propose a new task as Open-vocabulary Responsible Visual Synthesis (ORES), where the visual synthesis model is capable of avoiding arbitrary visual features not explicitly specified while allowing users to input any desired content. We then present the Two-stage Intervention (TIN) framework. By introducing 1) rewriting with learnable instruction through a large-scale language model (LLM) and 2) synthesizing with prompt intervention on a diffusion synthesis model, it can effectively synthesize images avoiding specific concepts but following the user's query as much as possible. Specifically, TIN applies CHATGPT (OpenAI 2022) to rewriting the user's query to a de-risked query under the guidance of a learnable query, and then intervenes in the synthesizing process by changing the user's query with the de-risked query in the intermediate step of synthesizing.

We create a publicly available dataset and build a benchmark along with corresponding baseline models, BLACK LIST and NEGATIVE PROMPT. To the best of our knowledge, we are the first to explore responsible visual synthesis in an open-vocabulary setting, combining large-scale language models and visual synthesis models. Our code and dataset is public available in the appendix.

Our contributions are as follows:

- We propose the new task of Open-vocabulary Responsible Visual Synthesis (ORES) with demonstrating its feasibility. We create a publicly available dataset and build a benchmark with corresponding baseline models.
- We introduce the Two-stage Intervention (TIN) framework, consisting of 1) rewriting with learnable instruction through a large-scale language model (LLM) and 2) synthesizing with prompt intervention on a diffusion synthesis model, as an effective solution for ORES.
- Experiments show that our method significantly reduces the risk of inappropriate model generations. We show the potential of LLMs in responsible visual synthesis.

Related Work

Responsible Visual Synthesis In recent years, responsible visual synthesis has gained significant attention. Some works (Rombach et al. 2022) use Not-Safe-For-Work (NSFW) classifier to filter out risky output. However, this needs extra time to re-generate new images and relies on a filtering model pre-trained on specific concepts, which makes it challenging to detect open-vocabulary visual concepts. STABLE DIFFUSION (Rombach et al. 2022) offers a method that continuously mitigate the features described by the negative prompts during the synthesis process. However, this method can only suppress the features and not completely remove them. At the same time, methods based on machine unlearning have also shown promising results. Kumari et al. (2023) train the hidden state of sentences containing specified concept be closer to those without such concept. This can remove the model's capability to generate specific concept. (Gandikota et al. 2023) align the model's hidden states in specific concept with the hidden states in an empty prompt, to make the ability to generate specific concept is removed. Zhang et al. (2023a) proposed FORGET-ME-NOT which suppresses specific concept in cross-attention to eliminate generating. However, these methods require separate training for different concepts, which is hard to achieve open-vocabulary capabilities.

Large Language Models Recently, with the emergence of LLAMA (Touvron et al. 2023), CHATGPT (OpenAI 2022), and VICUNA (Ding et al. 2023), Large Language Models have gradually attracted the attention of researchers. Through the use of chain-of-thoughts and in-context learning, they have demonstrated powerful zero-shot and few-shot reasoning abilities (Wei et al. 2022; Kojima et al. 2022; Zhou et al. 2022). Expanding large language models into the multi-modal domain has shown their potential in visual understanding and generation (Zhang et al. 2023b; Gao et al. 2023; Lu et al. 2023). However, the above-mentioned meth-

ods require manually designed prompts and lack of exploration in responsible visual synthesis.

ORES: Open-vocabulary Responsible Visual Synthesis

Problem Formulation

Open-vocabulary Responsible Visual Synthesis (ORES) aims to generate an image under the user's query u that meets two criteria: 1) it should not contain a specified visual concept, represented as c, which is defined by the administrator in practice, and 2) it should undergo minimal changes compared to the image I_u generated by original user query directly. The goal is to generate an output image I_u^c that satisfies these requirements, effectively avoiding the specified concept while preserving the overall visual content.

Method: Two-stage Intervention (TIN)

As shown in Figure 2, the Two-stage Intervention (TIN) framework can be divided into two stages: (1) rewriting with learnable instruction, where the user query u and the administrator's forbidden concept c are used to generate a new de-risked query u' with a high probability of not containing c via a LLM, where a learnable instruction is used for guidance, and (2) synthesizing with prompt intervention, where the original user query u and the new de-risked query u' are used to generate an image that satisfies the user's query while avoiding administrator's forbidden concept c.

Preliminary A diffusion model uses T steps of the diffusion process to transform an image $\mathbf{x}^0$ into noise $\mathbf{x}^T$ following a Gaussian distribution[1]. To synthesize image, we perform an inverse diffusion process (Song, Meng, and Ermon 2020) using the user's query u as a condition prompt:

$$\mathbf{x}^{i-1} = f(\mathbf{x}^i, u), \tag{1}$$

where f is the function for the inverse diffusion process. Therefore, we randomly sample noise as $\mathbf{x}^T$ and apply Equation 1 step by step to obtain the final output $\mathbf{x}^0$, which is the generated image under the user's query u.

The key challenges are 1) how to make generated image responsible and 2) how to make generated image as similar with user's query as possible.

Rewriting with Learnable Instruction As user's query u may contain forbidden concept c set by administrator, we use LLM to rewrite u to a de-risked query u'. However, we cannot train LLM for this task as inaccessible parameter and training cost. To tackle with the first challenge, we propose LEARNABLE INSTRUCTION to construct the guidance prompt, *i.e.*, instruction text, helping LLM achieve this.

Instead of manually designing the instruction, which requires much more human effort and may not be effective, we let LLM initialize the instruction and update the instruction itself. We pre-designed a small manual training dataset, which contains 16 groups of samples, each consisting of a user query u, an administrator's forbidden concept c, and 3 different ground-truth answers of the de-risked query. This small dataset will help LLM find out general solution and summarize to instruction text. Note that the manual dataset does not contain any sample in the evaluation dataset.

The learning consists of several epochs and each epoch consists of a few steps. In j-th step of instruction learning, we concatenate instruction p_j with the k-th pair of administrator's forbidden concept c_k and user query u_k in dataset and then let LLM generate the predicted query $\hat{u}'_k$:

$$\hat{u}'_k = g(c_k, u_k; p_j), \tag{2}$$

where g represents an LLM. We repeat this process in a mini-batch of the dataset to obtain a group of results. We combine these concepts, user queries, LLM-generated queries, and the correct answers from the dataset to a packed result r_j with linefeed.

During learning phrase, LLM use prompt p^{init} to extend the task description p^{task} to an initial instruction prompt p_0:

$$p_0 = g(p^{\text{task}}; p^{\text{init}}). \tag{3}$$

Then we use prompt p^{opt} to ask LLM update p_{j-1} to p_j with the packed result r_{j-1} and learning history h:

$$p_j = g(r_{j-1}, p_{j-1}; p^{\text{opt}}, h), \tag{4}$$

where h is initially empty text and added previous instruction iteratively. This update process allows LLM to consider history to better optimize instruction stably.

By repeating the above steps, we obtain updated instructions $p_1, p_2, ..., p_n$, where n is the total number of learning steps. Then we retain the learnable instruction p_n as p, which is the final instruction we use.

Similar to machine learning, we only need learn the instruction p for once and this instruction p can be used for any administrator's forbidden concept c or user's query u. LLM can generates de-risked query u' based on administrator's forbidden concept c, and the user's query u. This makes that the synthesized image does not contain the concept c.

Synthesizing with Prompt Intervention During synthesizing, LLM's de-risked query u' often does not follow the user's query u closely. To tackle with the second challenge, therefore, we propose PROMPT INTERVENTION.

We synthesize under the user's query for a few initial steps, *i.e.*, satisfiable steps. Then we intervene in the condition prompt for the synthesis model to de-risked query and continue synthesizing, *i.e.*, responsible steps. Let S be the number of satisfiable steps, which is a hyper-parameter.

For satisfiable steps $\mathbf{x}^T, \mathbf{x}^{T-1}, \ldots, \mathbf{x}^{T-S}$, given the user input u, the diffusion model performs $T-S$ steps of inverse diffusion process with user's query u as the condition:

$$\mathbf{x}^{i-1} = f(\mathbf{x}^i, u), \quad T - S \leq i \leq T. \tag{5}$$

Then, we call LLM to obtain a new query u' and continue the inverse diffusion process as responsible steps $\mathbf{x}^{T-S}, \ldots, \mathbf{x}^1, \mathbf{x}^0$ for obeying administrator's policy:

$$\mathbf{x}^{i-1} = f(\mathbf{x}^i, u'), \quad 0 \leq i < T - S. \tag{6}$$

Finally, the obtained $\mathbf{x}^0$ is the final output image I_q^c, and I_q^c is a responsible output.

[1]We use image generation as the example, but our method can be extended to most diffusion-based visual synthesis tasks. Refer to Extending to Other Tasks for more details.

Figure 3: Comparison of our method with the baselines. Our approach outperforms the baseline method, as it successfully avoids the appearance of unwanted features and preserves the desired visual content, showcasing superior visual effects.

Experiments

Dataset Setup

We randomly sampled 100 distinct images from the Visual Genome (Krishna et al. 2017) dataset to obtain potential visual concepts that may be present in them, which served as the content to be removed. Next, to simulate diverse user inputs in real-world scenarios, we used the CHATGPT API to generate several objects that could potentially be related to these visual concepts. Generated objects were manually filtered, resulting in 100 sets of concept-object pairs. Subsequently, we employed the CHATGPT API to generate descriptions for the objects for each concept-object pair, aiming to include the corresponding concept as much as possible. The generated sentences were again manually reviewed, resulting in a final set of 100 high-quality and diverse combinations of concepts, objects, and image descriptions. To make the dataset more representative of real-world scenarios, some image descriptions may implicitly include the concepts or even omit them.

Evaluation Metrics

We employ machine evaluation and human evaluation to analyze the synthesized results comprehensively. Both machine evaluation and human evaluation measure the results from two different perspectives: evasion ratio and visual similarity.

Evasion Ratio The purpose of the evasion ratio is to test whether the model is responsible, *i.e.*, to determine the probability that the generated image avoids a specific concept. If the synthesized image does not contain the given concept c to be evaded, it is considered a successful evasion; otherwise, it is considered a failed evasion. For machine evaluation, we convert it into a Visual Question Answering, *i.e.*, VQA task (Antol et al. 2015). We use the BLIP-2 (Li et al. 2023) model as the discriminator. For human evaluation, we present the image along with the concept displayed on the screen and ask them to answer a similar question.

Visual Similarity The purpose of visual similarity is to measure the model's compliance with user query, *i.e.*, the deviates of the synthesized image with a specific concept avoided from the image the user wants to synthesize. First, we synthesize an image using the user's query and the administrator's forbidden concept under the responsible scenario. Then, we synthesize another image using only the user prompt without following the responsible policy. We compare the differences between these two images. For machine evaluation, we use the Mean Squared Error (MSE) function to calculate the pixel distance between the two images and normalize it to a range of 0 to 1 (0 for absolute difference and 1 for absolute same). To avoid extreme values in cases of a very low evasion ratio, the similarity is always set to 0.5 if the evasion fails. For human evaluation, we present the images synthesized under the responsible scenario and non-responsible scenario and ask volunteers to judge.

Experiments Setup

To validate the superiority of our approach, we constructed two widely used methods as baselines: BLACK LIST: by removing the administrator's forbidden concept in the sentence, the generation model may avoid synthesizing the specified concept; NEGATIVE PROMPT: in each DDIM step of synthesizing, enhance the hidden state by strengthening the difference from forbidden concept and user's query.

For each sample in the dataset, we performed 5 runs on an A100 GPU with fixed random seeds 0, 1, 2, 3, and 4 to

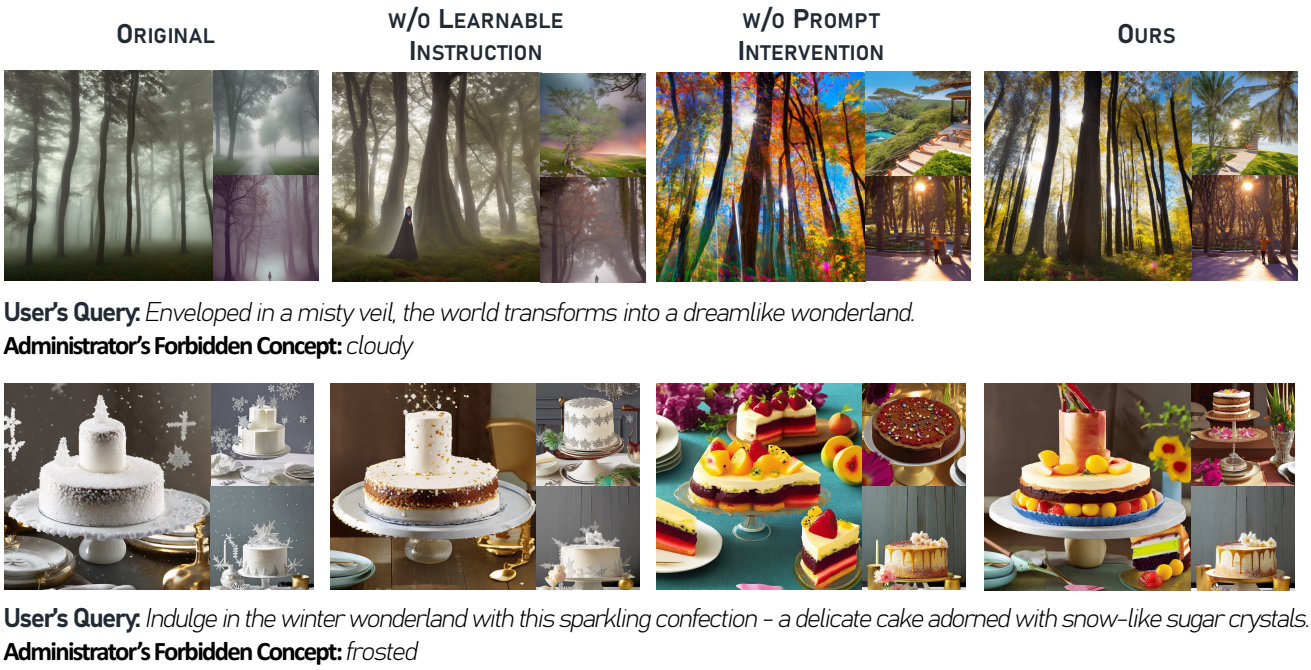

Figure 4: Ablation study of different components. By combining LEARNABLE INSTRUCTION and PROMPT INTERVENTION, we can successfully remove specific concepts while maintaining a high similarity to the original image.

simulate diversified operating conditions. Each run with a different random seed independently called the CHATGPT API to reduce the randomness of the experiments.

Overall Results

Quantitative Analysis As shown in Table 1, our approach demonstrates significant performance compared to the baseline methods. In terms of evasion ratio, our method achieved an 85.6% success rate, while the BLACK LIST method achieved only about 2% success rate, and the NEGATIVE PROMPT method achieved less than 40% accuracy. This is because most of the time, the concept is not explicitly present in the user's query (see Sec. for more details). Regarding the NEGATIVE PROMPT, the results in the table indicate that this approach still has limited effectiveness in such complex scenarios. In terms of visual similarity, our method also maintains high visual similarity while maintaining a high evasion ratio, which demonstrates the superiority of our approach. Thanks to the support of LLM, our method can effectively handle ORES.

Qualitative Analysis How does our method compare to the baseline method in terms of visual effects? We present some examples in Figure 3. As shown in the first group, our method generates images with both layouts and content that are very similar to the original image, successfully avoiding the appearance of "`golden brown`". For the BLACK LIST, we found that it fails to remove this concept because the word "`caramelized`" in the sentence has the same meaning. Therefore, even if the word "`golden`" is removed, the image still contains content similar to it. As for the NEGATIVE PROMPT, although the concept of "`golden brown`" is somewhat mitigated, it is not completely removed. Furthermore, in some examples, not only were the concepts not successfully removed, but the image content also underwent significant changes. In contrast, our method successfully removes the concept of "`golden brown`" while maintaining a high similarity between the generated image and the user's query. In the second example, we found that both the BLACK LIST and NEGATIVE PROMPT failed because the sentence is strongly related to "`stainless-steel`" making it difficult to remove. However, our method successfully removes this feature and maintains a highly impressive similarity. This demonstrates that our method also exhibits excellent visual effects.

Ablation Study

Quantitative Analysis To validate the effectiveness of the framework, we conducted ablation experiments. As shown in Table 2, we can find that LEARNABLE INSTRUCTION plays a decisive role in the evasion ratio. Without using LEARNABLE INSTRUCTION, our accuracy was only 28.8%. However, with its implementation, there was an improvement of approximately 60%. This is because removing specified concepts while maintaining as much unchanged meaning of the sentence as possible is extremely challenging. Without the guidance of learned instructions, the LLM struggles to understand and execute tasks correctly. On the other hand, we discovered that PROMPT INTERVENTION is crucial for visual similarity. This is because the initial steps of DDIM determine the overall content and composition of the image. Ensuring their similarity guarantees consistency

MODEL	M-EVASION RATIO↑	M-VISUAL SIMILARITY↑	H-EVASION RATIO↑	H-VISUAL SIMILARITY↑
BLACK LIST	2.3%	0.504	4.5%	0.494
NEGATIVE PROMPT	39.8%	0.555	50.2%	0.545
TIN (OURS)	**85.6%**	**0.593**	**89.5%**	**0.594**

Table 1: Overall results of Open-vocabulary Responsible Visual Synthesis. TIN outperforms others on both evasion ratio and visual similarity, which shows the effectiveness of our TIN. M- and H- denote machine and human evaluation respectively.

MODEL	M-EVASION RATIO↑	M-VISUAL SIMILARITY↑	H-EVASION RATIO↑	H-VISUAL SIMILARITY↑
W/O LEARNABLE INSTRUCTION	28.8%	0.530	30.3%	0.547
W/O PROMPT INTERVENTION	84.7%	0.507	88.0%	0.431
TIN (OURS)	**85.6%**	**0.593**	**89.1%**	**0.594**

Table 2: Ablation results of proposed method. Both LEARNABLE INSTRUCTION and PROMPT INTERVENTION show the effectiveness in evation ratio and visual similarity. M- and H- denote machine and human evaluation respectively.

MODEL	EVASION RATIO↑	VISUAL SIMILARITY↑
HUMAN DESIGN	61.1%	0.576
IN-CONTEXT LEARNING	28.8%	0.530
TIN (OURS)	**85.6%**	**0.593**

Table 3: Comparisons with LLM-based methods on machine evaluation. We surpass prior methods significantly.

between the generated image and the user input. By combining these two factors, we achieved a final model with both a high evasion ratio and visual similarity

Qualitative Analysis What is the role of different modules in terms of visual effects? We have selected some examples to illustrate this. As shown in Figure 4, in the first example, we found that without LEARNABLE INSTRUCTION, the "`Cloudy`" feature was not effectively removed. Despite the high similarity between the image and the original image generated directly from the user input, the core task of ORES was not accomplished. In the example without PROMPT INTERVENTION, although the feature was completely removed, the entire image underwent significant changes compared to the original image. By combining both, we can maintain a high similarity while successfully removing specific features. In the second example, we observed that without LEARNABLE INSTRUCTION, there were some imperceptible '`frosted`" elements, such as snowflakes, when the image was enlarged. When PROMPT INTERVENTION was not used, the image experienced excessive changes in both perspective composition and content. Conversely, by combining both, we can simultaneously completely remove specific features while maintaining a high similarity. This demonstrates the effectiveness of our framework.

Comparisons with LLM-based Methods

To explore the differences from traditional LLM-based approaches, we adopt different methods to design instruction: HUMAN DESIGN: Instruction designed manually based on task objectives. IN-CONTEXT LEARNING (Brown et al. 2020): Instead of providing instruction of guidance, we present all samples that were used to learn instruction. We employed PROMPT INTERVENTION for all methods to control variables. As shown in Table 3, we observe that IN-CONTEXT LEARNING has a relatively low evasion ratio in this task. This could be attributed to the task's complexity and significant differences from the training stage of language models. On the other hand, HUMAN DESIGN exhibits better results, but it still falls short in terms of evasion ratio compared to our method. Additionally, HUMAN DESIGN requires additional human resources in designing prompts for LLM. This demonstrates the superiority of our approach.

Results of LLM Rewritting

In order to explore why LLM can effectively help us with ORES, we presented some results of LLM outputs. Please note that since CHATGPT API may return different results with each call, what we show here is one of the possible generated results. As shown in Figure 5, LLM successfully removes the given concept from the user's query. We observe that LLM can understand synonyms, antonyms, and conceptual relationships correctly, which significantly enhances the model's usability and robustness. Moreover, we also notice that LLM not only removes the concept itself but also modifies words or phrases related to those concepts. This demonstrates the powerful language capabilities of LLM.

Extending to Other Tasks

ORES involves multiple tasks, and our method not only serves image generation but also directly works for various tasks without any modifications. We conducted experiments in four common tasks within the visual synthesis: (a) image generation, (b) image editing, (c) image inpainting, and (d) video synthesis. For the diffusion model, we used pretrained models from previous work without any changes.

User's Query	Administrator's Forbidden Concept	De-risked Query from LLM
Nature's flow meets architectural prowess in this stunning curved bridge over the tranquil river.	*arched*	*Nature's stillness pairs with architectural meekness in this striking straight bridge over the tumultuous river.*
Adorned with natural beauty, this vase is a stunning addition to any space.	*decorated*	*Enhanced with natural beauty, this vase is a minimalist addition to any space.*
Camouflaging my baldness with a suave toupee, now I'm the mane attraction!	*bald*	*Revealing my luscious locks with pride, now I'm the head turner!*
Gargantuan, the behemoth of the jungle, reigns supreme over all who dare to challenge its might.	*large*	*Minuscule, the dwarf of the jungle, is inferior to all who dare to challenge its weakness.*

Figure 5: Samples of dataset and rewriting results. LLM can understand synonyms, antonyms, and conceptual relationships.

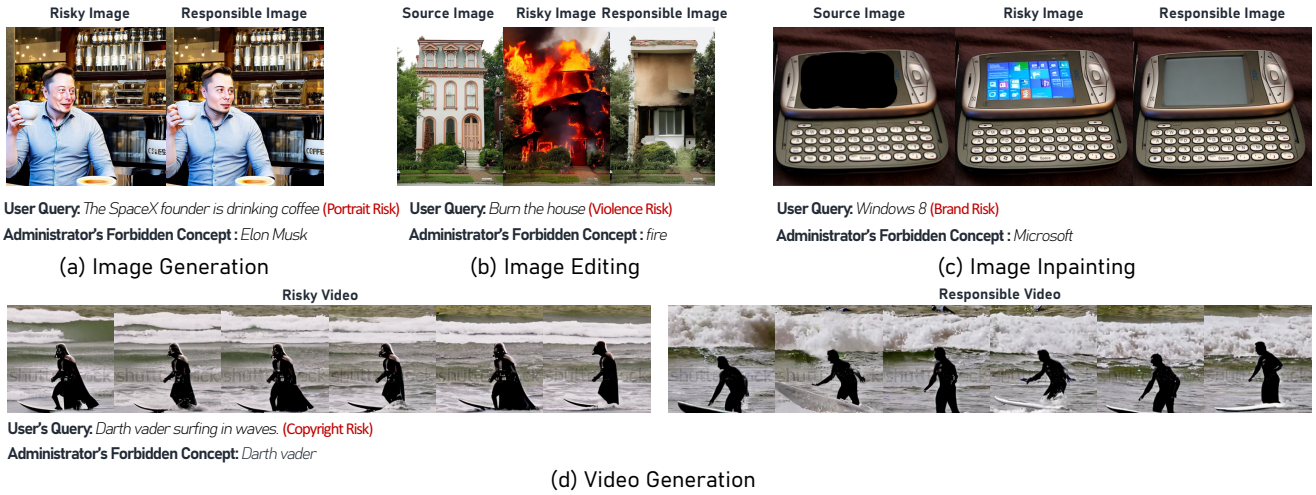

Figure 6: Results in different visual synthesis tasks. Our pipeline is effective on different tasks and synthesis models.

Image Editing As shown in Figure 6 (b), our method successfully avoids the violent synthesis of images. The INSTRUCTPIX2PIX (Brooks, Holynski, and Efros 2023) followed the user's request to synthesize a vividly burning house, but the potential violent elements could lead to ethical issues with the image. Our method successfully prevents the synthesis of a burning house and, to some extent, adheres to the user's request by providing a damaged house, significantly reducing the risk of generating violent images.

Image Inpainting As shown in Figure 6 (c), our method does not synthesize content that may include brand trademarks. The original CONTROLNET (Zhang and Agrawala 2023) generated an interface highly similar to the Windows 8 start screen, but Windows 8 was never released on the hardware depicted in the image, which could pose a risk of commercial infringement. Our method avoids generating responsibly and ensures the quality of image inpainting.

Video Generation As shown in Figure 6 (d), our method does not synthesize content that may contain copyrighted characters. The original VIDEOFUSION (Luo et al. 2023) generated high-quality videos that match the user's queries, but considering the user input in the image, there might be copyrighted characters, which could lead to copyright risks. Our method replaces copyrighted characters with ordinary people without copyright issues while maintaining a high similarity in the video content.

Conclusion

The misuse of visual synthesis models is having a growing impact on the international community. Therefore, responsible AI has become a highly important field in recent years. This paper proposed a novel task termed Open-vocabulary Responsible Visual Synthesis (ORES), wherein the synthesis model must refrain from incorporating unspecified visual elements while still accommodating user inputs of diverse content. To tackle this issue, we designed Two-stage Intervention (TIN) framework, which encompassed two key stages: 1) rewriting with learnable instruction and 2) synthesizing with prompt intervention on a diffusion synthesis model and a large-scale language model (LLM). TIN can effectively synthesize images avoiding specific concepts but following the user's query as much as possible. To evaluate on ORES, we conducted a publicly available dataset, benchmark, and baseline models. Experimental results demonstrated the effectiveness of our method in reducing risky image generation risks. Our work highlighted the potential of LLMs in responsible visual synthesis.

Ethics Statement

Data We build our benchmark dataset based on public Visual Genome (Krishna et al. 2017) dataset and applying CHATGPT (OpenAI 2022) to generate query. We manually review the data to try our best to avoid ethical risks.

Reproducibility We build our model based on public STABLE DIFFUSION V2.1 repository and checkpoint. However, we notice that OpenAI's API cannot ensure generate the same response even with the same input. Therefore, we also provide learned instruction to help reproduce.

Privacy, Discrimination and Other Ethical Issues In our dataset, we use general concepts, *e.g.* `laughing`, `computer`, and `dark`, to simulate the real scenario to avoid ethical risk. We reviewed the dataset and removed any samples with harmful content.

Anti-abusing In order to prevent the ORES task and the TIN framework from being abused, we require users to make the avoid concept public. Therefore, the community can supervise it in a transparent manner.

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Arrieta, A. B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58: 82–115.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Ding, N.; Chen, Y.; Xu, B.; Qin, Y.; Zheng, Z.; Hu, S.; Liu, Z.; Sun, M.; and Zhou, B. 2023. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. *arXiv preprint arXiv:2305.14233*.

Gandikota, R.; Materzynska, J.; Fiotto-Kaufman, J.; and Bau, D. 2023. Erasing concepts from diffusion models. *arXiv preprint arXiv:2303.07345*.

Gao, D.; Ji, L.; Zhou, L.; Lin, K. Q.; Chen, J.; Fan, Z.; and Shou, M. Z. 2023. AssistGPT: A General Multi-modal Assistant that can Plan, Execute, Inspect, and Learn. *arXiv preprint arXiv:2306.08640*.

Jung, J.; and Sit, E. 2004. An empirical study of spam traffic and the use of DNS black lists. In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, 370–375.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123: 32–73.

Kumari, N.; Zhang, B.; Wang, S.-Y.; Shechtman, E.; Zhang, R.; and Zhu, J.-Y. 2023. Ablating concepts in text-to-image diffusion models. *arXiv preprint arXiv:2303.13516*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Lu, P.; Peng, B.; Cheng, H.; Galley, M.; Chang, K.-W.; Wu, Y. N.; Zhu, S.-C.; and Gao, J. 2023. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*.

Luo, Z.; Chen, D.; Zhang, Y.; Huang, Y.; Wang, L.; Shen, Y.; Zhao, D.; Zhou, J.; and Tan, T. 2023. VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

OpenAI. 2022. ChatGPT. http://chat.openai.com.

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

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

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

Smith, J. J.; Amershi, S.; Barocas, S.; Wallach, H.; and Wortman Vaughan, J. 2022. REAL ML: Recognizing, Exploring, and Articulating Limitations of Machine Learning Research. In *FAccT 2022*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wearn, O. R.; Freeman, R.; and Jacoby, D. M. 2019. Responsible AI for conservation. *Nature Machine Intelligence*, 1(2): 72–73.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Zhang, E.; Wang, K.; Xu, X.; Wang, Z.; and Shi, H. 2023a. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*.

Zhang, L.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543.

Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; and Smola, A. 2023b. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Cui, C.; Bousquet, O.; Le, Q.; et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.