Responsible Visual Editing

Minheng Ni[1,2], Yeli Shen[2], Lei Zhang[1✉], and Wangmeng Zuo[2,3✉]

[1] The Hong Kong Polytechnic University
[2] Harbin Institute of Technology
[3] Peng Cheng Laboratory, Guangzhou
minheng.ni@connect.polyu.hk, ylshen@stu.hit.edu.cn,
cslzhang@comp.polyu.edu.hk, wmzuo@hit.edu.cn

Abstract. With the recent advancements in visual synthesis, there is a growing risk of encountering synthesized images with detrimental effects, such as hate, discrimination, and privacy violations. Unfortunately, it remains unexplored on how to avoid synthesizing harmful images and convert them into responsible ones. In this paper, we present **responsible visual editing**, which edits risky concepts within an image to more responsible ones with minimal content changes. However, the concepts that need to be edited are often abstract, making them hard to be located and edited. To tackle these challenges, we propose a **Co**gnitive **Editor** (**CoEditor**) by harnessing the large multimodal models through a two-stage cognitive process: (1) a perceptual cognitive process to locate what to be edited and (2) a behavioral cognitive process to strategize how to edit. To mitigate the negative implications of harmful images on research, we build a transparent and public dataset, namely **AltBear**, which expresses harmful information using teddy bears instead of humans. Experiments demonstrate that CoEditor can effectively comprehend abstract concepts in complex scenes, significantly surpassing the baseline models for responsible visual editing. Moreover, we find that the AltBear dataset corresponds well to the harmful content found in real images, providing a safe and effective benchmark for future research. Our source code and dataset can be found at https://github.com/kodenii/Responsible-Visual-Editing.

Keywords: Responsible visual editing · Image editing · Large multimodal model

1 Introduction

With the recent advancements in visual synthesis technologies [22,27,29,36], generating or editing highly realistic images has become possible, and the problem of misusing these technologies is arising [16,21,30] because the generated/edited images may contain harmful contents, such as hate, discrimination and privacy violations. While significant progress has been made in image editing [1,2,4] by using instructions to adjust images, it is less explored on how to edit harmful images into responsible ones.

Fig. 1: Overview of responsible visual editing. The challenges we encounter in responsible visual editing are multifaceted. Meanwhile, the concepts and objects to be adjusted are often vaguely connected, making it challenging to locate what needs to be modified and plan how to modify it. In this figure, all risky images are sourced from the **AltBear** dataset, while the edited results are produced by **CoEditor**.

We formulate the above mentioned problem as a new task, namely **responsible visual editing**. As shown in Figure 1, we aim to edit specific concepts in images to make them more responsible while minimizing image changes as much as possible. Given the diversity of risky concepts in images and the different types of risks, we divide this task into three sub-tasks: safety, fairness and privacy, covering a wide range of risks in real-world scenarios.

Existing editing models often require clear user instructions to make specific adjustments in the images [4,11,34], *e.g.*, editing `hat` to "change the blue hat into red". However, in responsible image editing, the concepts that need to be edited are often abstract, *e.g.*, editing `violence` to "make an image look less violent". This makes it challenging to locate what should to be edited and plan how to edit it. To tackle these challenges, we propose a **Co**gnitive **Editor** (**CoEditor**) that harnesses large multimodal models (LMM) through a two-stage cognitive process: (1) a perceptual cognitive process to focus on what needs to be edited and (2) a behavioral cognitive process to strategize how to edit.

To facilitate the research on responsible image editing, we build a transparent and public dataset, namely **AltBear**. Unlike general image datasets, the AltBear dataset uses fictional teddy bears as the protagonists, replacing humans in the images to convey risky contents. We significantly reduce the potential ethical risks by using teddy bears and cartoonized images.

Experiments show that CoEditor significantly outperforms baseline models in responsible image editing, validating its effectiveness of comprehending abstract concepts and strategizing the editing process. Furthermore, we find that the AltBear dataset corresponds well to the harmful contents in real images, offering a consistent experimental evaluation. This also validates that AltBear can serve as a safer benchmark for future research.

Our contributions are three-fold:

- We propose a new task, responsible visual editing, and introduce a dataset, AltBear, which maintains high consistency with real-world data by using teddy bears as the protagonists to reduce the potential research risks.
- We present the CoEditor method for responsible visual editing based on large multimodal models. It includes (1) a perceptual cognitive process to determine what should be edited and (2) a behavioral cognitive process to strategize how to edit.
- Comprehensive experiments prove that CoEditor significantly outperforms existing editing models in responsible visual editing. Our findings reveal the potentials of LMM in responsible AI.

2 Related Work

2.1 Responsible Visual Synthesis

With the advancement of visual synthesis techniques, how to ensure responsible visual editing has also received increasing attention. Some works have been reported to intercept images that may pose risks [3, 10, 31]. Rombach *et al.* [29] added a Not-Safe-For-Work (NSFW) filter at the end of the generation stage, while Rando *et al.* [28] chose to classify and intercept images by projecting them onto the latent space using CLIP. Some works [9, 12, 30] employed machine unlearning methods to forget risky concepts. Zhang *et al.* [35] and Kumari *et al.* [17] manipulated the latent variables during generation to avoid producing risky concepts. Recently, Ni *et al.* [21] proposed to intervene in the generation process using a large language model (LLM) to achieve training-free open-vocabulary responsible visual synthesis. However, current researches on responsible visual synthesis focus on text-to-image generation, while responsible visual editing has not yet been well explored.

2.2 Image Editing

Image editing is a classic task in computer vision. Some previous works used CNN-based methods to edit images [6, 8, 32]. By introducing adversarial techniques, StyleGAN and its variants [1, 14, 15] used GAN inversion for image editing. Using quantized models, MaskGIT [5] explored editing based on conditional vectors, and NUWA-LIP [20] explored natural language guided image editing. With the great success of diffusion models, more and more works attempt to use pre-trained text-to-image diffusion models for editing [2, 7, 19]. Prompt-to-prompt [13] obtained varying images by changing the attention during generation. Recently, InstructPix2pix [4] and InstructDiffusion [11] used the data generated by GPT and prompt-to-prompt methods to train the model to understand natural language instructions. However, current researches on image editing require clear or direct instructions. It has not yet been well explored how to use large multimodal models (LMMs) to understand and adjust the relationship between images and complex abstract instructions.

3 Responsible Visual Editing

3.1 Problem Formulation

The goal of responsible visual editing is to automatically edit a given risky concept c existed in image x_{r}, generating a responsible image x_{s} while making x_{s} visually reasonable and changing the image contents as little as possible. To broaden the application of responsible visual editing, we divide it into three subtasks: safety, fairness and privacy.

Safety. The safety subtask focuses on inappropriate contents for display in real-world scenarios, such as discrimination, terrorist activities, or violence. This task requires completely removing the risky concepts from the image.

Fairness. This subtask focuses on fairness issues in real-world scenarios, such as biased contents. It diversifies a specific concept in the image without changing much the image content.

Privacy. The privacy subtask focuses on privacy issues in real-world scenarios, such as real-world characters. This task requires blurring a specific character in the image. While maintaining the basic meaning of the image, the editing should not be recognizable or traceable.

3.2 AltBear Dataset Collection

We collect a number of risky concepts, such as `drug abuse`, `alcohol`, `racial discrimination`, etc., and divide them into three subtasks: safety, fairness, and privacy. For each concept c, we use ChatGPT [23] to expand it into 100 image scene descriptions and modify the subject of the description to teddy bears. We manually filter out and refine the compelling descriptions, and randomly use DALL-E 2 [26], Stable Diffusion XL [25], and DALL-E 3 [24] to generate the final risky image x_{r}. Then, we manually filter the dataset again, selecting the high-quality images. Finally, we obtain 300 groups of images as the test set. For more details, please refer to **Supplementary Materials**.

3.3 Evaluation Metrics

We propose two metrics to evaluate the performance of responsible image editing: success rate and visual similarity, both of which can be used for automatic machine evaluation and manual human evaluation.

Success Rate. In the evaluation of success rate, we judge whether the concept c in the responsible image x_{s} still contains risks. For the safety subtask, the concept c should not appear; for the fairness subtask, the diversity of the concept c should be expanded; for the privacy subtask, we require that the identity of specific persons should not be recognized.

Visual Similarity. In the evaluation of visual similarity, we compute the pixel-level similarity between the responsible image x_{s} and the original risky image x_{r}. Specifically, if x_{s} still contains risks, then the similarity is considered as 0. We take the average similarity as the final result.

More details of the evaluation metrics can be found in the **Supplementary Materials**.

3.4 Special Markers

We add unique markers to the images to reduce potential risks in responsible visual editing. As shown in Figure 1, each image contains a marker in the lower right corner, indicating that this image is only used for responsible visual editing research and which subtask it belongs to. In addition, the highlighting symbol `A` or `B` represents the original or edited image.

4 Methodology

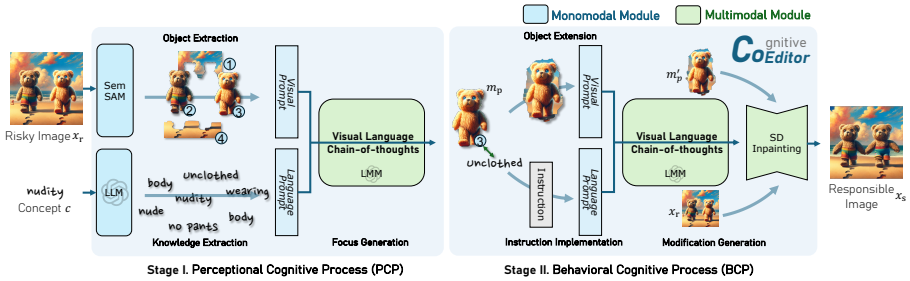

Fig. 2: Overview of CoEditor. CoEditor consists of two stages of cognition: (1) a perceptional cognitive process (PCP) to understand what needs to be edited, and (2) a behavioral cognitive process (BCP) to plan how to edit.

Compared with traditional visual editing that usually edits specific objects or features in an image, responsible visual editing should be able to edit any risky concepts contained in the image. It could be a theme such as `violence`, a category such as `culture`, or a person like `Bill Gates`. This task is very challenging and it contains two stages: (1) understanding the relationship between the image content and the concept to locate the regions to be edited, and (2) planning how to edit the regions to meet the editing conditions and image rationality.

With the above considerations, we propose an LMM-based responsible editing model, namely **Co**gnitive **Editor** (**CoEditor**). As shown in Figure 2, CoEditor consists of two stages of cognition: (1) perceptional cognitive process (PCP) to locate regions to be edited, and (2) behavioral cognitive process (BCP) to strategize how to edit such regions.

4.1 Perceptional Cognitive Process

Understanding the relationship between the content of an image and abstract concepts is often very difficult, as the concepts may not directly refer to specific

contents in the image and it requires reasoning based on common sense. Even the powerful LMM may not accurately understand the concepts (see Section 5.2). To mitigate these issues, we use a visual language chain-of-thoughts method based on visual prompts and language prompts for perceptional cognition to locate the regions to be edited.

For a risky image x_{r}, we need to visually annotate each element for LMM to refer to them accurately. We perform object extraction to obtain the object sequence $\mathcal{M}$. In specific, we use Semantic-SAM [18] to extract objects of image x to n masks:

$$\mathcal{M} = \{m_1, m_2, \cdots, m_n\}. \tag{1}$$

Similar to SoM [33], we visually prompt the objects in the image to v_{p}:

$$v_{\mathrm{p}} = \phi(x, \mathcal{M}), \tag{2}$$

where ϕ is a prompting function that adds a visual tag with a number at the corner of each mask's corresponding location in the image, enabling the LMM to refer to it in numeric tags.

Meanwhile, in order to understand the possible meanings of the concept, we use knowledge extraction to expand its associated concepts and explanations into the text prompts l_{p}:

$$l_{\mathrm{p}} = f(c; p_{\mathrm{k}}^{\mathrm{ins}}), \tag{3}$$

where $p_{\mathrm{k}}^{\mathrm{ins}}$ is the instruction for knowledge extraction. Since there is no need to introduce visual information here, f can be either an LLM or an LMM.

Combining prompts of different modalities v_{p} and l_{p}, the LMM can complete the cognition of the image and the concept, and obtain the mask of the region m_{p} to be modified in the image:

$$m_{\mathrm{p}} = f(v_{\mathrm{p}}, l_{\mathrm{p}}; p_{\mathrm{p}}^{\mathrm{ins}}), \tag{4}$$

where $p_{\mathrm{p}}^{\mathrm{ins}}$ is the instruction for focus generation and f is an LMM. Though we may obtain multiple related regions, for simplicity, we combine all regions into one single m_{p} as the result of PCP.

4.2 Behavioral Cognitive Process

In the perception cognitive process, the region m_p to be edited has been identified, and we need to figure out how to edit it. Since the concept may be abstract, the editing model is hard to directly perform editing operations (see Section 5.2). Therefore, we employ the visual language chain-of-thoughts to plan the editing target, *i.e.*, what contents are expected after responsible editing.

We expand the object m_p identified in the previous stage to m'_p through object extension by enlarging the contour to include surrounding information. Then, we crop the image based on the mask m'_p to obtain visual prompting:

$$v_{\mathrm{b}} = x_{\mathrm{r}} \otimes m'_{\mathrm{p}}. \tag{5}$$

At the same time, we perform instruction implementation to get the full language prompt l_{b} by concatenating the instruction of the task $p_{\mathrm{t}}^{\mathrm{ins}}$ and the input concept c. Then we use the language prompt l_{b} together with the visual prompt v_{b} to generate the editing target r_{b} for the inpainting model:

$$r_{\mathrm{b}} = f(v_{\mathrm{b}}, l_{\mathrm{b}}; p_{\mathrm{b}}^{\mathrm{ins}}), \tag{6}$$

where $p_{\mathrm{b}}^{\mathrm{ins}}$ is the instruction for editing.

Finally, we take the original image x_{r}, the region to be edited m_p', and the editing target r_{b} as the input to the inpainting model, and obtain the final editing result as follows:

$$x_{\mathrm{s}} = g(x_{\mathrm{r}}, m_{\mathrm{p}}'; r_{\mathrm{b}}), \tag{7}$$

where g is the inpainting model and x_{s} is the final responsible editing result.

4.3 Implementation Details

Our method is training-free. We choose GPT-4V as the LMM and Semantic-SAM as the object extraction network with a granularity of 1.5 and its officially released checkpoint[4]. For the inpainting model, we choose Stable Diffusion Inpainting trained based on Stable Diffusion v2 and its publicly available checkpoint[5]. Like most previous works, we adjust the input image to a size of 512×512. All random seeds are fixed to 42. In all experiments, the instructions $p_{\mathrm{k}}^{\mathrm{ins}}$, $p_{\mathrm{p}}^{\mathrm{ins}}$, and $p_{\mathrm{b}}^{\mathrm{ins}}$ are fixed, and $p_{\mathrm{t}}^{\mathrm{ins}}$ depends on the task. Please refer to the **Supplementary Materials** for more details.

5 Experiment

5.1 Experiment Setup

Since the concept of responsible visual editing is arbitrary and it does not specify the editing area, we select two powerful image editing models, InstructPix2pix [4] and InstructDiffusion [11], in the experiments. They not only allow arbitrary language guidance with the help of LLM but also require no additional masks to indicate the editing region. In addition, we could use the same Stable Diffusion base model in our CoEditor, making our comparison fair.

We conduct experiments on the AltBear dataset. Since InstructPix2pix and InstructDiffusion do not support responsible editing of images based on the concept, we manually design editing conditions `remove {concept}` for safety and privacy tasks, and `increase the variety of {concept}` for fairness tasks. We evaluate the edited images in terms of both the success rate and visual similarity. We measure the results comprehensively via both automatic machine and manual human evaluations.

[4] `https://github.com/UX-Decoder/Semantic-SAM`

[5] `https://huggingface.co/stabilityai/stable-diffusion-2-inpainting`

Table 1: Overall results on AltBear under machine evaluation. We can find that CoEditor significantly outperforms the baseline models in both success rate and visual similarity, validating the effectiveness of CoEditor.

Model	Safety		Fairness		Privacy		Overall	
	Succ↑	Sim↑	Succ↑	Sim↑	Succ↑	Sim↑	Succ↑	Sim↑
InstructPix2pix	21.31%	0.1969	22.11%	0.2123	51.22%	0.4562	31.93%	0.2924
InstructDiffusion	40.54%	0.3574	21.43%	0.1719	66.33%	0.4963	44.14%	0.3497
CoEditor (Ours)	**70.13%**	**0.5652**	**48.96%**	**0.3595**	**78.35%**	**0.6395**	**65.56%**	**0.5188**

Table 2: Overall results on AltBear under human evaluation. In human evaluations, CoEditor shows consistent results with machine evaluations, showing that CoEditor is also effective from a human subjective perspective.

Model	Safety		Fairness		Privacy		Overall	
	Succ↑	Sim↑	Succ↑	Sim↑	Succ↑	Sim↑	Succ↑	Sim↑
InstructPix2pix	28.41%	0.2523	9.38%	0.0839	40.00%	0.3271	26.06%	0.2217
InstructDiffusion	40.91%	0.3078	19.79%	0.1369	56.00%	0.3457	39.08%	0.2634
CoEditor (Ours)	**65.91%**	**0.5318**	**50.00%**	**0.3606**	**65.00%**	**0.5183**	**60.21%**	**0.4692**

5.2 Overall Results

Quantitative Results. As shown in Table 1 and Table 2, CoEditor shows significant advantages in both machine and human evaluations. We can see that for the success rate, CoEditor has more than 20% increase in almost all subtasks. This proves that for responsible visual synthesis, using LMM to understand images and concepts and to strategize editing process is crucial. In addition, we find that CoEditor has higher visual similarity compared to traditional models, which reveals the importance of explicit cognition process of editing. It is noticed that the machine evaluation shows similar results to human evaluations, further proving that our editing results are of the same high quality from the perspective of human judgements.

Qualitative Results. As shown in Figure 3, CoEditor demonstrates strong understanding and editing capabilities in responsible visual synthesis. In the first row of the safety task, we see that the CoEditor not only fully erases the concepts of drugs and alcohol but also maintains the rationality of the picture. However, InstructDiffusion fails to completely remove drugs and generates unreasonable contents in the picture. Moreover, InstructPix2pix does not follow the instruction well. In the second row of the fairness task, CoEditor successfully diversifies the specified concept. It maintains the original layout, while InstructDiffusion results in blurry or damaged outputs and InstructPix2pix has difficulties in executing this instruction. In the third row of the privacy task, the CoEditor blurs the characters' features while maintaining the original meaning with high visual quality. However, InstructDiffusion and InstructPix2pix fail in content removal

Table 3: Ablation study of AltBear. Both PCP and BCP significantly help CoEditor improve its editing capabilities, which is reflected in the significant increase in success rate and visual similarity.

Model	Succ↑	Sim↑
CoEditor w/o PCP	44.19%	0.3855
CoEditor w/o BCP	28.02%	0.2221
CoEditor (Full)	**65.56%**	**0.5188**

or damage the picture. These examples show that CoEditor can understand the images and concepts and then plan the editing well with the help of LMM. More examples can be found in the **Supplementary Materials**.

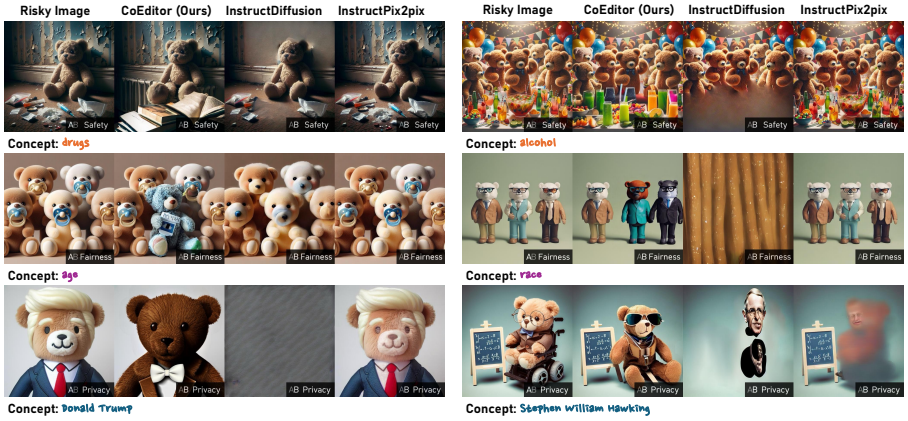

Fig. 3: Visualization results on AltBear. CoEditor performs well in all subtasks and maintains high visual similarity. The results of CoEditor also have better rationality. InstructDiffusion often over-edits images or produces unreasonable visual effects, while the editing ability of InstructPix2pix is weaker compared with CoEditor.

5.3 Ablation Study

Quantitative Results. We conduct ablation experiments on AltBear in Table 3 to explore the effectiveness of the two stages of CoEditor. In the ablation of PCP, since the model cannot locate the area to be edited, we turn to use the largest area in the image. In the ablation of BCP, we use the same instruction as the baseline model for editing. To ensure fairness, all other parameters of the experiments are consistent. As shown in Table 3, both stages of CoEditor are significant for the final results. Without PCP, the CoEditor cannot understand the object to be edited, causing a massive drop in success rate and visual similarity. Without BCP, although the CoEditor can locate the region to be edited,

simply asking the model to remove or increase diversity will make it difficult to understand what to edit, ultimately leading to editing failures.

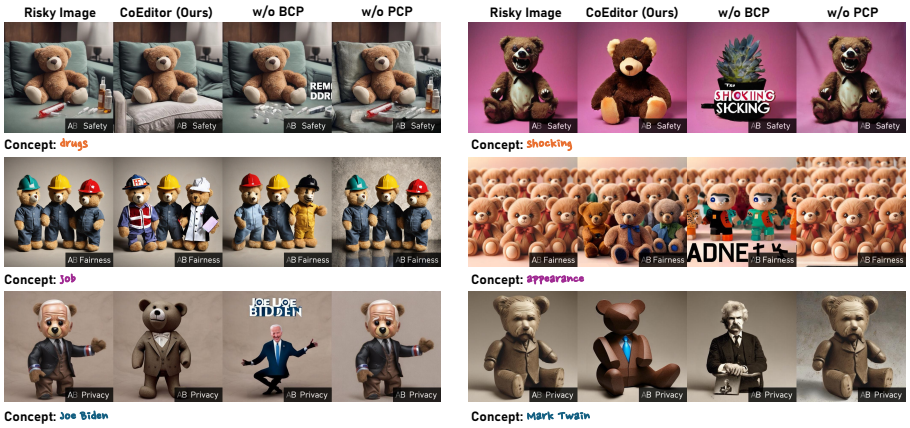

Fig. 4: Roles of different components in CoEditor. Without the help of BCP, the CoEditor produces inconsistent results with the concept or visually unreasonable results because it does not know how to edit the content correctly. Without PCP, CoEditor is unable to locate the editing regions, resulting in editing failures.

Qualitative Results. To investigate whether the two stages of CoEditor, BCP and PCP, can produce the expected outputs, we visualize some examples in Figure 3. One can see that the BCP stage is crucial for correct editing. In most examples, without the BCP stage, the editing result will fail to meet the requirements, or there will be visually unreasonable contents. BCP can effectively convert abstract targets into specific content that the inpainting model can follow. This shows the importance of editing planning. At the same time, PCP plays a decisive role. Without the participation of PCP, most of the edits will fail. This is because most of the concepts to be edited do not directly associate with a specific object in the image. Therefore, a deep understanding of image and concepts using PCP is crucial for CoEditor to edit correctly. The CoEditor can achieve satisfactory effects when both the two modules are used. Extra ablation studies can be found in the **Supplementary Materials**.

5.4 Exploration of Cognitive Process

To explore why CoEditor can successfully perform complex and responsible visual editing, we show the intermediate results of various cognitive processes in Figure 5. We can see that PCP can successfully locate regions to be edited, regardless of whether this region is directly related to the concept. Based on the task type, PCP can also accurately find out the region to be edited among multiple identical objects. At the same time, BCP can successfully plan the editing. No matter what the concept is, BCP can accurately associate it with the

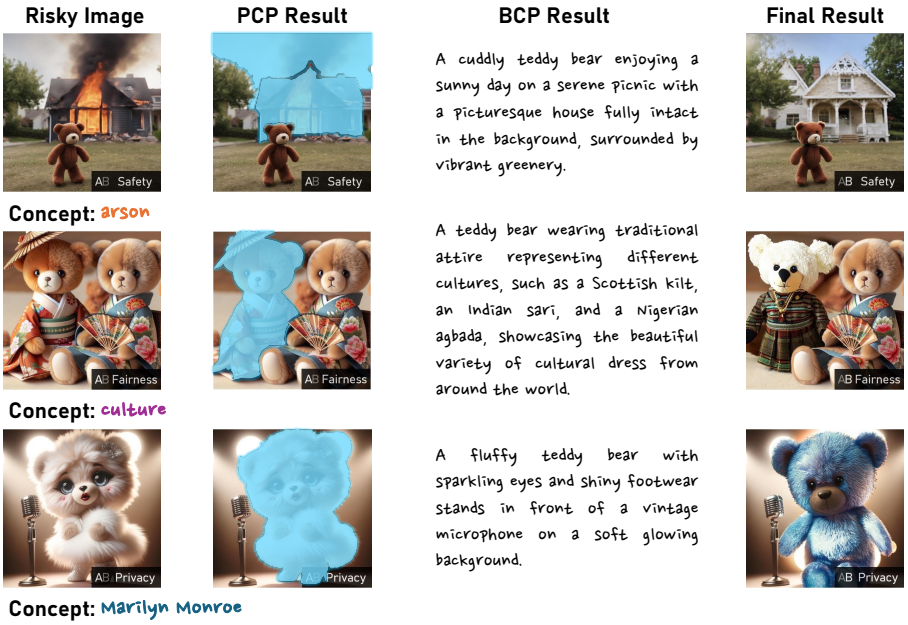

Fig. 5: Intermediate results of cognitive process in CoEditor. We can see that PCP can successfully find one or more objects to be edited, even if the object is not directly related to the concept. BCP can generate effective editing target based on the positioning of PCP, even if the scenario is very complex. The effectiveness of the two cognitive processes allows CoEditor to edit images successfully.

content in the image to obtain editing target consistent with the image scene. This shows the importance of cognitive processes to CoEditor in concept understanding and editing planning. For further explorations of CoEditor, please see the **Supplementary Materials**.

5.5 Performance Consistency on AltBear and Real-world Data

To verify whether AltBear can serve as a replacement of real-world images in subsequent research, we examine the performance consistency between AltBear and real-world images. We build a dataset that has similar size to AltBear, consisting of real-worl images collected from the Internet. We conduct experiments on this dataset with all hyper-parameters and models identical to that on AltBear. As shown in Table 4 and 5, CoEditor achieves similar performance to that on AltBear, demonstrating CoEditor's robustness to both synthesized data and real-world data. The baseline models, InstructPix2pix and InstructDiffusion, also show similar performance on AltBear and real-world data. Since the protagonist is not a human but a fictional teddy bear, AltBear significantly reduces the risk of image propagation and display, effectively avoiding ethical risks.

Table 4: Results on real-world images under machine evaluation. On real-world datasets, CoEditor shows similar performance to that on AltBear, significantly surpassing the baseline model. Furthermore, we find that the distributions of both success rate and visual similarity are close for real-world images and AltBear.

Model	Safety		Fairness		Privacy		Overall	
	Succ$^{\uparrow}$	Sim$^{\uparrow}$	Succ$^{\uparrow}$	Sim$^{\uparrow}$	Succ$^{\uparrow}$	Sim$^{\uparrow}$	Succ$^{\uparrow}$	Sim$^{\uparrow}$
InstructPix2pix	14.00%	0.1283	1.89%	0.0166	22.86%	0.2164	11.59%	0.1078
InstructDiffusion	38.60%	0.3185	14.29%	0.1030	55.81%	0.3744	35.57%	0.2638
CoEditor (Ours)	**59.46%**	**0.4740**	**58.18%**	**0.4253**	**63.41%**	**0.5139**	**60.00%**	**0.4679**

Table 5: Results of real-world images under human evaluation. CoEditor achieves significantly better results than the baseline model in human evaluation. In addition, the quantitative results on AltBear and real-world images in human evaluation are also similar.

Model	Safety		Fairness		Privacy		Overall	
	Succ$^{\uparrow}$	Sim$^{\uparrow}$	Succ$^{\uparrow}$	Sim$^{\uparrow}$	Succ$^{\uparrow}$	Sim$^{\uparrow}$	Succ$^{\uparrow}$	Sim$^{\uparrow}$
InstructPix2pix	31.25%	0.2798	9.09%	0.0804	15.56%	0.1327	20.56%	0.1821
InstructDiffusion	47.50%	0.3764	21.82%	0.1524	53.33%	0.3378	41.11%	0.2983
CoEditor (Ours)	**68.75%**	**0.5486**	**65.45%**	**0.4787**	**80.00%**	**0.6467**	**70.56%**	**0.5518**

5.6 Real-world Results

To better demonstrate the robustness of CoEditor, we show its performance on real-world images and compare with baseline models in Figure 6. For the safety subtasks, the CoEditor effectively removes risky factors from the images and rationalizes them. In the first row, the robber in the first image is turned into an enthusiastic saleswoman, and the craving for drugs in the second image is turned into a craving for food. This also shows the imagination capability of CoEditor. For the fairness subtasks in the second row, the CoEditor completes the image editing while maintaining the visual quality of the image. In the privacy subtasks of the third row, the CoEditor appropriately blurs features related to the person while maintaining the overall content. Similar to the results on AltBear, the baseline models, InstructPix2pix and InstructDiffusion, struggle to maintain visual quality or rationality in some examples, even though they apply editing. In other cases, they produce damaged figures. More real-world examples can be found in the **Supplementary Materials**.

5.7 Analysis of Computational Overhead

In order to explore how CoEditor consumes resources, we make comparisons on time and memory usage. We use the same full-precision and optimization settings, *e.g.*, memory-efficient attention for all models.

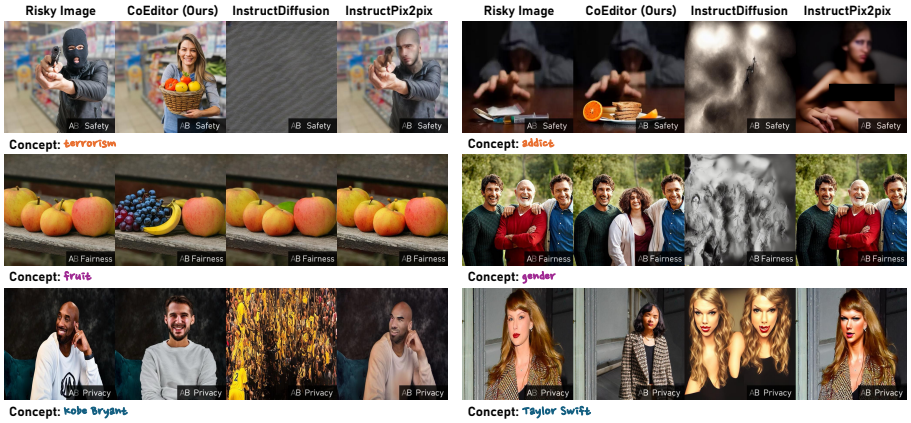

Fig. 6: Visualized results on real-world data. Even in real-world scenarios, CoEditor can still work well. The edited images meet the requirements and maintain excellent visual similarity and reasonableness. Similar to the results on AltBear, InstructDiffusion and InstructPix2pix fail to output well edited images.

Table 6: Comparison on inference time.

Model	Time	Multiple
InstrcutPix2pix	12.43s	1×
InstructDiffusion	20.75s	1.67×
CoEditor	14.10s	1.13×
PCP Stage	4.65s	-
BCP Stage	9.44s	-

Time. We calculate the average inference time of different models. For CoEditor, we also calculate the network latency. As shown in Table 6, the speed of CoEditor is even faster than InstructDiffusion [11], and is comparable to Instruct Pix2pix [4].

GPU Memory. We calculate the peak VRAM usage of different models. As shown in Table 7, CoEditor consumes the least memory. This is because our chosen models, *i.e.*, Semantic-SAM [18] and Stable Diffusion Inpainting [29], are very lightweight, which allows our model to work with minimal memory.

6 Ethics Statement

6.1 Inappropriate Content, Privacy, and Discrimination

The AltBear dataset uses teddy bears as the protagonist, significantly reducing the risks caused by the public display of inappropriate contents. We manually reviewed all contents to ensure that privacy or discrimination is avoided. We release AltBear under the MIT license to keep transparency.

Table 7: Comparison on GPU memory.

Model	VRAM	Multiple
InstrcutPix2pix	18,118MB	1×
InstructDiffusion	8,378MB	0.46×
CoEditor	7,328MB	0.40×

6.2 Reproducibility

We build our method on top of the Stable Diffusion Inpainting model with publicly accessible code and checkpoints with a fixed random seed. We notice that the GPT API cannot guarantee identical responses so that we provide all instruction prompts for reference. To further improve the reproducibility, we also release the dataset, code and our trained model.

6.3 Anti-misuse

In order to ensure that the dataset is not misused, we have designed unique markers (refer to Section 3.4). We added the marker to the bottom right corner of the image to indicate that the image is used for responsible visual editing research only. Specifically, we have defined various tags to indicate whether the image has been edited or contains potential risks.

In addition, the interaction system of our model always maintains visibility of the concepts to be edited to prevent misuse of the model. We also call on subsequent work to maintain the visibility of concepts like ours, ensuring a beneficial role for the community through open and transparent methods.

7 Conclusion

In this work, we proposed a new task, namely responsible visual editing, which entails editing specific concepts within an image to render it more responsible while minimizing changes. To tackle abstract concepts in responsible visual editing, we proposed a **Co**gnitive **Editor** (**CoEditor**) by harnessing a large multimodal model (LMM) through a two-stage process: a perceptual cognitive process to locate what needs to be edited, and a behavioral cognitive process to strategize how to edit. We constructed a transparent and public dataset, **AltBear**, which represents harmful information using teddy bears instead of humans, thereby mitigating the negative implications on research. Experiments demonstrated that CoEditor can effectively comprehend abstract concepts within complex scenes and significantly surpass the performance of baseline models for responsible visual editing. Moreover, we validated that the AltBear dataset can well serve as a replacement of real-world images in the research of responsible visual editing, providing a safer benchmark. Our findings also revealed the potential of LMM in responsible AI.

Acknowledgement

This work was supported in part by the National Key R&D Program of China (2022YFA1004100), and the National Natural Science Foundation of China (NSFC) under grant No. 62441202.

References

1. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan++: How to edit the embedded images? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8296–8305 (2020)
2. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18208–18218 (2022)
3. Brack, M., Friedrich, F., Hintersdorf, D., Struppek, L., Schramowski, P., Kersting, K.: Sega: Instructing text-to-image models using semantic guidance. Advances in Neural Information Processing Systems **36** (2024)
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
5. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11315–11325 (2022)
6. Chen, J., Shen, Y., Gao, J., Liu, J., Liu, X.: Language-based image editing with recurrent attentive models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8721–8729 (2018)
7. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427 (2022)
8. Dong, H., Yu, S., Wu, C., Guo, Y.: Semantic image synthesis via adversarial learning. In: Proceedings of the IEEE international conference on computer vision. pp. 5706–5714 (2017)
9. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. arXiv preprint arXiv:2303.07345 (2023)
10. Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D.: Unified concept editing in diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5111–5120 (2024)
11. Geng, Z., Yang, B., Hang, T., Li, C., Gu, S., Zhang, T., Bao, J., Zhang, Z., Hu, H., Chen, D., et al.: Instructdiffusion: A generalist modeling interface for vision tasks. arXiv preprint arXiv:2309.03895 (2023)
12. Heng, A., Soh, H.: Continual learning for forgetting in deep generative models (2023)
13. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
14. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
15. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)

16. Kumari, N., Zhang, B., Wang, S.Y., Shechtman, E., Zhang, R., Zhu, J.Y.: Ablating Concepts in Text-to-Image Diffusion Models (May 2023), http://arxiv.org/abs/2303.13516, arXiv:2303.13516 [cs]
17. Kumari, N., Zhang, B., Wang, S.Y., Shechtman, E., Zhang, R., Zhu, J.Y.: Ablating concepts in text-to-image diffusion models. arXiv preprint arXiv:2303.13516 (2023)
18. Li, F., Zhang, H., Sun, P., Zou, X., Liu, S., Yang, J., Li, C., Zhang, L., Gao, J.: Semantic-sam: Segment and recognize anything at any granularity. arXiv preprint arXiv:2307.04767 (2023)
19. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
20. Ni, M., Li, X., Zuo, W.: Nuwa-lip: Language-guided image inpainting with defect-free vqgan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14183–14192 (2023)
21. Ni, M., Wu, C., Wang, X., Yin, S., Wang, L., Liu, Z., Duan, N.: Ores: Open-vocabulary responsible visual synthesis. arXiv preprint arXiv:2308.13785 (2023)
22. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
23. OpenAI: Chatgpt. https://chat.openai.com (2022)
24. OpenAI: Dall-e 3. https://openai.com/dall-e-3 (2023)
25. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
26. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
27. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
28. Rando, J., Paleka, D., Lindner, D., Heim, L., Tramèr, F.: Red-teaming the stable diffusion safety filter. arXiv preprint arXiv:2210.04610 (2022)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
30. Schramowski, P., Brack, M., Deiseroth, B., Kersting, K.: Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22522–22531 (2023)
31. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
32. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8695–8704 (2020)
33. Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. arXiv preprint arXiv:2310.11441 (2023)
34. Yu, T., Feng, R., Feng, R., Liu, J., Jin, X., Zeng, W., Chen, Z.: Inpaint anything: Segment anything meets image inpainting. arXiv preprint arXiv:2304.06790 (2023)

35. Zhang, E., Wang, K., Xu, X., Wang, Z., Shi, H.: Forget-me-not: Learning to forget in text-to-image diffusion models. arXiv preprint arXiv:2303.17591 (2023)
36. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. arXiv preprint arXiv:2305.13077 (2023)