

A MORE VISUALIZATION

Figure 5 showcases comparison of inference with and without the refinement, showing that the refinement is critical for achieving high-quality outputs. If ReDiff inferences without the refinement, errors tend to accumulate, such as repeated words or symbols and incoherent sentences, ultimately degrading the quality of the caption. This highlights the importance of the model’s refinement capability.

Beyond correcting the model’s own errors during generation, ReDiff also demonstrates a powerful, generalizable ability to revise disturbing inputs. As shown in Figure 6, we provide the model with an image and a user-provided caption containing either syntactic chaos or a factual hallucination. In both cases, our model successfully corrects the initial erroneous text and proceeds to generate a coherent and accurate completion, highlighting the strong revision ability of ReDiff.

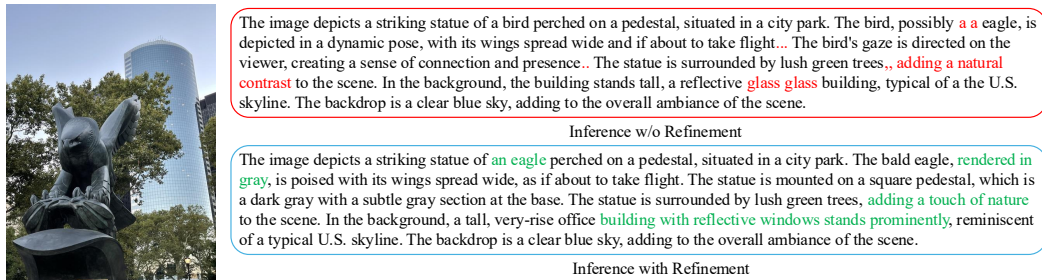


Figure 5: Generation results of ReDiff with or w/o refinement during inference.

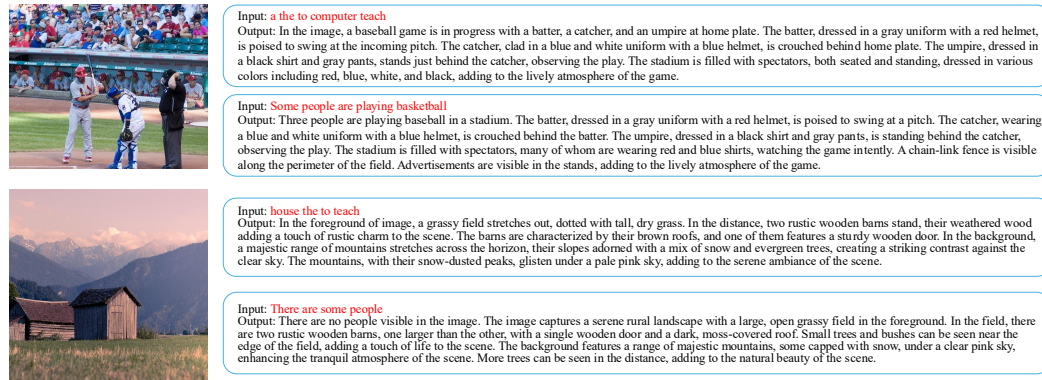


Figure 6: ReDiff can revise wrong input answers.

B ETHICS STATEMENT

All datasets and models used in this study are publicly available and open-source. No proprietary, private, or personally identifiable information was collected or used. The images employed are either natural scenes or normal human activities, without any violent, explicit, or otherwise harmful content. Therefore, the research meets relevant considerations regarding privacy, ethics and copy-right.