# Supplementary Materials: Appendices

Anonymous Authors

**Table 1: Twenty prompts for the image captioning task.**

| Image captioning task's prompt |
|---|
| a photo of |
| provide a brief overview of the image content |
| a picture of |
| describe the image content |
| offer a rich description of the image |
| please provide a detailed description of this picture |
| give a detailed description of the picture |
| an image of |
| this picture depicts |
| a photograph of |
| describe this picture |
| a snapshot of |
| describe the content of the image |
| description of the painting |
| this picture captures |
| describe the painting |
| for this picture, please provide a detailed description |
| introduce the content of the picture |
| provide a detailed description of the image |
| please describe what you see in this picture |

## A DATASET CONSTRUCTION

We employ a set of manually designed prompts to generate rephrased text. Table 1 presents the prompts utilized in our study. For each sample, we randomly selected one prompt from the twenty options to construct the rephrased text.

## B SEQUENTIAL MLLM DEBIAS EDITING RESULTS

Sequential debias editing is also an important scenario. In this section, we employ four methods (FT-L, FT-V, MEND, SERAC) for debias editing on the BLIP-2 OPT model. The reasons why we do not use KE and IKE are that they only support single editing. We conduct experiments on the image captioning task and randomly select one-third of the dataset as the experimental data. The results of different model editing methods on Reliability, T-Generality, V-Generality, and M-Locality for the IC task are shown in Figure 1.

We can observe that the overall performance of the four methods is consistent with the performance of single-edit. It is easy to see that the performance of the fine-tuning method (FT-L, FT-V) in terms of Reliability and T-Generality first improves and then decreases as the number of edits increases. Besides, the performance of the FT-V method gradually decreases in terms of V-Generality. This may be attributed to the fact that the impact of its modifications on the visual module parameters increases as the scale of editing increases. The model editing methods (MEND, SERAC) demonstrate relatively stable and superior performance across the three metrics of Reliability, T-Generality, and V-Generality. Regarding M-Locality, we can observe a significant downward trend in the performance of the MEND method as the scale of editing increases.

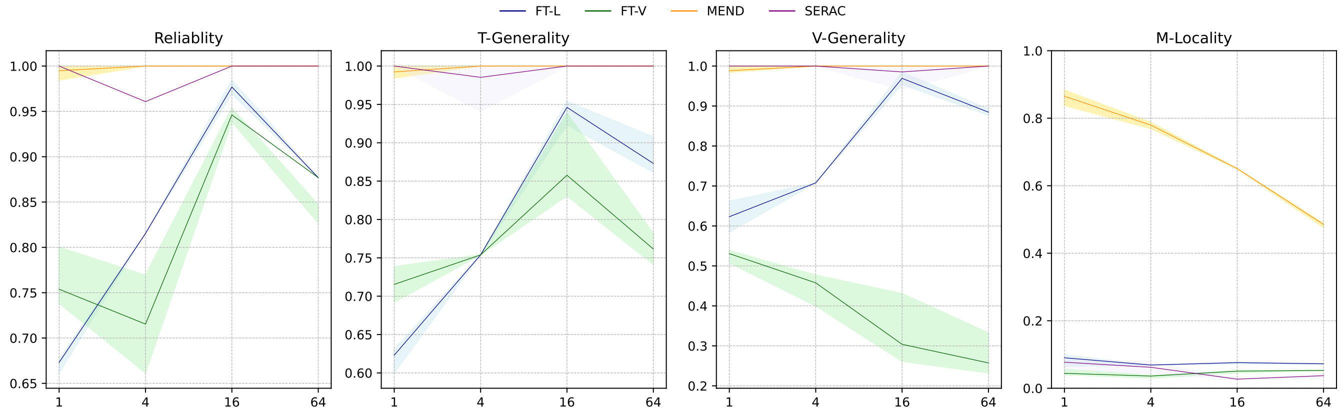

**Figure 1: Main results of multimodal model sequential debias editing. The gray shading represents the variance of each method. To reduce the variance, we ran 4 repetitions for edits within the range of [1,4] and 2 repetitions for edits within the range of [16,64].**