A. Appendix

The appendix is structured as follows - In Appendix A.1 we include additional details of leveraging LLaVA to generate optimal edit instructions. We provide further descriptions of the metrics used for our quantitative analysis in Appendix A.2. Appendix A.3 contains additional qualitative comparisons between REEDIT and the various baselines. Finally, Appendix A.4 contains additional examples showing poor and ambiguous samples from the InstructPix2Pix dataset, and Appendix A.5 discusses the current limitations of REEDIT.

A.1. Additional details of LLaVA-based edits

Below are the prompts $p1$, $p2$, and $p3$ that are used as input to LLaVA in various stages of our method and baselines. Fig. 6 summarizes the pipeline for generating captions and edit instruction using LLaVA.

1. **p1** *The given image is a 2x1 grid of two individual images. The image on the right is an edited version of the image on the left. Give a detailed explanation of the edits required to obtain the second image starting from the first image. The suggested edits can include addition/removal of objects, replacement of objects, change of style, change of background, motion, etc. Describe ONLY the edits, and do not mention any elements that don't require editing. Ignore minor changes and focus on a broad holistic view of the required edit. Give an answer in 100 words or less. Your answer should be in a single paragraph. Strictly adhere to this format.*

2. **p2** *Generate a one line description of an image generated after applying the following edit on this image - "<Response from LLaVA using p1>". Generate the caption in one line based on the content of the input image. If any part of the mentioned edit is not applicable to the given image, ignore it. Make sure that your caption completely describes the final image that would be obtained after applying this edit on the given image. The generated caption should be in one line, and should contain less than 20 words. Do not exceed 20 words.*

3. **p3** *Generate a one line edit instruction to edit the given image. The edit should follow the instruction in this longer edit - "<Response from LLaVA using p1>" Generate the edit instruction in a single line based on the content of the input image. If any part of the mentioned image is not applicable to the given image, ignore it. Make sure that your instruction is sufficient to replicate the describe edit. The generated instruction should be in one line, and should contain less than 20 words. Do not exceed 20 words.*

A.2. Details about Metrics

In this work, we use several image quality assessment metrics. Each metric provides a measure of a different aspect of the generation, refer to Table 2 for the average performance of REEDIT on our entire dataset of 1500 images. $\downarrow, \uparrow$ denote that a lower value of the metric is better and a higher value of the metric is better respectively.

a. LPIPS (↓). The Learned Perceptual Image Patch Similarity [61] calculates perceptual similarity between two images (here, $\hat{y}_{\text{edit}}, y_{\text{edit}}$) by comparing the deep features of two images. Traditionally, VGG [44] has been used to compute these features. This makes LPIPS more aligned with human visual perception, capturing subtle differences that traditional metrics like PSNR and SSIM might miss. Lower LPIPS values indicate higher similarity between images.

b. SSIM (↑). [53] is a measure of Structural Similarity between two images. A higher SSIM score generally indicates higher structural similarity. In Table 2, we report the structural similarity (SSIM) between $\hat{y}_{\text{edit}}, y_{\text{edit}}$. A higher value of SSIM indicates that the edit has been performed correctly on y.

c. CLIP Score (↑). This score [16] is a reference-free metric that measures the alignment between images and textual descriptions. Specifically, in our paper, it corresponds to the cosine similarity (normalized dot product) of $\hat{y}_{\text{edit}}, \mathcal{E}_{\text{text}}(g_{\text{caption}})$ where $\mathcal{E}_{\text{text}}(g_{\text{caption}})$ is the clip text embedding of the generated caption and the generated image.

d. Directional Similarity (↑). StyleGAN-Nada [12] proposed a directional CLIP similarity measure that measures the cosine similarity between the difference of edited and un-edited image ($\hat{y}_{\text{edit}} - y$), and the caption ($\mathcal{E}_{\text{text}}(g_{\text{caption}})$). A higher similarity indicates that the edit performed is in the direction of the text.

e. S-Visual (↑). Metric proposed in the baseline VISII [34] which computes the cosine similarity between the difference between the clip embeddings of the exemplar pair, and the difference between the clip embeddings of test image y and the generated image $\hat{y}_{\text{edit}}$. It is noteworthy that VISII optimizes the same function they use as a metric.

A.3. Additional Qualitative Results

Figs. 9 9 provides additional qualitative comparisons, highlighting the efficacy of REEDIT in exemplar-based image editing. Specifically, REEDIT *outperforms strong baselines across various types of edits*, including **a.** global style transfer, **b.** local style transfer, **c.** object replacement, and **d.** object addition.

A.4. Examples of poor samples in IP2P dataset

We present additional examples of poor and ambiguous samples from the InstructPix2Pix dataset in Fig. 7. We noticed a number of these samples, necessitating the manual curation of our evaluation dataset, as described in Sec. 4.

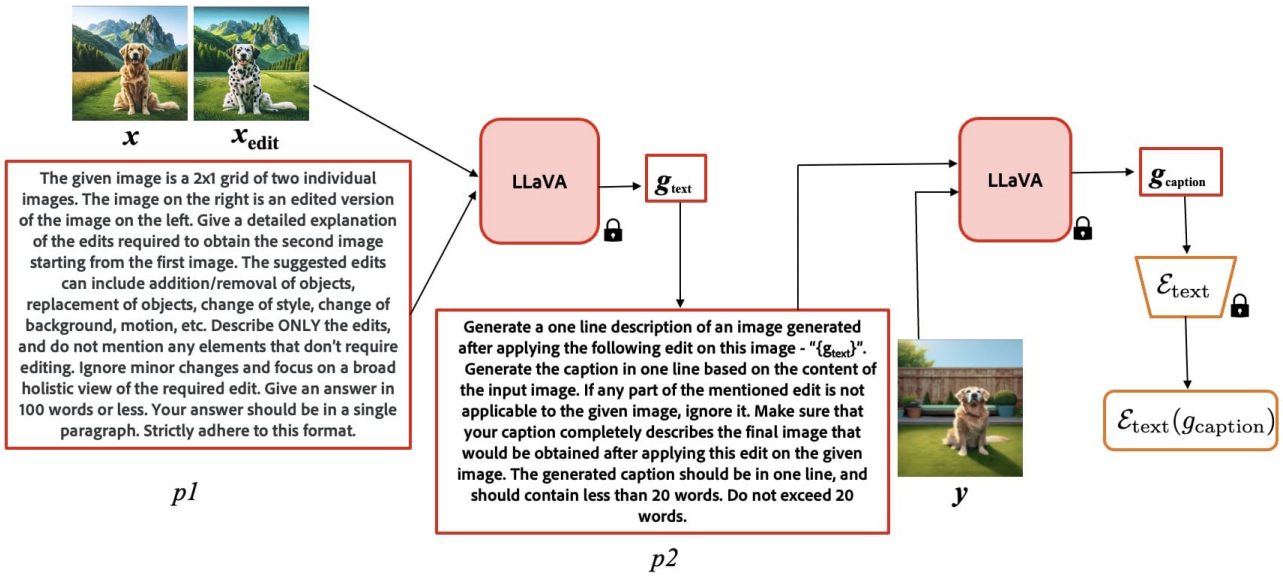

Figure 6. Overview of generating text-based edits using multimodal VLMS. **a.** In the first step, we input a detailed prompt $p1$, and a grid of exemplar pairs. The output g_{text} is then curated in the form of another prompt $p2$ which is passed as input to LLaVA with image y to generate g_{caption}. Note that all models are frozen and are used in inference mode.

A.5. Limitations of REEDIT

We present a novel approach for exemplar-based image editing that addresses several limitations of existing methods, such as over-reliance on models like InstructPix2Pix [4] (VISII). Our method produces state-of-the-art results approximately four times faster than strong baselines. However, it has some limitations. We illustrate some of these limitations in Fig. 8. For edits like *object addition*, our method's performance can be poor, especially when the objects are extremely small. Additionally, as seen in Row 2 of the same figure, REEDIT also fails to remove the large lake. However, all the remaining baselines also fail in these cases, producing high levels of distortions to produce the edit. We attribute of REEDIT in these cases due to the over-reliance on the guidance (f, Q, K), which prevents large changes in structure. A key area of exploration is *selective guidance* to circumvent this problem, which is part of our future work.

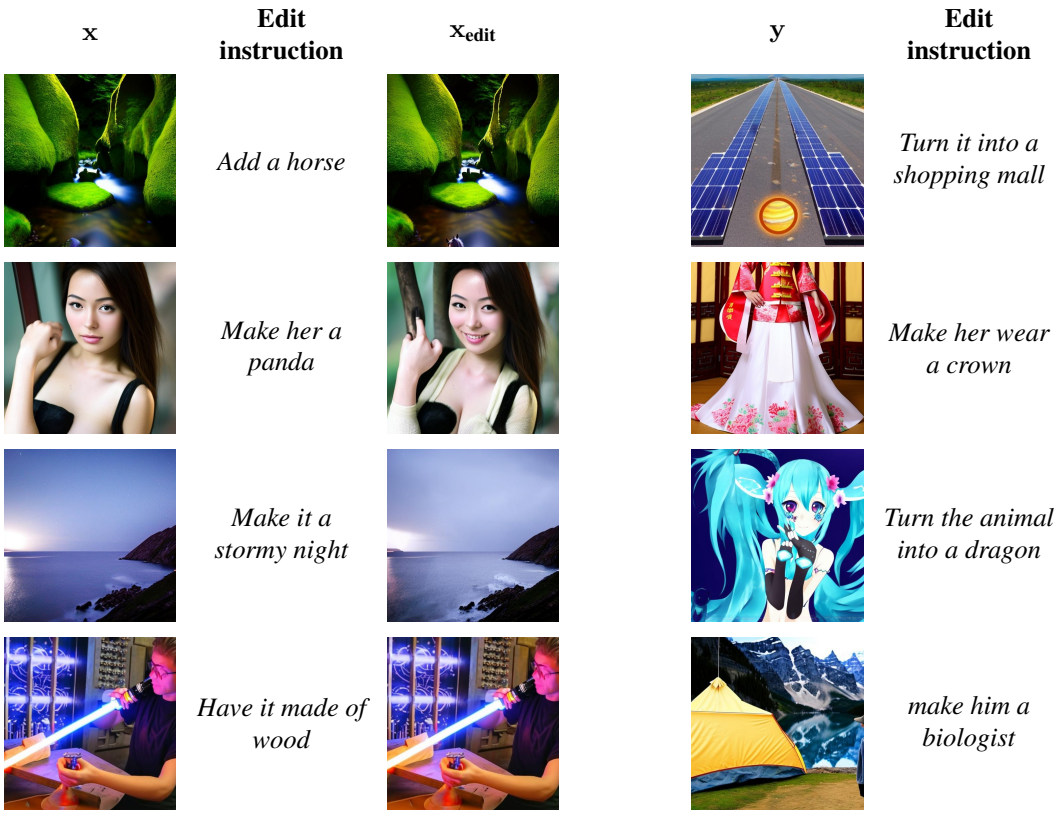

Figure 7. Images illustrating failure of automated dataset generation. **a**: Cases where exemplar pair x, x_{edit} does not represent expected edit. **b**: Cases where test image y does not conform with edit

Figure 8. Illustration of failure cases of REEDIT. REEDIT struggles most in addition or removal of objects. However, baselines also produce undesirable results in these cases.

Figure 9. Overview of additional qualitative comparisons: We show additional results across different edit types. REEDIT clearly outperforms the baselines consistently, by both maintaining the structure of the test image y and being faithful to the edit illustrated in the exemplar pair.). View at high magnification to observe subtle edits.

Figure 9. Overview of additional qualitative comparisons: We show additional results across different edit types. REEDIT clearly outperforms the baselines consistently, by both maintaining the structure of the test image y and being faithful to the edit illustrated in the exemplar pair. View at high magnification to observe subtle edits.