HIVE: Harnessing Human Feedback for Instructional Visual Editing

Shu Zhang*[1], Xinyi Yang*[1], Yihao Feng*[1], Can Qin[1], Chia-Chih Chen[1], Ning Yu[1], Zeyuan Chen[1], Huan Wang[1], Silvio Savarese[1,2], Stefano Ermon[2], Caiming Xiong[1], Ran Xu[1]
[1]Salesforce AI Research, [2]Stanford University

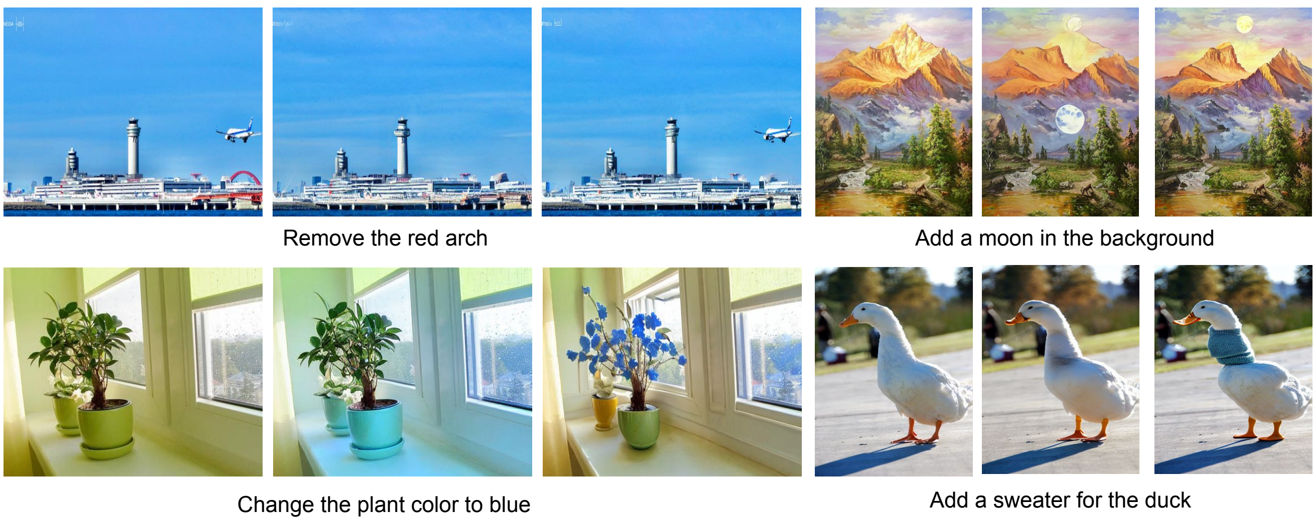


Figure 1. We show four groups of representative results. In each triplet, from left to right are: the original image, InstructPix2Pix [7] using our data (IP2P-Ours), and HIVE. We observe that HIVE leads to more acceptable results than the model without human feedback. For instance, in the left two examples, IP2P-Ours understands the editing instruction "remove" and "change to blue" individually, but fails to understand the corresponding objects. Human feedback resolves this ambiguity, as shown in other examples as well.

Abstract

Incorporating human feedback has been shown to be crucial to align text generated by large language models to human preferences. We hypothesize that state-of-the-art instructional image editing models, where outputs are generated based on an input image and an editing instruction, could similarly benefit from human feedback, as their outputs may not adhere to the correct instructions and preferences of users. In this paper, we present a novel framework to harness human feedback for instructional visual editing (HIVE). Specifically, we collect human feedback on the edited images and learn a reward function to capture the underlying user preferences. We then introduce scalable diffusion model fine-tuning methods that can incorporate human preferences based on the estimated reward.

Besides, to mitigate the bias brought by the limitation of data, we contribute a new 1.1M ***training dataset****, a 3.6K* ***reward dataset*** *for rewards learning, and a 1K* ***evaluation dataset*** *to boost the performance of instructional image editing. We conduct extensive empirical experiments quantitatively and qualitatively, showing that HIVE is favored over previous state-of-the-art instructional image editing approaches by a large margin.*

1. Introduction

State-of-the-art (SOTA) text-to-image generative models have shown impressive performance in terms of both image quality and alignment between output images and captions [1, 44, 46]. Thanks to the impressive generation abilities of these models, *instructional image editing* has emerged as one of the most promising application scenarios for content generation [7]. Different from traditional image editing [3, 16, 16, 30, 55, 55], where both the input and the edited caption are needed, *instructional image editing* only requires human-readable instructions. For instance, classic image editing approaches require an input caption "a dog is playing a ball", and an edited caption "a cat is playing a ball". In contrast, instructional image editing only needs editing instruction such as "change the dog to a cat". This

*Denotes equal contribution. Primary contact: shu.zhang@salesforce.com. Our project page: https://shugerdou.github.io/hive/.

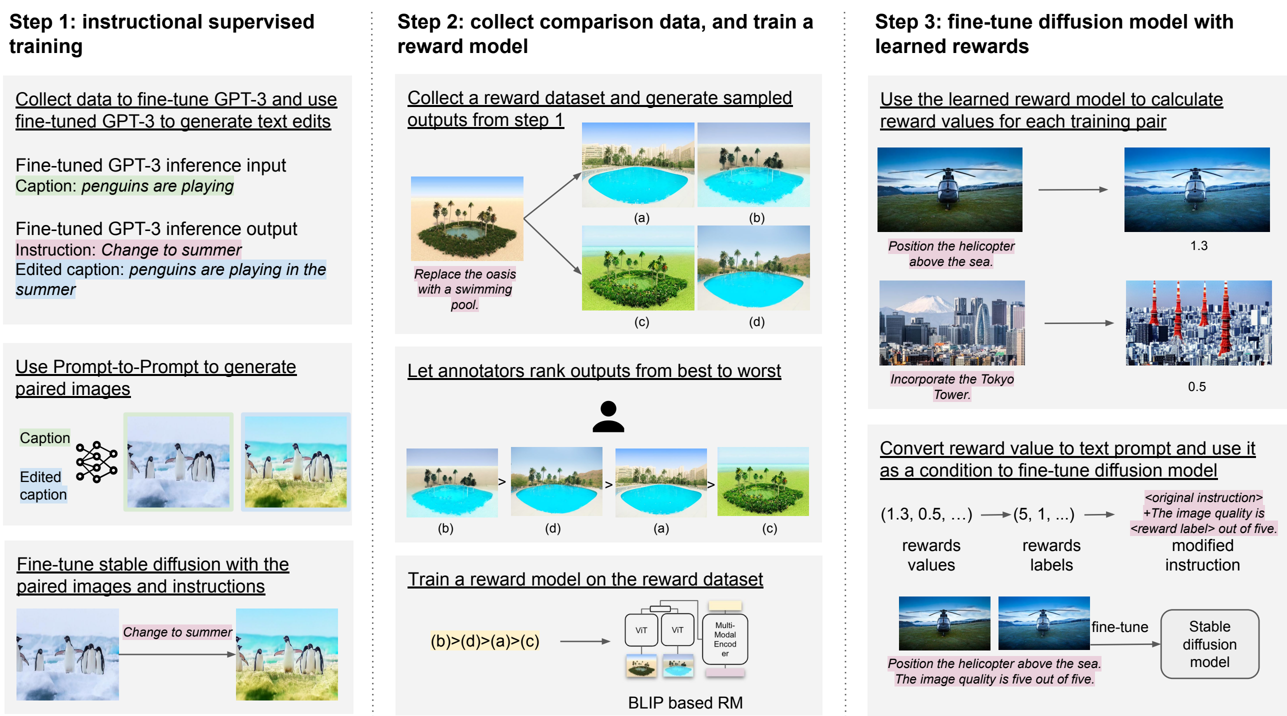


Figure 2. Overall architecture of HIVE. The first step is to train a baseline HIVE without human feedback. In the second step, we collect human feedback to rank variant outputs for each image-instruction pair, and train a reward model to learn the rewards. In the third step, we fine-tune diffusion models by integrating the estimated rewards.

experience mimics *how humans naturally perform image editing*.

Instructional image editing was first proposed in InstructPix2Pix [7], which fine-tunes a pre-trained stable diffusion [46] by curating a triplet of the original image, instruction, and edited image, with the help of GPT-3 [8] and Prompt-to-Prompt image editing [16]. Though achieving promising results, the training data generation process of InstructPix2Pix lacks explicit alignment between editing instructions and edited images.

Consequently, the modified images may only align to a certain extent with the editing instructions, as shown in the second column of Fig. 4. Furthermore, since these editing instructions are provided by human users, it's crucial that the final edited images accurately reflect the users' true intentions and preferences. Typically, humans prefer to make selective changes to the original images, which are usually not factored into the training data or objectives of InstructPix2Pix [7]. Considering this observation and the recent successes of ChatGPT [35], we propose to refine the stable diffusion process with human feedback. This adjustment aims to ensure that the edited images more closely correspond to editing instructions provided by humans.

For large language models (LLMs) such as InstructGPT [35, 37], we often first learn a reward function to reflect what humans care about or prefer on the generated text output, and then leverage reinforcement learning (RL) algorithms such as proximal policy optimization (PPO) [50] to fine-tune the models. This process is often referred to as *reinforcement learning with human feedback* (RLHF). Leveraging RLHF to fine-tune diffusion-based generative models, however, remains challenging. Applying on-policy algorithms (*e.g.*,PPO) to maximize rewards during the fine-tuning process can be prohibitively expensive due to the hundreds or thousands of denoising steps required for each sampled image. Moreover, even with fast sampling methods [21, 31, 52, 57], it is still challenging to back-propagate the gradient signal to the parameters of the U-Net. [1]

To address the technical issues described above, we propose Harnessing **H**uman Feedback for **I**nstructional **V**isual **E**diting (**HIVE**), which allows us to fine-tune diffusion-based generative models with human feedback. As shown in Fig. 2, HIVE consists of three steps:

1) We perform instructional supervised fine-tuning on the dataset that combines our newly collected 1.1M training data and the data from InstructPix2Pix. Since observing failure cases and suspecting the grounding visual components from image to instruction is still a challenging problem, we collect 1.1M training data.

2) For each input image and editing instruction pair, we ask human annotators to rank variant outputs of the fine-tuned model from step 1, which gives us a reward learning

[1] We present a rigorous discussion on the difficulty in Appendix C.1.

dataset. Using the collected dataset, we then train a reward model (RM) that reflects human preferences.

3) We estimate the reward for each training data used in step 1, and integrate the reward to perform human feedback diffusion model finetuning using our proposed objectives presented in Sec. 3.4.

Our main contributions are summarized as follows:

- To tackle the technical challenge of fine-tuning diffusion models using human feedback, we introduce two scalable fine-tuning approaches in Sec. 3.4, which are computationally efficient and offer similar costs compared with supervised fine-tuning. Moreover, we empirically show that human feedback is an essential component to boost the performance of instructional image editing models.
- To explore the fundamental ability of instructional editing, we create a new dataset for HIVE including three subdatasets: a new 1.1M **training dataset**, a 3.6K **reward dataset** for rewards learning, and a 1K **evaluation dataset**.
- To increase the diversity of the data for training, we introduce cycle consistency augmentation based on the inversion of editing instruction. Our dataset has been enriched with one pair of data for bi-directional editing.

2. Related Work

Text-To-Image Generation. Text-to-image generative models have achieved tremendous success in the past decade. Generative adversarial nets (GANs) [15] is one of the fundamental methods that dominated the early-stage works [45, 59, 61]. Recently, diffusion models [17, 51–53] have achieved state-of-the-art text-to-image generation performance. [12, 29, 34, 43, 44, 46, 47, 60]. As a result, instead of training a text-to-image model from scratch, our work focuses on *fine-tuning* existing stable diffusion model [46], by leveraging additional human feedback.

Image Editing. Similarly, diffusion models based image editing methods, *e.g*. SDEdit [32], BlendedDiffusion [3], BlendedLatentDiffusion [2], DiffusionClip [22], EDICT [55] or MagicMix [30], have garnered significant attention in recent years. To leverage a pre-trained image-text representation (*e.g*., CLIP [42], BLIP [28]) and text-to-image diffusion based pre-trained models [44, 46, 47], most existing works focus on text-based localized editing [5, 16, 33]. Prompt-to-Prompt [16] edits the cross-attention layer in Imagen and stable diffusion to control the similarity of image and text prompt. ControlNet [62] and UniControl [41] adopt controllable conditions to control image editings. Recently, InstructPix2Pix [7] tackle the problem via a different approach, requiring only human-readable editing instruction to perform image editing. Our work follows the same direction as InstructPix2Pix[7] and leverages human feedback to address the misalignment between editing instructions and resulting edited images.

Learning with Human Feedback. Incorporating human feedback into the learning process can be a highly effective way to enhance performance across various tasks such as fine-tuning LLMs [4, 35, 37, 48, 54], robotic simulation [10, 18], computer vision [40], and to name a few. Many existing works leverage PPO [50] to align to human feedback, however on-policy RL algorithms are not suitable for diffusion-based model fine-tuning (See more discussion in Appendix C.1).

Simultaneously, several concurrent works [25, 56, 58] study the text-to-image generation problem using human feedback. ReFL [58] investigates how to back-propagate the reward signal to random latter denoising step in the diffusion process, while [56] explores how to design fine-grained human preference score to improve the generation quality. [25] leverages human feedback to align text-to-image generation, where they naively view reward as weights to perform maximum likelihood training. Different from the above works, our work tackles the problem of *instructional image editing*, where there are *little or even no ground truth data* for the alignment between *human-readable* editing instructions and edited images. In addition, the conditions on both image input and instructions make the human feedback more valuable than standard text-to-image tasks, since the conditions make the training harder than standard text-to-image tasks.

3. Methodology

In this section, we introduce the new datasets we collected in Sec. 3.1, and explain the three major steps of HIVE in the rest of the section. Concretely, we introduce the instructional supervised training in Sec. 3.2, and describe how to train a reward model to score edited images in Sec. 3.3, then present two scalable fine-tuning methods to align diffusion models with human feedback in Sec. 3.4.

3.1. Dataset

Instructional Edit Training Dataset. We follow the same method of [7] to generate the training dataset. We collect 1K images and their corresponding captions. We ask three annotators to write three instructions and corresponding edited captions based on the collected input captions. Therefore, we obtain 9K prompt triplets: input caption, instruction, and edited caption. We fine-tune GPT-3 [8] with OpenAI API v0.25.0 [36] with them. We use the fine-tuned GPT-3 to generate five instructions and edited captions per input image-caption pair in Laion-Aesthetics V2 [49]. We observe that the captions from Laion are not always visually descriptive, so we use BLIP [28] to generate more diverse types of image captions. Later stable diffusion based Prompt-to-Prompt [16] is adopted to generate paired images. In addition, we design a cycle-consistent augmentation method (Sec. 3.2.1) to generate additional training data. We generate 1.17M training triplets in total. Combining the 281K training data from [7], we obtain 1.45M training image pairs along with instructions.

Reward Fine-tuning Dataset. We collect 3.6K image-instruction pairs for the task of reward fine-tuning. Among them, 1.6K image-instruction pairs are manually collected, and the rest are from Laion-Aesthetics V2 with GPT-3 generated instructions. We use this dataset to ask annotators to rank various model outputs.

Evaluation Dataset. We use two evaluation datasets: the test dataset in [7] for quantitative evaluation and a new 1K dataset collected for the user study. The quantitative evaluation dataset is generated following the same method as the training dataset, which means that the dataset does not contain real images. Our collected 1K dataset contains 200 real images, and each image is annotated with five human-written instructions. More details of annotation tooling, guidelines, and analysis are in Appendix A.

3.2. Instructional Supervised Training

We follow the instructional fine-tuning method in [7] with two major upgrades on dataset curation (Sec. 3.1) and cycle consistency augmentation (Sec. 3.2.1). A pre-trained stable diffusion model [46] is adopted as the backbone architecture. In instructional supervised training, the stable diffusion model has two conditions $c = [c_I, c_E]$, where c_E is the editing instruction, and c_I is the latent space of the original input image. In the training process, a pre-trained auto-encoder [23] with encoder $\mathcal{E}$ and decoder $\mathcal{D}$ is used to convert between edited image $\tilde{\boldsymbol{x}}$ and its latent representation $z = \mathcal{E}(\tilde{\boldsymbol{x}})$. The diffusion process is composed of an equally weighted sequence of denoising autoencoders $\epsilon_\theta(z_t, t, c)$, $t = 1, \cdots, T$, which are trained to predict a denoised variant of their input z_t, a noisy version of z. The objective of instructional supervised training is:

$$L = \mathbb{E}_{\mathcal{E}(\tilde{\boldsymbol{x}}),c,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c))\|_2^2\right].$$

3.2.1 Cycle Consistency Augmentation

Cycle consistency is a powerful technique that has been widely applied in image-to-image generation [19, 63]. It involves coupling and inverting bi-directional mappings of two variables X and Y, $G : X \to Y$ and $F : Y \to X$, such that $F(G(X)) \approx X$ and vice versa. This approach has been shown to enhance generative mapping in both directions.

While Instructpix2pix [7] considers instructional image editing as a single-direction mapping, we propose adding cycle consistency. Our approach involves a forward-pass editing step, $F : x \xrightarrow{inst} \tilde{\boldsymbol{x}}$. We then introduce instruction reversion to enable a reverse-pass mapping, $R : \tilde{\boldsymbol{x}} \xrightarrow{\sim inst} x$. In this way, we could close the loop of image editing as: $x \xrightarrow{inst} \tilde{\boldsymbol{x}} \xrightarrow{\sim inst} x$, *e.g.* "add a dog" to "remove the dog".

To ensure the effectiveness of this technique, we need to separate invertible and non-invertible instructions from the dataset. We devised a rule-based method that combines speech tagging and template matching. We found that most instructions adhere to a particular structure, with the verb appearing at the start, followed by objects and prepositions. Thus, we grammatically tagged all instructions using the Natural Language Toolkit (NLTK) [2]. We identified all invertible verbs and pairing verbs, and also analyzed the semantics of the objects and the prepositions used. By summarizing invertible instructions in predefined templates, we matched desired instructions. Our analysis revealed that 29.1% of the instructions in the dataset were invertible. We augmented this data to create more comprehensive training data, which facilitated cycle consistency. For more information, see Appendix B.1.

3.3. Human Feedback Reward Learning

The second step of HIVE is to learn a reward function $\mathcal{R}_\phi(\tilde{\boldsymbol{x}}, c)$, which takes the original input image, the text instruction condition $c = [c_I, c_E]$, and the edited image $\tilde{\boldsymbol{x}}$ that is generated by the fine-tuned stable diffusion as input, and outputs a scalar that reflects human preference.

Unlike InstructGPT which only takes text as input, our reward model $\mathcal{R}_\phi(\tilde{\boldsymbol{x}}, c)$ needs to measure the alignment between instructions and the edited images. To address the challenge, we present a reward model architecture in Fig. 3, which leverages pre-trained vision-language models such as BLIP [28]. More specifically, the reward model employs an image-grounded text encoder as the multi-modal encoder to take the joint image embedding and the text instruction as input and produce a multi-modal embedding. A linear layer is then applied to the multi-modal embedding to map it to a scalar value. More details are in Appendix B.2.

With the specifically designed network architecture, we train the reward function $\mathcal{R}_\phi(\tilde{\boldsymbol{x}}, c)$ with our collected reward fine-tuning dataset $\mathcal{D}_{\text{human}}$ induced in Sec. 3.1. For each input image c_I and instruction c_E pair, we have K edited images $\{\tilde{\boldsymbol{x}}\}_{k=1}^K$ ranked by human annotators, and denote the human preference of edited image $\tilde{\boldsymbol{x}}_i$ over $\tilde{\boldsymbol{x}}_j$ by $\tilde{\boldsymbol{x}}_i \succ \tilde{\boldsymbol{x}}_j$. Then we can follow the Bradley-Terry model of preferences [6, 37] to define the pairwise loss function:

$$\ell_{\text{RM}}(\phi) := -\sum_{\tilde{\boldsymbol{x}}_i \succ \tilde{\boldsymbol{x}}_j} \log\left[\frac{\exp(\mathcal{R}_\phi(\tilde{\boldsymbol{x}}_i, c))}{\sum_{k=i,j}\exp(\mathcal{R}_\phi(\tilde{\boldsymbol{x}}_k, c))}\right],$$

where $(i, j) \in [1 \ldots K]$ and we can get $\binom{K}{2}$ pairs of comparison for each condition c. Similar to [37], we put all the $\binom{K}{2}$ pairs for each condition c in a single batch to learn the reward functions. We provide a detailed reward model training discussion in Appendix B.2.

3.4. Human Feedback based Model Fine-tuning

With the learned reward function $\mathcal{R}_\phi(c, \tilde{\boldsymbol{x}})$, the next step is to improve the instructional supervised training model by

[2]https://www.nltk.org/

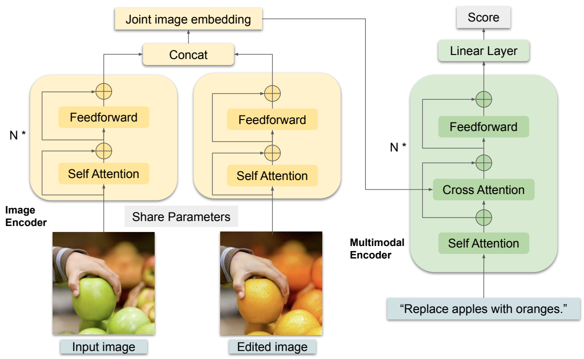


Figure 3. Model architecture for reward $R(\tilde{\boldsymbol{x}}, c)$. Here the reward model evaluates human preference for an edited image of a hand selecting an orange compared to the original input image of the hand selecting an apple. The input to the reward model includes both images and a text instruction. The output is a score indicating the degree of preference for the edited image based on the input image and instruction.

reward maximization. As a result, we can obtain an instructional diffusion model that aligns with human preferences.

The RL fine-tuning techniques we present are built upon recent offline RL techniques [9, 20, 27, 38] With an input image and editing instruction condition $c = [c_I, c_E]$, we define the edited image data distribution generated by the instructional supervised diffusion model as $p(\tilde{\boldsymbol{x}}|\ c)$, and the edited image data distribution generated by the current diffusion model we want to optimize as $\rho(\tilde{\boldsymbol{x}}|\ c)$, then under the pessimistic principle of offline RL, we can optimize ρ by the following objectives:

$$J(\rho) := \max_\rho \mathbb{E}_c \big[\mathbb{E}_{\tilde{\boldsymbol{x}}\sim\rho(\cdot|c)}[\mathcal{R}_\phi(\tilde{\boldsymbol{x}}, c)] - \eta \mathrm{KL}(\rho(\tilde{\boldsymbol{x}}|c)||p(\tilde{\boldsymbol{x}}|c))\big]\,, \quad (1)$$

where η is a hyper-parameter. The first term in Eq. (1) is the standard reward maximization in RL. The second term is a regularization to stabilize learning, which is a widely used technique in offline RL [24], and is also adopted for PPO fine-tuning of InstructGPT (*a.k.a* "PPO-ptx") [37].

To avoid using sampling-based methods to optimize ρ, we can differentiate $J(\rho)$ *w.r.t* $\rho(\tilde{\boldsymbol{x}}|c)$ and solve for the optimal $\rho^*(\tilde{\boldsymbol{x}}|c)$, resulting the following expression for the optimal solution of Eq. (1):

$$\rho^*(\tilde{\boldsymbol{x}}|c) \propto p(\tilde{\boldsymbol{x}}|c)\exp\left(\mathcal{R}_\phi(\tilde{\boldsymbol{x}}, c)/\eta\right)\,, \quad (2)$$

or $\rho^*(\tilde{\boldsymbol{x}}|c) = \frac{1}{Z(c)} p(\tilde{\boldsymbol{x}}|c)\exp\left(\mathcal{R}_\phi(\tilde{\boldsymbol{x}}, c)/\eta\right)$, with $Z(c) = \int p(\tilde{\boldsymbol{x}}|c)\exp\left(\mathcal{R}_\phi(\tilde{\boldsymbol{x}}, c)/\eta\right) d\tilde{\boldsymbol{x}}$ being the partition function. A detailed derivation is in Appendix C.2.

Weighted Reward Loss. The optimal target distribution $\rho^*(\tilde{\boldsymbol{x}}|c)$ in Eq. (2) can be viewed as an exponential reward-weighted distribution for $p(\tilde{\boldsymbol{x}}|c)$. Moreover, we have already obtained the empirical edited image data drawn from $p(\tilde{\boldsymbol{x}}|c)$ when constructing the instructional editing dataset, and we can view the exponential reward weighted edited image $\tilde{\boldsymbol{x}}$ from the instructional editing dataset as an empirical approximation of samples drawn from $\rho^*(\tilde{\boldsymbol{x}}|c)$. Formally, we can fine-tune a diffusion model thus it generates data from $\rho^*(\tilde{\boldsymbol{x}}|c)$, resulting in the weighted reward loss:

$$\ell_{\mathrm{WR}}(\theta) := \mathbb{E}_{\mathcal{E}(\tilde{\boldsymbol{x}}),c,\epsilon\sim\mathcal{N}(0,1),t}\left[\omega(\tilde{\boldsymbol{x}}, c)\cdot\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2\right]\,,$$

with $\omega(\tilde{\boldsymbol{x}}, c) = \exp\left(\mathcal{R}_\phi(\tilde{\boldsymbol{x}}, c)/\eta\right)$ being the *exponential reward weight* for edited image $\tilde{\boldsymbol{x}}$ and condition c. Different from RL literature [38, 39] using exponential reward or advantage weights to learn a policy function, our weighted reward loss is derived for fine-tuning stable diffusion.

Condition Reward Loss. We can also leverage the control-as-inference perspective of RL [26] to transform Eq. (2) to a conditional reward expression, thus we can directly view the reward as a conditional label to fine-tune diffusion models. Similar to [26], we introduce a new binary variable R^* indicating whether human prefers the edited image or not, where $R^* = 1$ denotes that human prefers the edited image, and $R^* = 0$ denotes that human does not prefer, thus we have $p(R^* = 1 \mid \tilde{\boldsymbol{x}}, c) \propto \exp\left(\mathcal{R}_\phi(\tilde{\boldsymbol{x}}, c)\right)$. Together with Eq. (2), and applying Bayes rules gives us the following derivation:

$$\begin{aligned} p(\tilde{\boldsymbol{x}}|c)\exp\left(\mathcal{R}_\phi(\tilde{\boldsymbol{x}}, c)/\eta\right) &:= q(\tilde{\boldsymbol{x}}|c)\left(p(R^* = 1 \mid \tilde{\boldsymbol{x}}, c)\right)^{1/\eta} \\ &= p(\tilde{\boldsymbol{x}}|c)\left(\frac{p(\tilde{\boldsymbol{x}}|\ R^* = 1, c)p(R^* = 1|\ c)}{p(\tilde{\boldsymbol{x}}|c)}\right)^{1/\eta} \\ &\propto p(\tilde{\boldsymbol{x}}|c)^{1-1/\eta} p(\tilde{\boldsymbol{x}}|\ R^* = 1, c)^{1/\eta}\,, \end{aligned}$$

where we drop $p(R^* = 1|\ c)$ since it is a constant w.r.t $\tilde{\boldsymbol{x}}$. We can now view the reward for each edited image as an additional condition. Define the new condition $\tilde{c} = [c_I, c_E, c_R]$, with c_R as the reward label, we can fine-tune the diffusion model with the condition reward loss:

$$\ell_{\mathrm{CR}}(\theta) = \mathbb{E}_{\mathcal{E}(x),\tilde{c},\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tilde{c})\|_2^2\right].$$

We quantize the reward into five categories, based on the quantile of the empirical reward distribution of the training dataset, and convert the reward value into a text prompt. For instance, if the reward value of a training pair lies in the bottom 20% of the reward distribution of the dataset, then we convert the reward value as a text prompt condition c_R :=*"The image quality is one out of five"*. And during the inference time to generate edited images, we fix the text prompt as c_R :=*"The image quality is five out of five"*, indicating we want the generated edited images with the highest reward. We empirically find this technique improves the stability of fine-tuning.

4. Experiments

This section presents the experimental results and ablation studies of HIVE's technical choices, demonstrating the effectiveness of our method. We adopt the default guidance

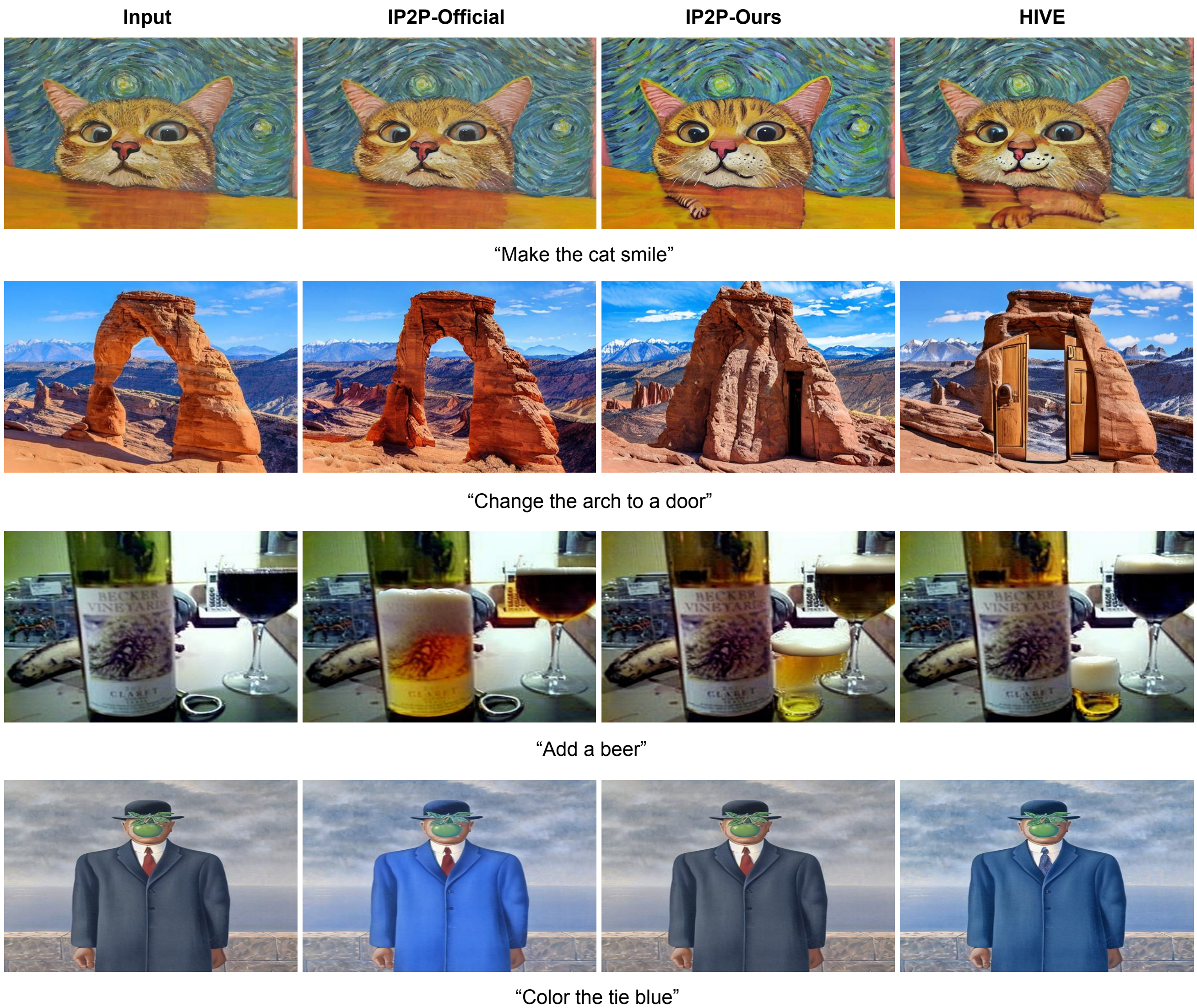


Figure 4. Comparisons between IP2P-Official (InstructPix2Pix official model), IP2P-Ours (InstructPix2Pix using our data) and HIVE. HIVE can boost performance by understanding the instruction correctly.

scale parameters in InstrcutPix2Pix for a fair comparison. Through our experiments, we discovered that the conditional reward loss performs slightly better than the weighted reward loss, and therefore, we present our results based on the conditional reward loss. The detailed comparisons can be found in Sec. 4.2 and Appendix D.3.

We evaluate our method using two datasets: a synthetic evaluation dataset with 15,652 image pairs from [7] and a self-collected 1K evaluation dataset with real image-instruction pairs. For the synthetic dataset, we follow InstructPix2Pix's quantitative evaluation metric and plot the trade-offs between CLIP image similarity and directional CLIP similarity [14]. For the 1K dataset, we conduct a user study where for each instruction, the images generated by competing methods are reviewed and voted by three human annotators, and the winner is determined by majority votes.

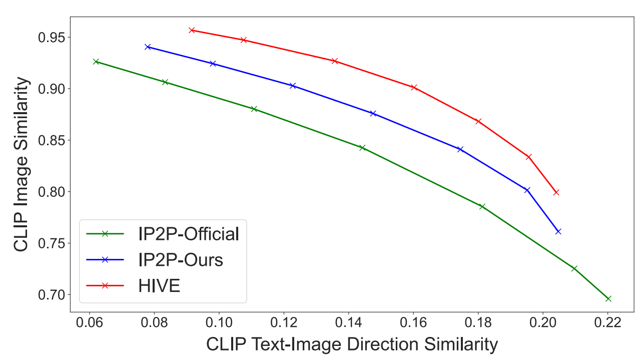
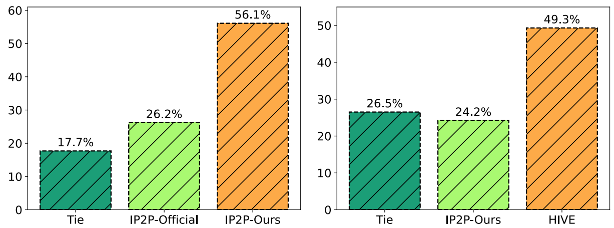


Figure 5. Comparisons between IP2P-Official, IP2P-Ours, and HIVE. It plots tradeoffs between consistency with the input image and consistency with the edit. The higher the better. For all methods, we adopt the same parameters as that in [7].



IP2P-Official vs IP2P-Ours IP2P-Ours vs HIVE

Figure 6. User study of comparison between (a) IP2P-Official vs IP2P-Ours and (b) IP2P-Ours and HIVE. IP2P-Ours obtains 30% more votes than IP2P-Official. HIVE obtains 25% more votes than that IP2P-Ours.

4.1. Baseline Comparisons

We perform experiments with the same setup as Instruct-Pix2Pix, where stable diffusion (SD) v1.5 is adopted. We compare three models: InstructPix2Pix official model (IP2P-Official), InstructPix2Pix using our data (IP2P-Ours) [3], and HIVE. We report the quantitative results on the synthetic evaluation dataset in Fig. 5. We observe that IP2P-Ours improves notably over IP2P-Official (blue curve vs. green curve). Moreover, human feedback further boosts the performance of HIVE (red curve vs blue curve) over IP2P-Ours by a large margin. In other words, with the same directional similarity value, HIVE obtains better image consistency than InstructPix2Pix.

To test the effectiveness of HIVE on real-world images, we report the user study results on the 1K evaluation dataset. We use "Tie" to represent that users think results are equally good or equally bad. As shown in Fig. 6(a), IP2P-Ours gets around 30% more votes than the IP2P-Official. The result is consistent with the user study on the synthetic dataset. We also demonstrate the user study outcome between HIVE and IP2P-Ours in Fig. 6(b). The user study indicates similar conclusions to the consistency plot, where HIVE gets around 25% more favorites than IP2P-Ours.

In Fig. 4, we present representative edits that demonstrate the effectiveness of HIVE. The results show that while using more data can partially improve editing instructions without human feedback, the reward model leads to better alignment between instruction and the edited image. For example, in the second row, IP2P-Ours generates a door-like object, but with the guidance of human feedback, the generated door matches human perception better. In the fourth row, the example of which is from the failure examples in [7], HIVE can locate the tie and change its color correctly.

Additionally, our visual analysis of the results (Fig. 7) indicates that the HIVE model tends to preserve the remaining part of the original image that is not instructed to be edited, while IP2P-Ours leads to excessive image editing more often. For instance, in the first example of Fig. 7, HIVE blends a pond naturally into the original image. The two Instruct-Pix2Pix models fulfill the same instruction, however, at the same time, alter the uninstructed part of the original background.

[3]It is the same to HIVE without human feedback.

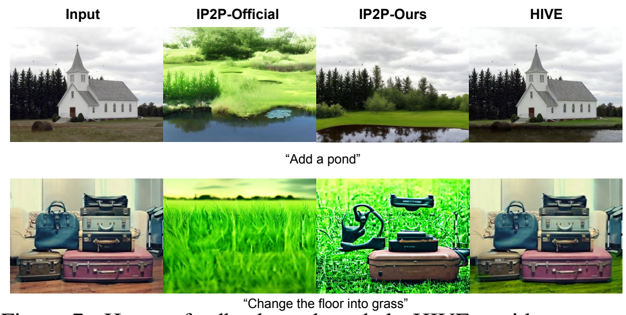
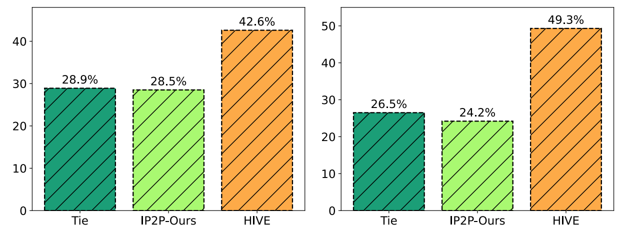
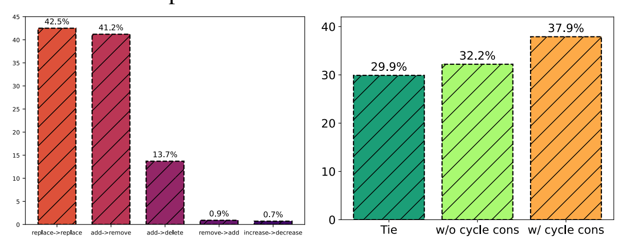


Figure 7. Human feedback tends to help HIVE avoid unwanted excessive image modifications.



HIVE with weighted reward loss HIVE with condition reward loss

Figure 8. User study of pairwise comparison between (a) HIVE with weighted reward loss and (b) HIVE with condition reward loss. The human preferences are close to each other.



Top five augmentations. User study of cycle consistency.

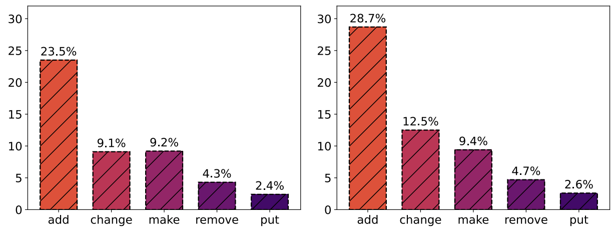
Figure 9. Cycle consistency analysis.

4.2. Ablation Study

Weighted Reward and Condition Reward Loss. We perform user study on HIVE with these two losses individually. As shown in Fig. 8, these two losses obtain similar human preferences on the evaluation dataset. More comparisons are in Appendix D.

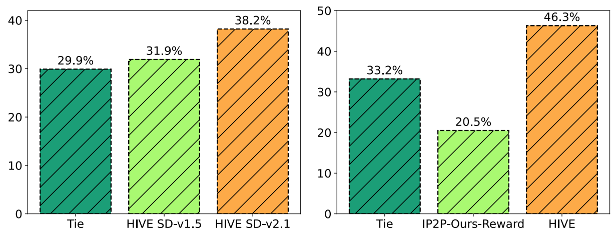
Cycle Consistency We analyze the impact of it which is introduced in Sec. 3.2.1. The top five augmentations in the cycle consistency are demonstrated in Fig. 9(a). We perform evaluation on both synthetic dataset and the 1K evaluation dataset. The user study in Fig. 9(b) shows that the cycle consistency augmentation improves the performance of HIVE by a notable margin.

Success Rate on Verbs It is observed that five verbs take around 85% of all verbs, where details can be found in



IP2P-Ours HIVE

Figure 10. Success rate of IP2P-Ours and HIVE on top five verbs.



HIVE SD v1.5 and v2.1 HIVE-Reward vs. HIVE

Figure 11. User study of pairwise comparison between (a) HIVE with SD v1.5 and v2.1 and (b) HIVE conditioning on reward score and HIVE. The human preferences are very close to each other.

Sec. A. We compare HIVE with IP2P-Ours on these five verbs, and report the success rate of these two methods on these verbs. It is seen in Fig. 10 that HIVE improves the most on "add" from 23.5% to 28.7%.

Other Baselines. To test the effectiveness of HIVE, we experiment two additional baselines. In Fig. 11(a), we upgrade the backbone of stable diffusion from v1.5 to v2.1. We observe that the upgraded backbone slightly improves the results. In Fig. 11(b), we directly use the reward scalar instead of the reward prompt as the condition for training, and the condition on the highest reward scalar for generating the image. We adopt the user study to compare it (named HIVE-reward) with HIVE. HIVE obtains 25.8 % more votes than the baseline model conditioned on the reward score. This is mainly because directly conditioning on the highest reward might cause overfiting.

Failure Cases and Limitations. We summarize representative failure cases in Fig. 12. First, some instructions cannot be understood. In the upper left example in Fig. 12, the prompt "zoom in" or similar instructions can rarely be successful. We believe the root cause is current training data generation method fails to generate image pairs with this type of instruction. Second, counting and spatial reasoning are common failure cases (see the upper right example in Fig. 12). We find that the instruction "one", "two", or "on the right" can lead to many undesired results. Third, the object understanding sometimes is wrong. In the bottom left example, the red color is changed on the wrong object. This is a common error in HIVE, where instructed edited objects are wrongly recognized.

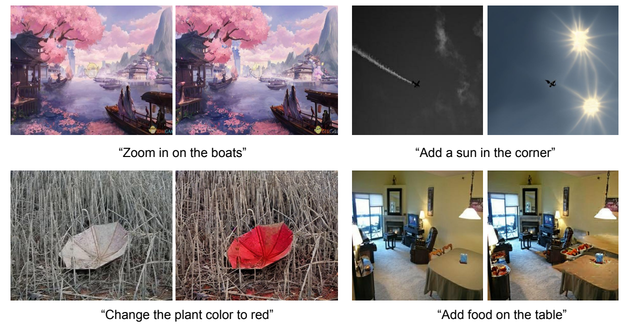


Figure 12. Failure examples.

We find some other limitations as well. One limitation of HIVE is that it cannot bring benefits to the cases where all outputs by the model without human feedback obtain the same wrong results. In such cases, user preferences cannot always be beneficial to the results. We believe that improving the data as well as the base model is an important step in the future. Another limitation is that compared to Prompt-to-Prompt [16], which is used to generate our training data, HIVE sometimes leads to some unstructured change in the image. We think that it is because of the limitation of the current training data. Instructed editing can have more diverse and ambiguous scenarios than traditional image editing problems. Using GPT-3 to finetune prompts to generate the training data is limited by the model and the labeled data. More ablation studies are in Appendix D.

5. Conclusion and Discussion

In our paper, we introduce a novel framework called HIVE that enables instructional image editing with human feedback. Our framework integrates human feedback, which is quantified as reward values, into the diffusion model finetuning process. We design two variants of the approach and both of them improve performance over previous state-of-the-art instructional image editing methods. Our work demonstrates instructional image editing with human feedback is a variable approach to align image generation with human preference, thus unlocking new opportunities and potential to scale up the model capabilities towards more powerful applications such as conversational image editing. While our method demonstrates impressive performance, we have also identified failure scenarios, as discussed in Sec. 4.2. In addition, it is possible that our trained model inherits bias and suffers from harmful content from pretrained foundation models such as Stable Diffusion, GPT3 and BLIP. These limitations would be considered when interpreting our results, and we expect red teaming with human feedback to mitigate some of the risks in future work.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 1

[2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022. 3

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 1, 3

[4] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. 3

[5] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European Conference on Computer Vision*, pages 707–723. Springer, 2022. 3

[6] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 4

[7] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 1, 2, 3, 4, 6, 7

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilyva Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2, 3

[9] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021. 5

[10] Paul Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *NeurIPS*, 2017. 3

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 14

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021. 3

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 13

[14] Rinon Gal, Or Patashnik, Haggai Maron, Amit Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics*, 41(4):1–13, 2022. 6

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 2014. 3

[16] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1, 2, 3, 8

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 3

[18] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in atari. *NeurIPS*, 2018. 3

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 4

[20] Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. In *Advances in Neural Information Processing Systems*, 2021. 5

[21] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. *URL https://arxiv. org/abs/2206.00364*, 2022. 2

[22] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2426–2435, 2022. 3

[23] Diederik Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[24] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020. 5

[25] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 3

[26] Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018. 5

[27] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020. 5

[28] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for uni-

fied vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 3, 4, 13

[29] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *arXiv:2301.07093*, 2023. 3

[30] Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. Magicmix: Semantic mixing with diffusion models. *arXiv preprint arXiv:2210.16056*, 2022. 1, 3

[31] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022. 2

[32] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3

[33] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 3

[34] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[35] OpenAI. Chatgpt. `https://openai.com/blog/chatgpt/`, . 2, 3

[36] OpenAI. Openaiapi. `https://platform.openai.com/docs/guides/fine-tuning`, . 3

[37] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. 2, 3, 4, 5

[38] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019. 5

[39] Jan Peters, Katharina Mulling, and Yasemin Altun. Relative entropy policy search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1607–1612, 2010. 5

[40] Andre Susano Pinto, Alexander Kolesnikov, Yuge Shi, Lucas Beyer, and Xiaohua Zhai. Tuning computer vision models with task rewards. *arXiv preprint arXiv:2302.08242*, 2023. 3

[41] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. In *NeurIPS*, 2023. 3

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 3, 15

[43] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831. PMLR, 2021. 3

[44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 3

[45] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text-to-image synthesis. In *ICML*, 2016. 3

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 2, 3, 4

[47] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 3

[48] Jeremy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. Training language models with language feedback. *arXiv preprint arXiv:2204.14146*, 2022. 3

[49] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2111.02114*, 2021. 3, 15

[50] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2, 3

[51] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of International Conference on Machine Learning*, pages 2256–2265, 2015. 3

[52] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020. 2

[53] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[54] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. *NeurIPS*, 2020. 3

[55] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. *arXiv preprint arXiv:2211.12446*, 2022. 1, 3

[56] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420*, 2023. 3

[57] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*, 2021. 2

[58] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023. 3

[59] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018. 3

[60] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 3

[61] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 3

[62] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 3

[63] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 4