A Implementation Details

A.1 Pixel Diffusion Model

For all experiments we use the pixel diffusion model DeepFloyd IF [33], as opposed to more common latent diffusion models. This is because the frequency subband, color space, and motion decompositions are not meaningful in latent space. For example, averaging channels in latent space does not correspond to an interpretable image manipulation. Interestingly, using our method to construct hybrid images with a latent diffusion model, by blurring latent codes, works to an extent but is easily susceptible to artifacts (see Appendix F), so we opt to use a pixel diffusion model which is more consistent and principled.

A.2 Hybrid Images

DeepFloyd IF [33] generates images in two stages. First at a resolution of 64×64 and then at 256×256. Because of this, we adopt the convention that our σ values are specified for the 64×64 scale, and are scaled by $4\times$ for the 256×256 images. We use a relatively large kernel size of 33 at both scales to minimize edge effects. We use σ values ranging from $\sigma = 1.0$ to $\sigma = 3.0$ for all hybrid images except for those in Fig. 3, in which we sweep the value of σ.

A.3 Triple Hybrids

Triple hybrids are quite difficult to synthesize, and as such we manually select the sigma values and prompts to generate high-quality samples. Specifically, we use σ_1 values from $\sigma_1 = 0.8$ to $\sigma_1 = 1.0$ and σ_2 values from $\sigma_2 = 1.2$ to $\sigma_2 = 2.0$ for all triple hybrids in Fig. 1 and Fig. 14.

A.4 Upscaling

DeepFloyd IF additionally uses a third stage which upscales from 256×256 to 1024×1024. We also use this stage, but because it is a latent model, we do not apply our method. We upscale using only the prompt corresponding to the highest frequency component or the color component.

B Human Studies

We use Amazon Mechanical Turk for the human study. 77 "master workers" were asked the following questions for each hybrid image pair:

- "Which image shows `[prompt_1]` clearer?"
- "Which image shows `[prompt_2]` clearer?"
- "Which image is of higher quality?"

For low frequency prompt questions, we downsample the images accordingly in order to help participants more easily see the content. For the high frequency prompt questions, as well as the quality questions, we display the images at full resolution. Participants were shown 8 hybrid image pairs in a random order.

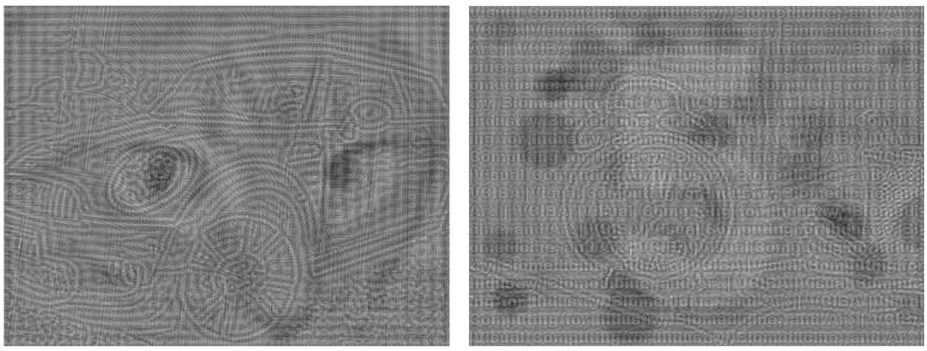

Fig. 10: Prior Work on Triple Hybrid Images. We show the triple hybrid results from prior work [59], which adapts the classic method of [46]. A description of what should be seen is provided underneath each image, going from high to low frequencies. As can be seen, these results are of lower quality than our results.

C Prior Triple Hybrid Methods

Prior work [59] attempts to create triple hybrid images by adapting the method of Oliva *et al.* [46]. As can be seen in Fig. 10, the results are not of high visual quality, and it can be hard to identify the three different subjects in the image, especially when compared to our results. This reflects the difficulty of creating these images.

D Metrics Implementation

In Tab. 2, we report the max CLIP score over multiple image downsampling factors. Specifically, for each hybrid image we downsample and then upsample by a factor f, where we choose f to be a linear sweep of 20 values between 1 and 8. These images are then preprocessed to a size of 224×224, which is the input resolution of the CLIP ViT-B/32 model which we use. We then take the normalized dot product between each resulting image embedding, and the text embedding for the corresponding prompt, and report the max. We report the max to account for the fact that different hybrid images are best seen at different downsampling factors.

E Connection to MultiDiffusion

In Sec. 4.2 we explore Factorized Diffusion with a spatial decomposition, and show that it allows targeting of prompts to specific spatial regions. We claim that this is a special case of MultiDiffusion [4]. MultiDiffusion updates a noisy image of arbitrary size by removing the consensus of multiple noise estimates over the image. Factorized Diffusion, with a spatial decomposition, also removes

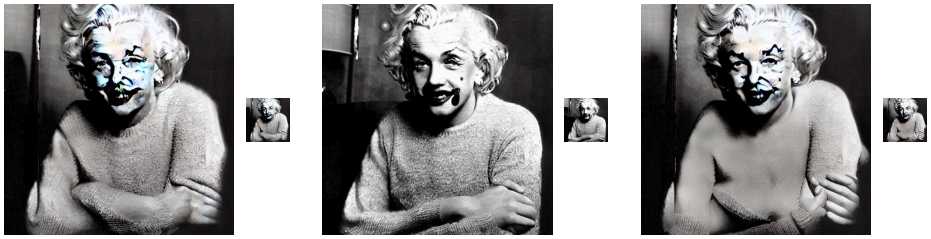

Fig. 11: Latent Hybrid Images. We provide hybrid image results using our method with Stable Diffusion v1.5, a latent diffusion model. As can be seen the results are passable, but suffer from artifacts, due to applying blurring and bandpass operations in the latent space.

a consensus of multiple noise estimates. However, in our setup this consensus is formed specifically by the disjoint union of multiple noise estimates, and our method operates only at the resolution for which the diffusion model is trained, as opposed to MultiDiffusion.

F Hybrid Images with Latent Diffusion Models

We show hybrid images resulting from using our method with Stable Diffusion v1.5, a latent diffusion model, in Fig. 11. As can be seen the results are recognizable, but have significant artifacts, due to applying bandpass filters in the latent space. We find that pixel diffusion models produce much higher quality samples.

G Synthesizing Hybrid Images with other Methods

We also attempt to generate hybrid images using two recent methods: Visual Anagrams [21] and Diffusion Illusions [7]. Results can be seen in Fig. 12. Both methods fail, which we describe and analyze below.

Diffusion Illusions works by minimizing an SDS [49] loss over multiple views of an image, paired with different prompts. We use the same high and low pass views as above. As can be seen in Fig. 12 the method produces a decent version of the low pass prompt, but fails to incorporate any of the high pass prompt. We believe this is because taking the high pass of an image moves it significantly out-of-distribution, rendering the SDS gradients unhelpful. Low passing an image alters its appearance, but keeps it relatively in-distribution, so as a result the method can still produce the low pass prompt.

Visual Anagrams works by denoising multiple transformations of an image, paired with different prompts. We use a high pass and low pass transformation, but this fails because these operations change the statistics of the noise in the noisy image. As a result, the diffusion model is being fed out-of-distribution images, and the reverse process fails to converge, as shown in Fig. 12.

Fig. 12: Other Illusion Methods. We attempt to create hybrid images using Visual Anagrams [21] and Diffusion Illusions [7], two recent methods designed to generate optical illusions. As can be seen, both methods fail. Please see Appendix G for analysis.

Table 3: Comparison to Visual Anagrams [21]. We use [21] to synthesize hybrids and color hybrids, and report the same metrics as [21]. We use prompt pairs built from the CIFAR-10 classes, with 10 prompts per pair for a total of 900 samples. Our method performs consistently better, as [21] is not designed to produce these kinds of illusions.

Task	Method	$\mathcal{A}\uparrow$	$\mathcal{A}_{0.9}\uparrow$	$\mathcal{A}_{0.95}\uparrow$	$\mathcal{C}\uparrow$	$\mathcal{C}_{0.9}\uparrow$	$\mathcal{C}_{0.95}\uparrow$
Hybrid Images	Visual Anagrams [21]	0.226	0.237	0.240	0.500	0.520	0.525
	Ours	**0.237**	**0.263**	**0.271**	**0.536**	**0.630**	**0.651**
Color Hybrids	Visual Anagrams [21]	0.223	0.232	0.234	0.500	0.537	0.547
	Ours	**0.231**	**0.260**	**0.269**	**0.512**	**0.562**	**0.586**

Finally, we also quantitatively evaluate hybrid and color hybrids generated using Geng *et al*. [21] against our proposed method, with results shown in Tab. 3. As prompts, we use all pairs of CIFAR-10 classes, and sample 10 images per prompt pair for a total of 900 samples. We use the same metrics as [21], and we find that our method does better consistently, as [21] was not designed to generate these illusions.

H Further Analysis of Factorized Diffusion

As discussed in Sec. 3.3, our analysis assumes that the update step is a linear combination of the noisy image, $\mathbf{x}_t$, and the noise estimate, ϵ_θ. However, many commonly used update steps also involve adding random noise $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$, such as DDPM [28]. To deal with this, we can view the update step as a composition of two steps:

$$\mathbf{x}_{t-1} = \texttt{update}(\mathbf{x}_t, \epsilon_\theta) \tag{21}$$

$$= \texttt{update'}(\mathbf{x}_t, \epsilon_\theta) + \sigma_z \mathbf{z}. \tag{22}$$

The first step is a linear combination of $\mathbf{x}_t$ and ϵ_θ, and the second adds in the noise $\mathbf{z}$. Our analysis then applies to just the `update'` function.

I Choosing Prompts

We find that carefully choosing prompts can generate higher quality illusions. For example, the success rate and quality of samples are much higher when at least one prompt is of a "flexible" subject, such as `"houseplants"` or `"a canyon"`. In addition, we found biases specific to decompositions. Prompts with the style `"photo of..."` performed better for hybrid and motion hybrid images. We suspect this is because photos tend to have ample amounts of both high and low frequency content, as opposed to styles such as `"oil paintings"` or `"watercolors"`, which tend to lack higher frequency content. For color hybrids, we found that using the style of `"watercolor"` produced better results, perhaps because of the style's emphasis on color.

Fig. 13: Colorization. Our method can also be used to solve inverse problems, such as colorization. We show grayscale images that we wish to colorize on the left. The color component is then generated conditioned on the text prompts displayed. Note that this is effectively prior work [12,32,58].

J Colorization

We also show colorization results in Fig. 13, using our method as an inverse problem solver, as discussed in Sec. 3.5. Specifically, we use the color space decomposition introduced in Sec. 3.4. During diffusion model sampling we hold

the grayscale component fixed to the grayscale component of a real image that we want to colorize, and generate the color component. Note that this is effectively prior work [12, 32, 58].

K Additional Results

In this section, we provide additional qualitative results. Additional results for hybrid images and triple hybrids are shown in Fig. 16 and Fig. 14 respectively. In Fig. 15 and Fig. 17, we provide more examples of motion and color hybrids, respectively. Finally, we provide more random samples for hybrid images, color hybrids, and motion hybrids in Fig. 18.

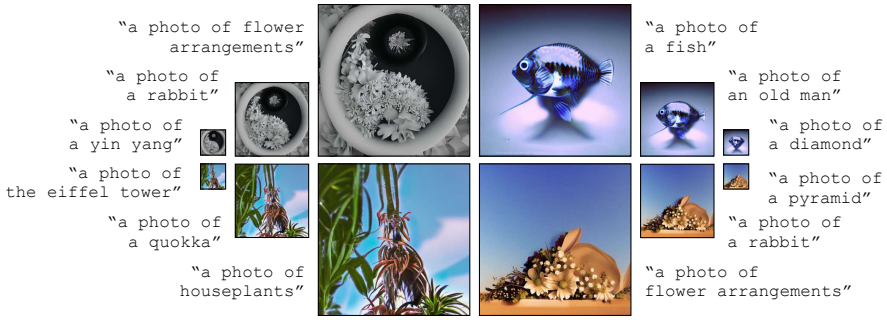

Fig. 14: Triple Hybrids. We provide more *triple hybrid* results. ***Best viewed digitally, using zoom.***

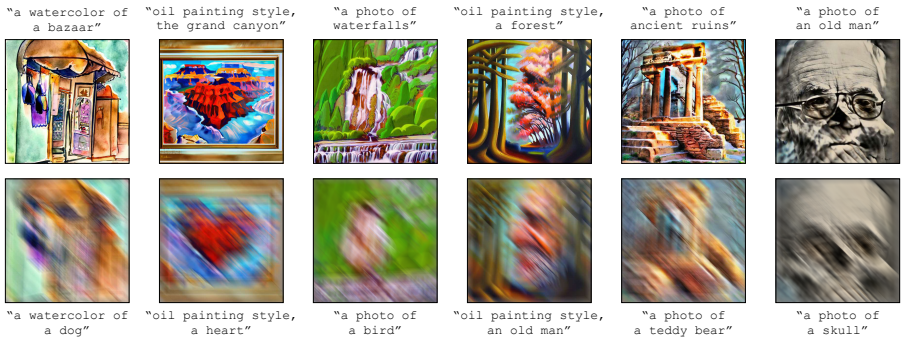

Fig. 15: Motion Hybrids. We show more *motion hybrid* results. These are images that change appearance when motion blurred. Here, the motion is from upper left to bottom right.

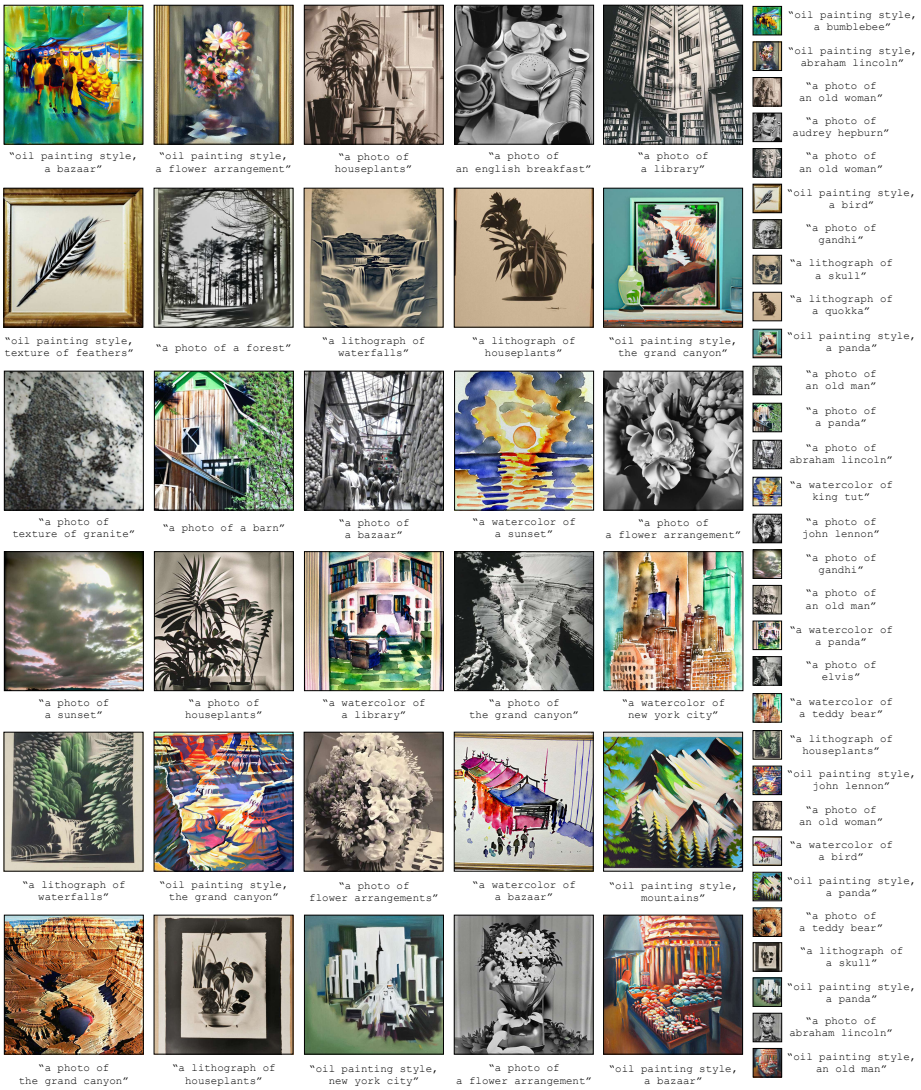

Fig. 16: Hybrid Images. We show more *hybrid image* results. For easier viewing, we provide insets of each hybrid image at lower resolution, along with the corresponding prompt. ***Best viewed digitally, with zoom.***

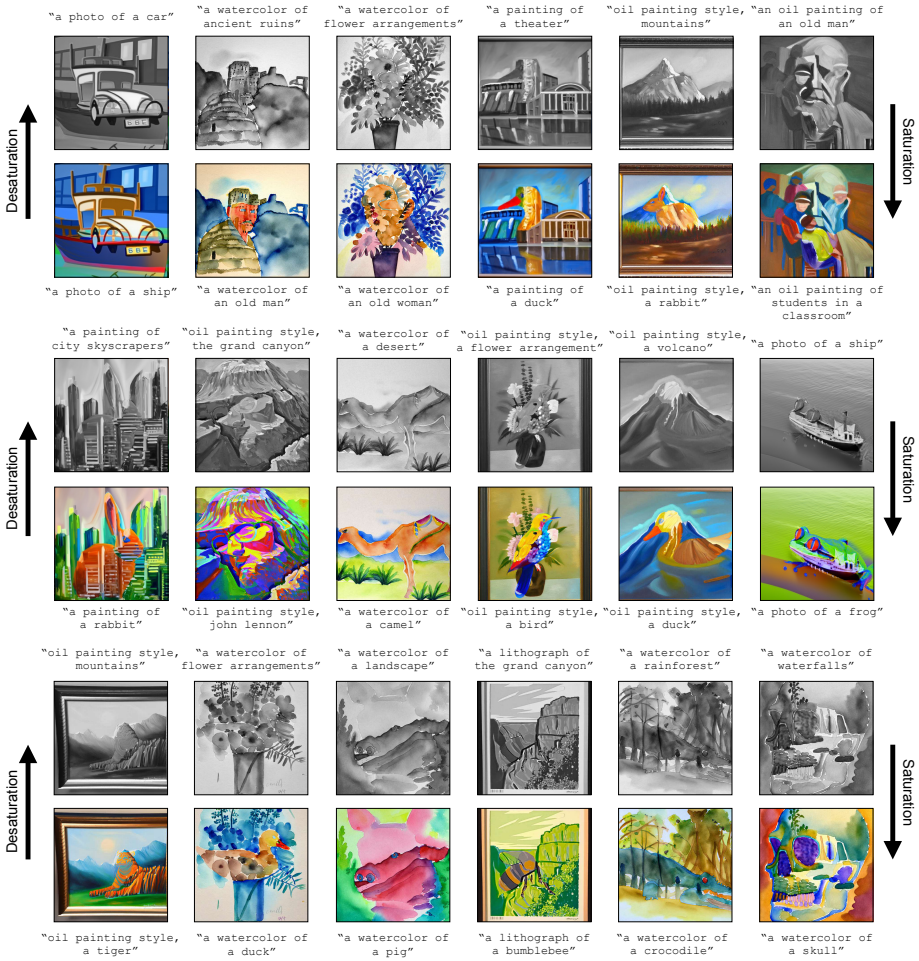

Fig. 17: Color Hybrids. We show more *color hybrid* results, with grayscale images placed above their colorized version.

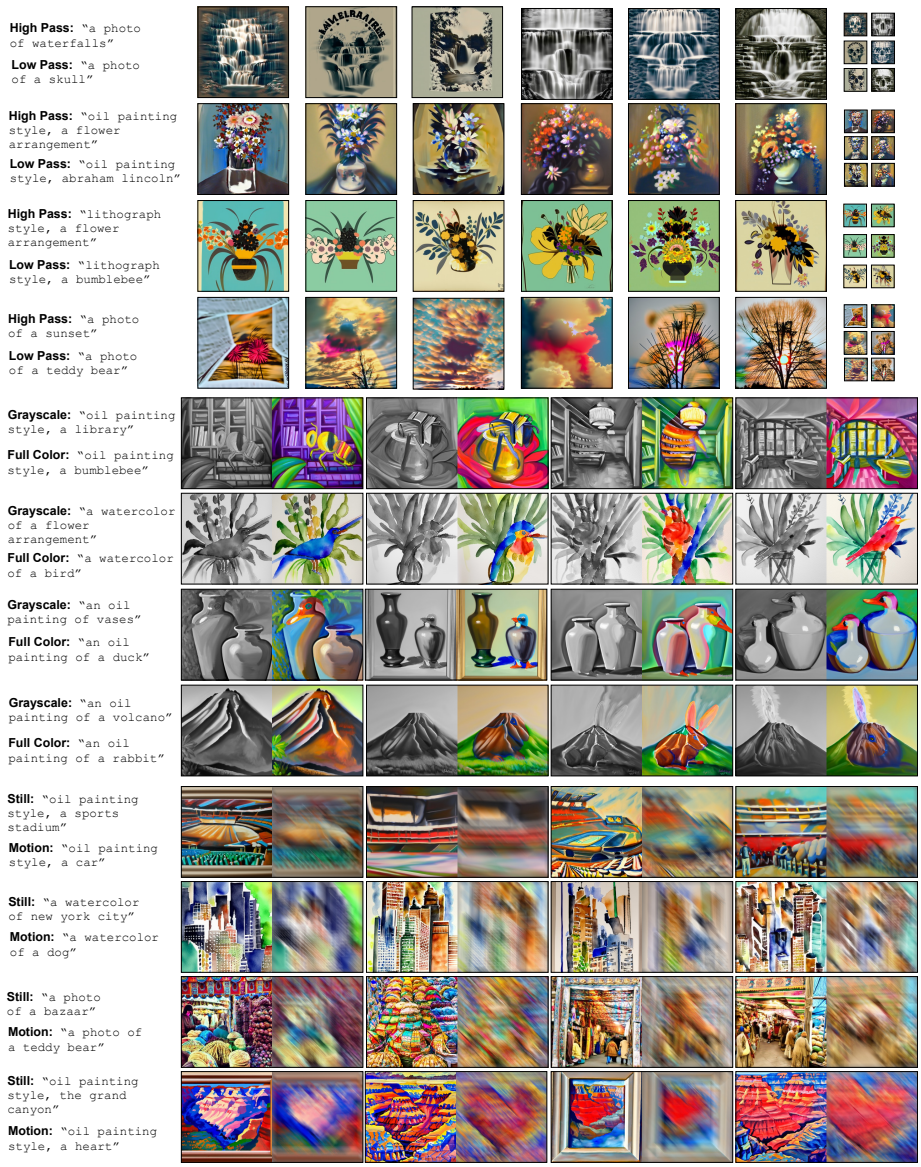

Fig. 18: Random Samples. We provide random samples of hybrid images, color hybrids, and motion hybrids for selected prompts.

L CVPR 2024 T-Shirt Design

We created an inverse hybrid image for the official CVPR 2024 T-shirt as part of the AI Art track. Our goal was for attendees to see only a watercolor of the Seattle skyline when they received the shirt. Then, as they see other people wearing the shirts in the conference center from a distance, the text "CVPR" would be revealed.

We took an existing photo of the Seattle skyline, and pasted the text "CVPR" over the image. We then used our technique to condition an image on the low frequencies of the edited photo, and fill in the high frequencies given the text prompt **`"a watercolor of the seattle skyline with mount rainier in the background"`**. The resulting image was then touched up by running Adobe Photoshop's generative fill in a few locations with artifacts to improve quality. We show the low frequency image and the hybrid image, before editing, in Fig. 19. We also show additional candidate T-shirt designs in Fig. 20, which all reveal the text "CVPR" when viewed from a distance.

Fig. 19: CVPR Hybrid Image T-shirt Design. a) The edited photo, from which we extract low spatial frequencies. **b)** The resulting hybrid image, after generating high spatial frequencies conditioned on the extracted low frequencies. For more details, please visit our website. Photo source: Pavol Svantner [60].

Fig. 20: CVPR Hybrid Image T-shirt Design Candidates. We show more CVPR T-shirt designs. For easier viewing, we provide insets of each hybrid image at lower resolution. ***Best viewed digitally, with zoom.***