Guided Denoising Diffusion for Prompt to Prompt Editing

Valeria Avino[1], Leonardo Rocci[2], and Riccardo Soleo[3]

[1]Department of Computer Science, University of La Sapienza

September 15, 2025

Abstract

This report details the implementation of advanced editing techniques based on **guided denoising diffusion models**, specifically utilizing the Stable Diffusion framework. We replicate the **Prompt-to-Prompt (P2P)** image editing method, which leverages cross-attention control to enable fine-grained manipulation of generated images. To extend this capability to real-world images, we implement **Null-text Inversion (NTI)**, a two-step process involving DDIM inversion and null-text embedding optimization. Furthermore, we demonstrate the versatility of this approach by applying the P2P technique to audio, treating spectrograms as images to perform **prompt-to-prompt audio editing** using the Riffusion model. This report highlights how these methodologies provide superior control and content preservation compared to traditional editing paradigms like SDEdit or global model steering. The code implementation can be found at `https://github.com/vaal4ds/Prompt-to-Prompt-Editing-with-SD`

1 Introduction

The rise of generative models, particularly diffusion models, has revolutionized content creation across various modalities. While these models excel at synthesizing novel content from text prompts, the challenge of editing and manipulating existing data remains complex. Traditional methods often fall short, either requiring explicit manual input (e.g., masks for inpainting) or failing to preserve the original content's structural and semantic integrity. This work explores a more sophisticated approach by intervening in the core mechanics of the diffusion process itself. By manipulating the cross-attention mechanism, which links the text prompt to the image's spatial features, we gain a precise and semantically aware tool for editing.

This report formalizes our implementation of advanced editing techniques based on guided denoising diffusion models, specifically utilizing the Stable Diffusion framework. We replicate the Prompt-to-Prompt (P2P) image editing method [1], which leverages cross-attention control to enable fine-grained manipulation of generated images. To extend this capability to real-world images, we implement Null-text Inversion (NTI) [2], a two-step process involving DDIM inversion and null-text embedding optimization. Furthermore, we demonstrate the versatility of this approach by applying the P2P technique to audio, treating spectrograms as images to perform prompt-to-prompt audio editing using the Riffusion model, which is a novel implementation of the core concepts. The report underscores the power of guided denoising diffusion and its ability to provide superior control across modalities.

2 Previous Work and Motivation

Prior to the attention-based editing methods, two common approaches were employed for content manipulation with diffusion models. The first is **SDEdit** (Stochastic Differential Editing) a technique that leverages the forward and reverse diffusion processes for image manipulation. Unlike unconditional generation which begins denoising from a pure noise latent, SDEdit starts from a partially noised version of the input image, x_t, obtained by applying a fixed number of forward diffusion steps to the original image, x_0. This controlled noising process embeds the source content into the model's latent space without completely destroying its structural integrity.

The subsequent denoising pass, guided by a new text prompt, is then executed with a crucial constraint. For localized edits, a user-provided binary mask, M, is used to delineate regions of interest. At each denoising step, the model's prediction is forced to preserve the original content in the unmasked area. This can be mathematically expressed by a projection-like step:

$$x_{t-1} \leftarrow (1 - M) \odot x_0 + M \odot x'_{t-1}$$

where x'_{t-1} is the denoised latent from the model, and $\odot$ denotes the element-wise Hadamard product. This ensures the unmasked regions remain faithful to the original image while the masked areas are freely reconstructed according to the new prompt.

While effective for some tasks, this reliance on an explicit mask makes the process less intuitive and more labor-intensive compared to prompt-based methods.

The second approach, **model steering**, operates on the principle that the internal activations of the text encoder can be manipulated to "steer" the model's output toward a desired concept. In order to compute a "direction" to follow in the latent space to go from generic latents to latents that embed the desired concept we defined a steering vector as a weighted average of activations differences from "negative" to "positive" prompts. A key enhancement was the use of weighted averaging to better capture the concept to inject, as shown in the following formula for the steering vector t_l at layer l:

$$t_l = \frac{1}{|D|} \sum_{i \in D} w_i \cdot (A_i(x^+) - A_i(x^-))$$

Here, $A_i(x^+)$ and $A_i(x^-)$ are the activation maps for the positive (x^+) and negative (x^-) prompts in the i-th pair from a dataset D, and w_i is a weight assigned to each pair.

During the image generation process, the calculated steering vector is injected by adding it to the original activation map $A_{i,l}(x)$ for each layer l:

$$A_{i,l}(x) \leftarrow A_{i,l}(x) + \epsilon \cdot t_l$$

The steering strength, ϵ, controls the magnitude of this effect. A key optimization strategy is to use an incremental steering strength, starting with a low ϵ in the early layers of the text encoder and gradually increasing it toward the final layers. This respects the hierarchical nature of the text encoder, where early layers capture low-level features and later layers capture more abstract information.

However, our tests revealed that this method is best suited for global, conceptual changes and often fails to preserve the intricate details of the original content. The inability to perform localized, semantic edits while maintaining the original image's pose, lighting, and non-prompted elements motivates the need for a more granular and precise control mechanism. The P2P method directly addresses this limitation by focusing on the attention layers, which are responsible for spatial layout and content.

3 Prompt-to-Prompt Editing

Prompt-to-Prompt (P2P) editing is a powerful technique for manipulating generated images by intervening in the diffusion model's core generative process. It is built upon the foundational principles of text-guided diffusion.

3.1 Text-Guided Diffusion

The foundation of modern text-to-image models is the text-guided diffusion process. The generative model operates by learning to reverse a gradual noising process. This is fundamentally divided into two parts: the forward process and the backward (or denoising) process.

The **forward diffusion process** gradually adds Gaussian noise to a data point (e.g., an image) over a series of T time steps. This process is defined as a Markov chain:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

where x_0 is the original data point, x_t is the noisy version at step t, and β_t is the noise schedule. The state at any time step t can be directly sampled from the original data x_0:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I)$$

where $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$.

The **backward (denoising) process** is the generative part. A neural network, typically a U-Net, is trained to predict the noise added at each step, $\epsilon_\theta(x_t, t)$. Using this prediction, the model iteratively reverses the forward process to generate a clean image from pure noise. The denoising step is formulated as:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right)$$

where $\alpha_t = 1 - \beta_t$. For text-guided generation, the model's noise prediction ϵ_θ is conditioned on a text prompt P.

3.2 Classifier-Free Guidance (CFG)

To enhance the model's adherence to the text prompt, **Classifier-Free Guidance (CFG)** is used during the denoising process. This technique combines a conditioned noise prediction with an unconditioned one. The model computes two noise predictions at each step: one guided by the prompt P, $\epsilon_\theta(z_t, t, P)$, and one unconditioned (guided by a null-text prompt, $\emptyset$), $\epsilon_\theta(z_t, t, \emptyset)$. The final guided noise prediction, ϵ_{CFG}, is a linear combination of these two, following the formula:

$$\epsilon_{CFG}(z_t, t, P) = \epsilon_\theta(z_t, t, \emptyset) + w(\epsilon_\theta(z_t, t, P) - \epsilon_\theta(z_t, t, \emptyset))$$

Here, z_t is the noisy latent representation at time step t, and w is the guidance weight. A higher value of w pushes the generation to align more closely with the text prompt, with a typical value of $w = 7.5$ providing a good balance between prompt adherence and image diversity.

3.3 Cross-Attention Mechanism and P2P Formulations

The core of the P2P method lies in its ability to manipulate the cross-attention maps within the Stable Diffusion U-Net. As illustrated in Figure 1, the cross attention maps represent the interaction between pixels and text, combining the spatial features of the latent representation of the image with the semantic information of the text prompt. The deep spatial features of the noisy latent image $\phi(z_t)$ are projected to a query matrix $Q = \ell_Q(\phi(z_t))$, and the textual embedding $\psi(P)$ is projected to a key matrix $K = \ell_K(\psi(P))$ and a value matrix $V = \ell_V(\psi(P))$ via

learned linear projections ℓ_Q, ℓ_K, ℓ_V. The cross-attention matrix $M_t \in R^{S \times L}$ at time t is then calculated by multiplying Q with K^T, where S is the number of spatial tokens and L is the number of text tokens. This matrix is then used to weight the value vectors (V).

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The P2P method introduces three core operations on these attention maps to control the editing process:

1. **Attention Replacement (*Rep*)**: Used for replacing one word in a prompt with another (e.g., changing "cat" to "dog"). This operation replaces the attention map of the source prompt (M_t) with the attention map of the target prompt (M_t^*) at early time steps, which are crucial for defining the image's overall structure.

$$Rep(M_t, M_t^*, t) = \begin{cases} M_t^* & \text{if } t < \tau, \\ M_t & \text{otherwise.} \end{cases}$$

   Here, τ is a time threshold that controls when the replacement occurs. This ensures that the global structure is determined by the new prompt while later, fine-grained details are kept consistent.

2. **Attention Reweighting (*Rew*)**: Used to change the emphasis of a specific word (e.g., from "a cloudy sky" to "a very cloudy sky"). This operation multiplies the attention values of a specific word (j^*) by a constant factor c, thereby increasing or decreasing its influence on the image generation.

$$Rew(M_t, M_t^*)_{i,j} = \begin{cases} w \cdot (M_t^*)_{i,j} & j = j^* \\ (M_t)_{i,j} & \text{otherwise.} \end{cases}$$

   The attention map for the source prompt M_t is used for all words except the one being reweighted, which uses the attention map of the new prompt M_t^*.

3. **Attention Refinement (*Ref*)**: Used for adding a new word to the prompt (e.g., "a cat" to "a fluffy cat"). This operation incorporates the attention of the new word without removing the structure established by the original words. This is achieved by using the attention maps from the new prompt only for the newly added tokens, and relying on a precomputed mapping '$A(j)$' from the original to the new tokens.

$$\text{Ref}(M_t, M_t^*, t)_{i,j} = \begin{cases} (M_t^*)_{i,j} & A(j) = \text{None} \\ (M_t)_{i,A(j)} & \text{otherwise.} \end{cases}$$

   The term $A(j) = \text{None}$ corresponds to the tokens in the new prompt that do not have a counterpart in the original prompt.

4. **Localized Edits (Attention-Based Blending)** To constrain edits to specific regions without relying on an external mask, the P2P method introduces a concept called local blend. This technique uses the cross-attention maps themselves to implicitly define the regions of interest. By analyzing the attention map of a specific word (or words), we can identify which pixels in the image are most influenced by that word. This creates a "soft mask" directly from the prompt's semantics.

   The core idea is to apply a linear combination of the denoising latents from both the source prompt (L) and the edited target prompt (L^*), weighted by this soft mask. The final blended latent, $L_{blended}$, at each time step t is calculated as:

$$L_{blended}(t) = \alpha L(t) + (1 - \alpha)L^*(t)$$

   where α is a blending weight derived from the attention map of the word(s) to be kept unchanged. This ensures that regions of the image not related to the edited words are preserved, while the target latent guides the change in the relevant areas. This allows for precise, localized edits without the need for manual inpainting or masking.

   The implementation of such strategies is showcased in figures 2,3, 4.

3.3.1 Choosing Layers for Control

The U-Net architecture of a diffusion model is composed of downsampling, mid-block, and upsampling layers. It is typical to target the **mid-block and upsampling layers**.

- The **mid-block** processes the smallest spatial resolution, capturing high-level, global compositional features. Intervening here allows for broad, structural changes.
- The **upsampling layers** progressively increase the spatial resolution and are responsible for adding fine-grained details. Controlling these layers enables more localized edits, preserving the overall composition while changing specific elements.

Manipulating both sets of layers ensures a balance between global coherence and local fidelity. The 'AttentionStore' utility is used to collect and apply these manipulations across the chosen layers in a coordinated manner.

4 Prompt-to-Prompt on Real Images with DDIM Text Inversion

To successfully apply the P2P method to a real image, we must first find a latent representation and a noise trajectory that accurately reconstructs the input image. This

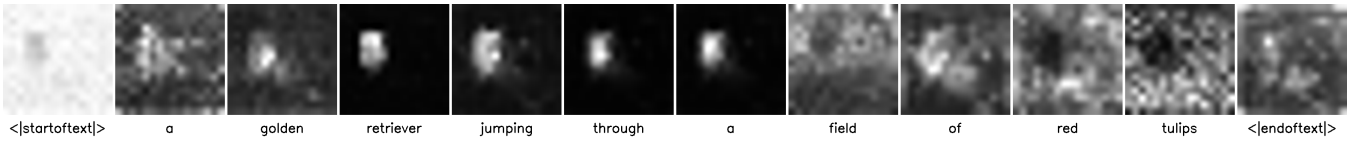

Figure 1: Visualization of the Cross-Attention Mechanism in action on the image generated from the prompt "A golden retriever jumping through a field of red tulips". Average attention masks for each word in the prompt which was used to synthesize the left image, highlighting the region-to-word correspondence.

is a non-trivial **inversion** problem. While fine-tuning the network's weights or optimizing the entire textual encoding are possible, they are often inefficient or result in non-interpretable representations. Instead, our approach exploits the key feature of Classifier-Free Guidance: the result is highly affected by the unconditional prediction. We propose to optimize only the null-text embedding, keeping the model weights and conditional textual embedding unchanged.

This methodology, referred to as **Null-text Inversion (NTI)**, provides high-quality reconstruction while still allowing for intuitive editing with Prompt-to-Prompt. It is a two-step process:

1. **DDIM Inversion**: We use a deterministic DDIM (Denoising Diffusion Implicit Models) scheduler to invert the forward diffusion process. A simple inversion technique for the DDIM sampling, which can be seen as reversing the ODE process, is given by:

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(z_t - \sqrt{1-\alpha_t} \epsilon_\theta(z_t, t) \right)$$

   By reversing the forward process $z_0 \to z_T$, we obtain a sequence of noisy latent codes $z_0, ..., z_T$ starting from the encoded real image z_0.

2. **Null-text Optimization**: The crucial second step is to find an optimal "null-text" embedding, e_u, that best reconstructs the image. We optimize a different "null embedding" e_t for each timestamp t, which significantly improves the reconstruction quality. This optimization minimizes the difference between a DDIM sampling step using a specific null-text embedding e_t and the inverted latent from the previous step. The optimization objective is:

$$\min_{e_t} \| z_{t-1}(z_t, e_t, C) - z_{t-1} \|_2^2$$

   For simplicity, $z_{t-1}(z_t, e_t, C)$ denotes applying a DDIM sampling step using latent z_t, the unconditional embedding e_t, and the conditional embedding C. This process is performed for timestamps from $T-1$ down to 1.

The full algorithm for Null-text Inversion is presented below. This approach results in a latent state and a set of unconditional embeddings that, when used for standard generation, perfectly reconstruct the original image, providing the necessary foundation for applying P2P edits.

5 Prompt-to-Prompt on Audios using Spectrograms

The P2P method's core principle of manipulating attention maps can be extended beyond the visual domain. For audio, we treat the spectrogram—a 2D visual representation of sound frequencies over time—as the "image" input to a diffusion model. We leverage the Riffusion model, which is a Stable Diffusion variant fine-tuned on music spectrograms.

The process for audio editing is as follows:

1. **Audio-to-Spectrogram Conversion**: An input audio file is converted into a high-resolution spectrogram image. This conversion is handled by a dedicated 'SpectrogramImageConverter'.

2. **Spectrogram Editing with P2P**: The spectrogram is then passed through the P2P framework, where we apply the same attention control mechanisms based on a new text prompt (e.g., changing "heavy metal" to "jazz"). The model edits the spectrogram's visual features, which correspond to changes in frequency, rhythm, and timbre.

3. **Spectrogram-to-Audio Conversion**: The edited spectrogram image is then converted back into an audio file, yielding an audio clip with the desired edits.

This demonstrates the power and flexibility of guided denoising diffusion models, confirming that their core principles are modality-agnostic and can be applied to any data that can be represented in a compatible format.

6 Conclusion

This report demonstrates a robust and versatile set of techniques for controlled content editing using guided denoising diffusion models. By intervening directly in the cross-attention mechanism, the P2P method provides a level of semantic and spatial control that surpasses traditional methods. The Null-text Inversion technique successfully extends this capability to real images, solving the complex problem of finding an invertible representation. Finally, the application to audio via spectrograms showcases the generalizability of these principles. The methodologies detailed herein offer a powerful new paradigm for

Algorithm 1: Null-text Inversion

Input: A source prompt embedding $C = \psi(P)$ and input image I.

Output: Noise vector z_T and optimized embeddings $\{e_t\}_{t=1}^{T}$.

Set guidance scale $w = 1$.

Compute the intermediate results $z_T \dots z_0$ using DDIM inversion over I.

Set guidance scale $w = 7.5$.

Initialize $z_T = z_T$ and $e_T = \psi(\text{null\_text})$.

for $t = T$ **down to** 1 **do**

  for $j = 0$ **to** $N - 1$ **do**

    $e_t \leftarrow e_t - \eta \cdot \nabla_{e_t} \| z_{t-1}(z_t, e_t, C) - z_{t-1} \|_2^2$

  end for

  Set $z_{t-1} = z_{t-1}(z_t, e_t, C)$

end for

Return $z_T, \{e_t\}_{t=1}^{T}$.

content creation and editing, blending the creative freedom of generative models with the precision required for practical applications.

7 Gallery

This section contains visual and auditory examples of the techniques described in this report.

References

[1] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. 2022.

[2] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models, 2022.

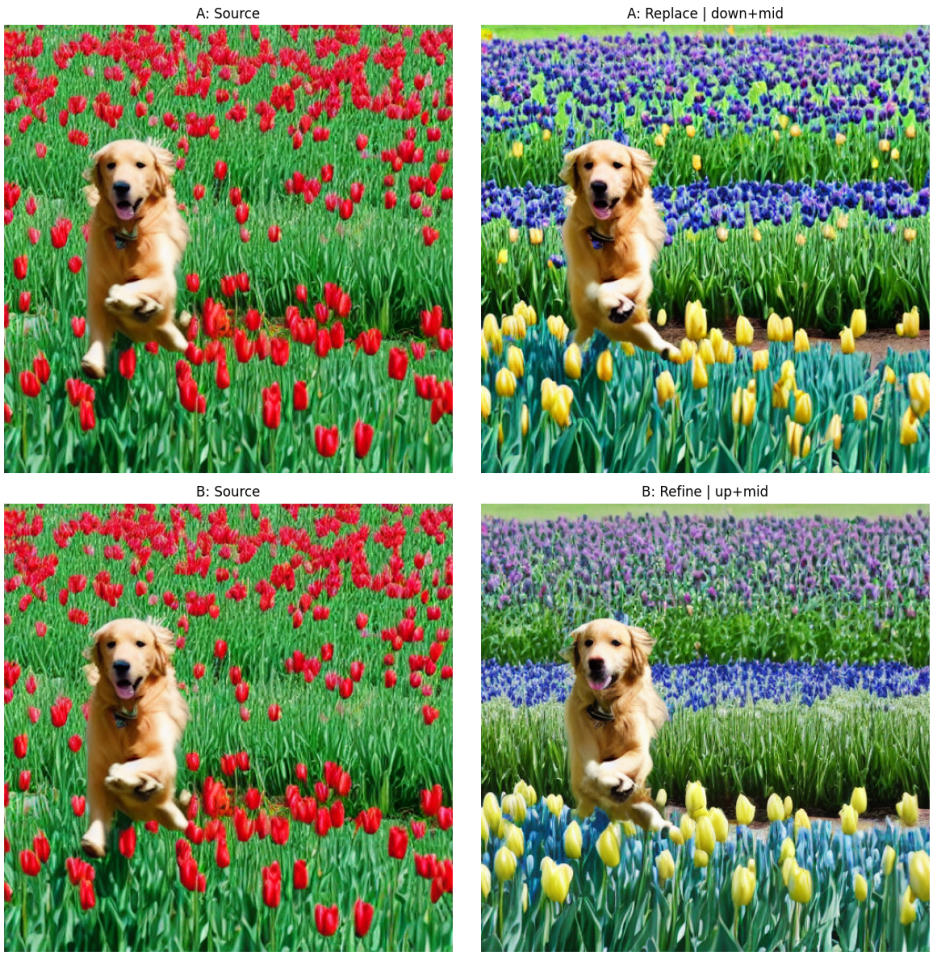

Figure 2: **Prompt to Prompt Reweight vs Refine:** On the left the generated image from the source prompt "A golden retriever jumping through a field of red tulips", on the right the p2p-edited image corresponding to the prompt: "A golden retriever jumping through a field of blue tulips". Up: applying the replacing strategy on the down-sampling blocks provides more structural change. Down: the refining strategy on the middle and up-sampling blocks provides a softer blend with the original image, thus results in a mixture of red and blue, violet.

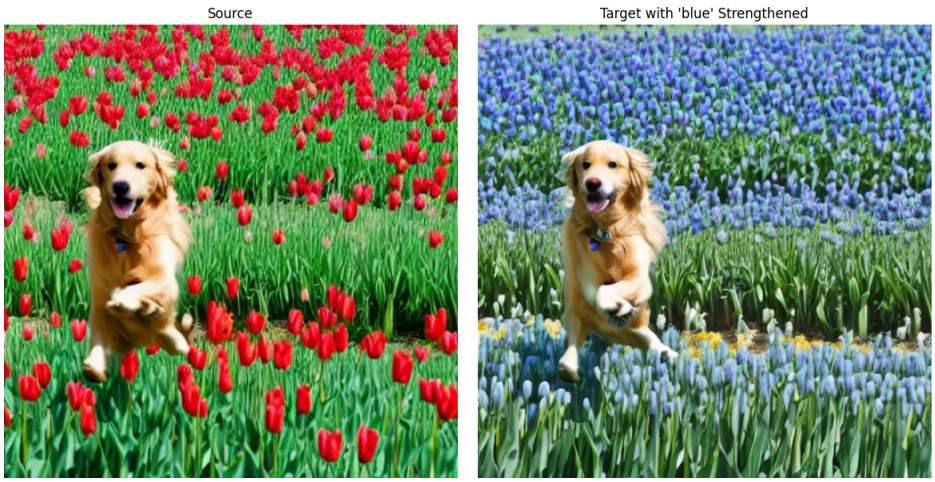

Figure 3: **Reweighting to strengthen a feature:** Example of enhancing one feature using the reweighting strategy on the token associated to "blue", by multiplying for a weight> 1.

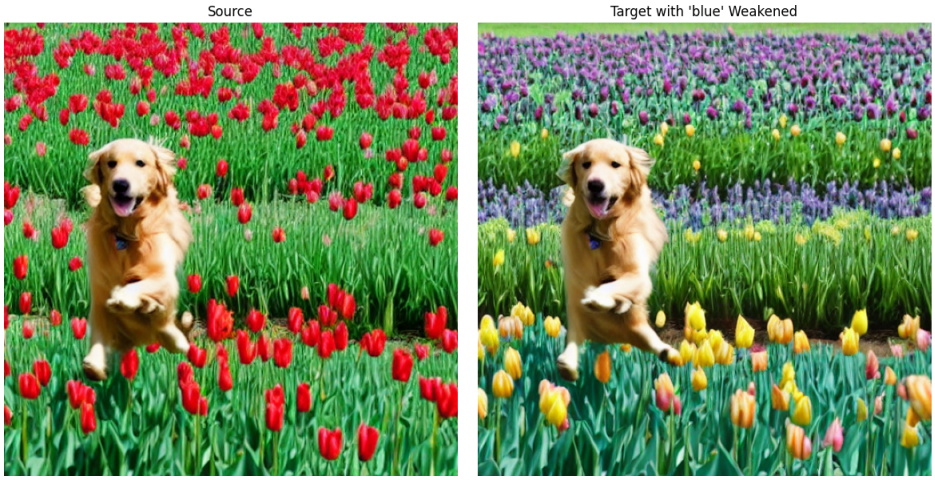

Figure 4: **Reweighting to weaken a feature:** Example of weakening one feature using the reweighting strategy on the token associated to "blue", by multiplying for a weight< 1.

Figure 5: **Prompt Refinement:** The left panel shows the source image generated from the prompt *"photorealistic landscape with a house near the river"*, while the right panel shows the target image generated from the refined prompt *"photorealistic landscape with a house near the river and a rainbow in the background"*. The refinement adds a visible rainbow in the sky, while the overall structure of the house and surrounding vegetation remains nearly unchanged, except for minor details below the rainbow.

Figure 6: **Style Transformation:** The left panel shows the source image generated from the prompt *"a photo of a forest"*, while the right panel shows the target image generated from the refined prompt *"Children drawing of a forest"*. The transformation converts the photorealistic forest into a child-style drawing, yet the overall arrangement of trees and bushes is preserved, so the scene remains clearly recognizable despite the change in artistic style.

Figure 7: **Word Swapping:** The left panel shows the source image generated from the prompt *"a photorealistic photo of a sunflower in a field"*, while the right panel shows the target image generated from the refined prompt *"a photorealistic photo of a rose in a field"*. All sunflowers in the original scene are replaced by roses, while the overall layout of the field and the spatial positions of the flowers remain unchanged.

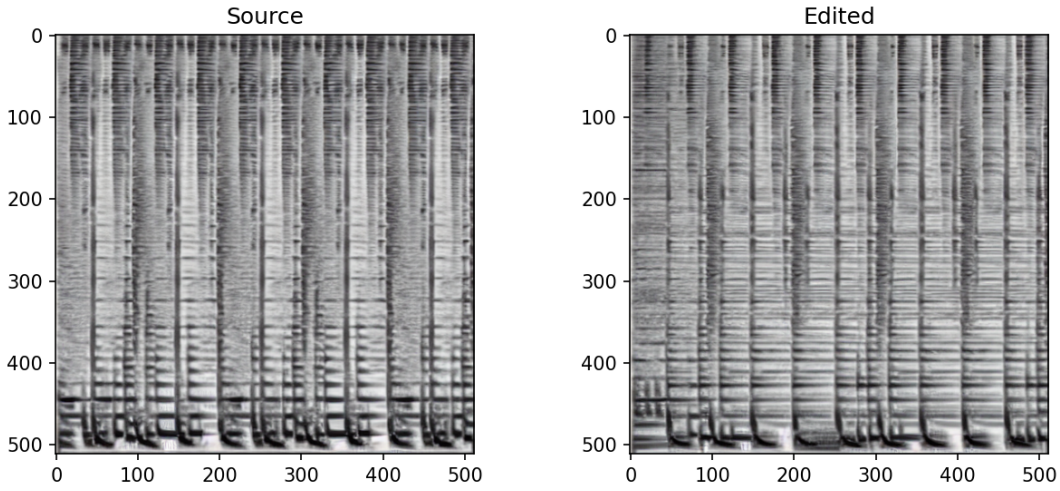

Figure 8: **Audio Prompt-to-Prompt:** A spectrogram of an original audio clip and its prompt-to-prompt edited version, showing changes corresponding to a new prompt Source prompt: "acid techno beat" (left), target prompt: "real tekno beat" (right).

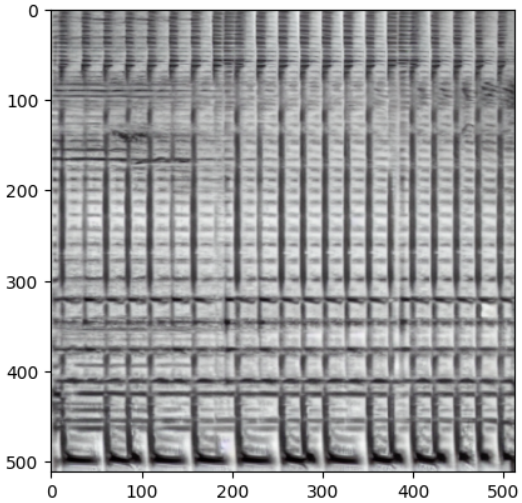

Figure 9: **Without Prompt-to-Prompt:** The spectrogram of a freshly generated target audio from the target prompt: "real tekno beat", showing how generation with no attention mechanism fails to be a honest edit of the original audio.

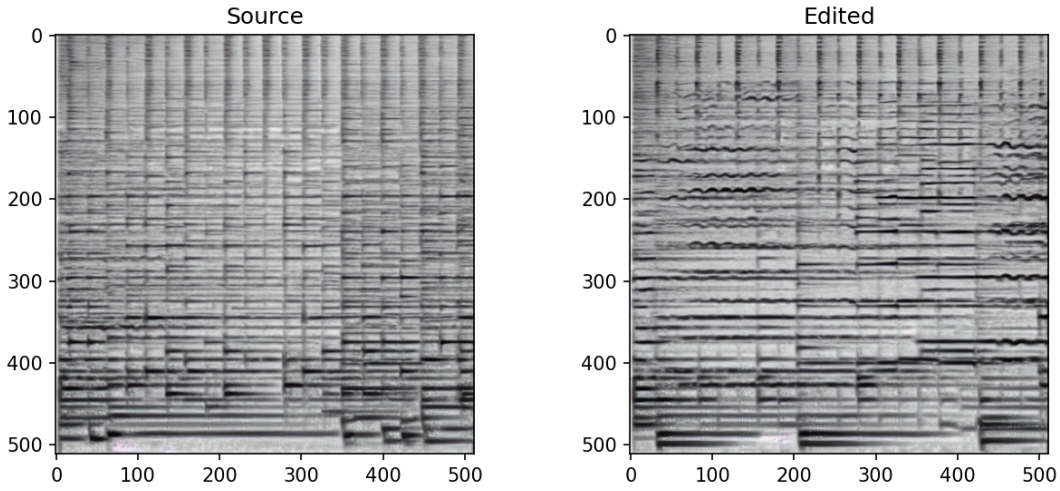

Figure 10: **Audio Prompt-to-Prompt with Local Blend:** A spectrogram of an original audio clip and its prompt-to-prompt edited version using Local Blend, showing changes corresponding to a new prompt Source prompt: "solo piano clean melody only piano, no vocals" (left), target prompt: "solo violin clean melody only violin, no vocals" (right).

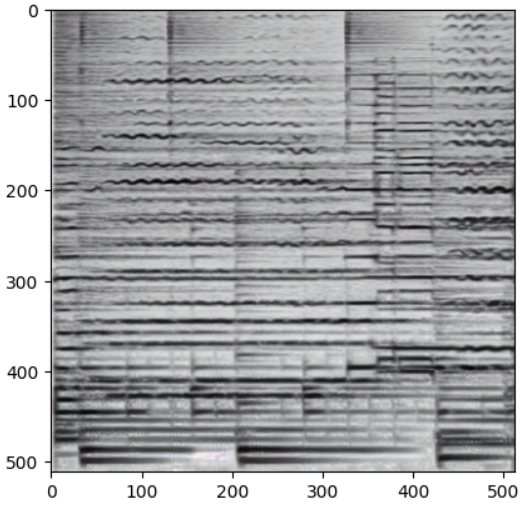

Figure 11: **Without Prompt-to-Prompt:** The spectrogram of a freshly generated target audio from the target prompt: "solo violin clean melody only violin, no vocals".

Figure 12: **Style Transformation:** A real photo of a pineapple plant is transformed into a drawing. The conversion preserves not only the objects in the foreground and background but also the structural details of the building windows and the surrounding vegetation, making the drawing closely adhere to the original scene.

Figure 13: **Geometric Editing:** Example of targeted shape modification: the pot of the pineapple plant is changed from circular to square while the rest of the image remains unaffected, demonstrating precise local editing.

Figure 14: **Multiple Subject Replacement:** The top–left panel shows the original photo of a kitten beside a mirror. In the remaining panels the kitten is successively replaced with, from left to right and top to bottom: a tiger, a small dog, a bulldog, a ferret, a rabbit, and a silver and gold sculpture of the original cat. The edits consistently affect both the primary subject and its reflection, highlighting the ability of the technique to replace the subject even in the mirrored image.