

## A Broader Impact

Our model’s capability to fuse images and text to generate new and creative object images holds significant potential across various fields, including entertainment, design, and education. However, it also raises important considerations regarding content safety and ethical use. In particular, if the input image or text contains inappropriate or offensive material, the generated images may similarly be inappropriate, leading to potentially unpleasant experiences for users.

To mitigate these risks, it is crucial to implement robust NSFW (Not Safe For Work) content detection mechanisms. While existing methods can address some cases of inappropriate content, we acknowledge the need for continuous improvement in this area. As part of our future work, we will incorporate advanced NSFW checking models to ensure the generated content adheres to safety standards and ethical guidelines. This proactive approach aims to safeguard users and promote responsible use of our image generation technology.

## B Limitation

Our method relies on the semantic correlation between the original and transformed content within the diffusion feature space. When the semantic match between two categories is weak, our method tends to produce mere texture changes rather than deeper semantic transformations. This limitation suggests that our approach may struggle with transformations between categories with weak semantic associations. Future work could focus on enhancing semantic matching between different categories to improve the generalizability and applicability of our method.

There are still some failure cases in our model, as shown in Fig. 9. These failures can be categorized into two types. The first row illustrates that when the content of the image is significantly different from the text prompt, the changes become implicit. The second row demonstrates that in certain cases, our adaptive function results in changes that only affect the texture of the original image. In our future work, we will investigate these situations further and analyze the specific items that do not yield satisfactory results.



Figure 9: Failure results of our ATIH model.

## C Text and Image Categories.

We selected 60 texts, as detailed in Table 4, and categorized them into 7 distinct groups. The 30 selected images are shown in Fig. 10, with each image corresponding to similarly categorized texts, as outlined in Table 5. Our model is capable of fusing content between any two categories, showcasing its strong generalization ability.

Table 4: List of Items by Category

Category	Items
Mammals	kit fox, Siberian husky, Australian terrier, badger, Egyptian cat, cougar, gazelle, porcupine, sea lion, bison, komondor, otter, siamang, skunk, giant panda, zebra, hog, hippopotamus, bighorn, colobus, tiger cat, impala, coyote, mongoose
Birds	king penguin, indigo bunting, bald eagle, cock, ostrich, peacock
Reptiles and Amphibians	Komodo dragon, African chameleon, African crocodile, European fire salamander, tree frog, mud turtle
Fish and Marine Life	anemone fish, white shark, brain coral
Plants	broccoli, acorn, brain coral
Fruits	strawberry, orange, pineapple, zucchini, butternut squash
Objects	triceratops, beach wagon, beer glass, bowling ball, brass, airship, digital clock, espresso maker, fire engine, gas pump, grocery bag, harp, parking meter, pill bottle, zucchini

Table 5: Origin Image Categories

Category	Items
Mammals	Sea lion, Dog (Corgi), Horse, Squirrel, Sheep, Mouse, Panda, Koala, Rabbit, Fox, Giraffe, Cat, Wolf, Bear
Birds	Owl, Duck, Bird
Insects	Ladybug
Plants	Tree, Flower vase
Fruits and Vegetables	Red pepper, Apple
Objects	Cup of coffee, Jar, Church, Birthday cake
Human	Man in a suit
Artwork	Lion illustration, Deer illustration, Twitter logo

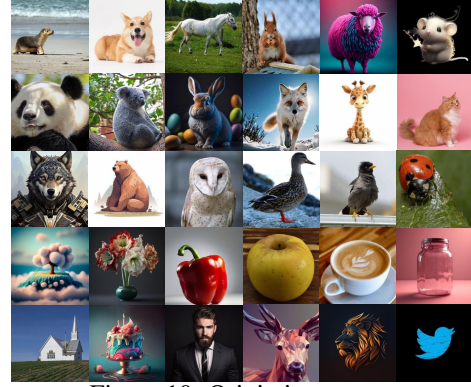


Figure 10: Origin images set

## 515 D Parameter Analysis.

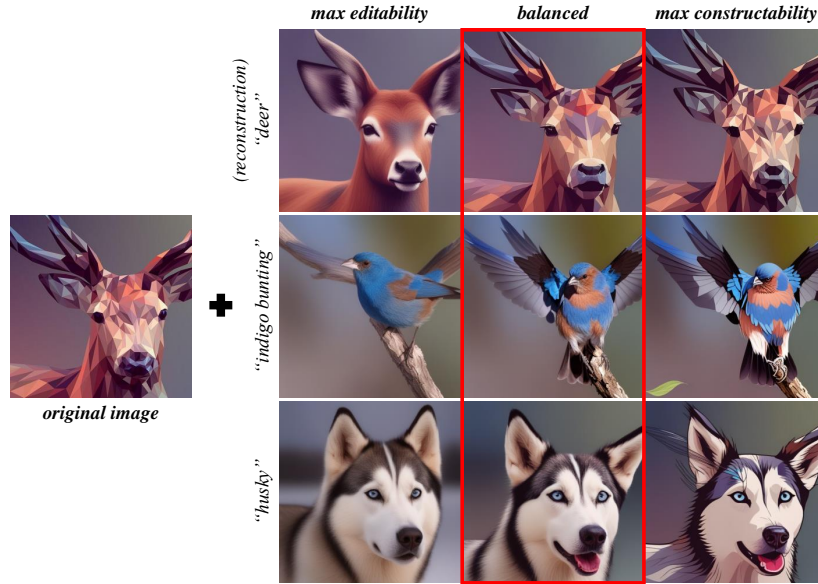


Figure 11: Image changes

516 **Analysis of  $\lambda$ .** Here, we provide a detailed explanation of the  
517 determination of  $\lambda$ . As shown in Fig. 11, we use the ratio  
518  $\lambda = \frac{L_r}{L_n}$  to balance editability and fidelity. We iteratively adjust  
519 this ratio in the range of  $[0, 400]$  with intervals of 10, measuring  
520 the Dino-I score between the reconstructed and original images,  
521 as well as the CLIP-T and AES scores for images directly edited  
522 with the inverse latent values at different ratios. These experi-  
523 ments were conducted on the class fusion dataset, using fusion  
524 text for direct image editing. Figs. 12, 13, and 14 indicate that as the ratio increases, image editability improves,  
525 peaking at a ratio of around 260, but with a decrease in quality. At a ratio of 125, both image fidelity and the  
526 AES score achieve an optimal balance. Therefore, we set  $\lambda$  to 125.

527 **Analysis of  $k$ .** The experimental analysis of parameter  $k$  was conducted using sdxturbo as the base model.  
528 The range for  $i$  was set to  $[0, 4]$ , and for each value of  $i$ ,  $\alpha$  was iterated from 0 to 2.2 in steps of 0.02 to  
529 observe changes in the fused image. The averaged experimental results produced a smooth curve, as shown in  
530 Fig.4. Based on these observations, the optimal range for  $k$  was determined to be between  $[2.1, 2.7]$ . In our  
531 experiments, we set the value of  $k$  to 2.3.

532 **Analysis of  $I_{sim}^{\min}$  and  $I_{sim}^{\max}$ .** As shown in Fig. 15, we visualized several specific node images generated during  
533 the variation of different  $\alpha$  factor values. When the image similarity with the original image exceeds 0.85, the

Table 6: Quantitative comparison results with different  $\lambda$ .

$\lambda$	AES $\uparrow$	CLIP-T $\uparrow$	Dino-I $\uparrow$
0	6.116	0.413	0.927
125	<b>6.153</b>	0.417	0.902
260	6.012	0.419	0.760

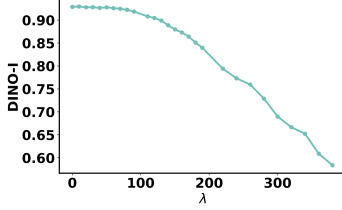


Figure 12: Dino-I changing with  $\lambda$

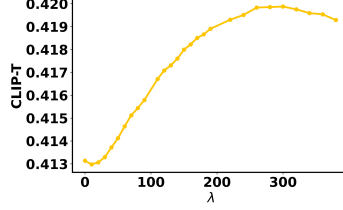


Figure 13: CLIP-T score changing with  $\lambda$

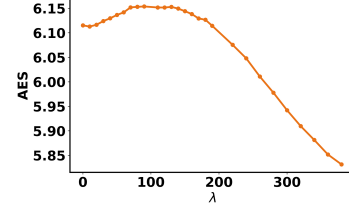


Figure 14: AES changing with  $\lambda$

images become overly similar. For example, in the dog-zebra fusion experiment, the dog’s texture remains largely unchanged, and no zebra features are visible. Conversely, when the image similarity falls below 0.45, the images overly conform to the text description. In this case, the entire head of the image turns into a zebra, representing an over-transformation phenomenon. Based on these observations, we set the minimum similarity threshold  $I_{sim}^{\min}$  to 0.45 and the maximum similarity threshold  $I_{sim}^{\max}$  to 0.85. This range helps us achieve a good balance between retaining original image information and integrating text features.



Figure 15: Similarity changes

## E Ablation Study.

We present another set of ablation study results in Fig. 16, where the two rows represent the cases without (w/o) and with (w) attention projection. The input image is a Corgi, and the text is Fire engine. The output images display the different transformations as  $\alpha$  varies. The top row shows the abrupt change in appearance without attention projection, resulting in a sudden transition from a Corgi to a fire engine. In contrast, with attention projection (bottom row), the change is smoother, achieving the desired blending result in the middle.



Figure 16: Results changing in Iteration w/ and w/o attention injection.

## F Algorithm.

Overall, our **novel object synthesis** comprises three key components: optimizing the noise  $\epsilon_t$  through a balance of fidelity and editability loss, adaptively adjusting the injection step  $i$ , and dynamically modifying the factor  $\alpha$ . These processes are detailed in Algorithm 1. Additionally, we utilize the Golden Section Search method to

---

**Algorithm 1** Novel Object Synthesis

---

```
1: Input: An object image  $O_I$ , a target prompt  $O_T$ , the number of inversion steps  $T$ , inject step  $i$ ,  
   sampled noise  $\epsilon_t$ , scale factor  $\alpha$ ,  $F(\alpha)$  is Eq.(9)  
2: Output: Object Synthesis  $O$   
3:  $\{z_T, \dots, \hat{z}'_{t-1}, \dots, z_0\} \leftarrow \text{scheduler\_inverse}(z_0)$   
4: for  $t = 1$  to  $T$  do  
5:    $\hat{z}_{t-1} \leftarrow \text{step}(\hat{z}_t)$   
6:    $\epsilon_{all}[t] \leftarrow \text{Balance-fidelity-editability}(\hat{z}_{t-1}, \hat{z}'_{t-1}, \hat{z}_t, \epsilon_t)$   
7: end for  
  
8:  $i_{init} \leftarrow T/2$   
9:  $i_{final} \leftarrow \text{Adjust-Inject}(z_T, \epsilon_{all}, O_T, i_{init})$   
10:  $\alpha_{good} \leftarrow \text{Golden-Section-Search}(F, \alpha_{min}, \alpha_{max})$   
11:  $O \leftarrow \text{DM}(z_T, \epsilon_{all}, O_T, i_{final}, \alpha_{good})$   
12: return  $O$   


---

13: function BALANCE-FIDELITY-EDITABILITY( $\hat{z}_{t-1}, \hat{z}_{t-1}, \hat{z}'_{t-1}, \epsilon_t$ )  
14:   while  $\mathcal{L}_r/\mathcal{L}_n > \lambda$  do  
15:      $\epsilon_t \leftarrow \epsilon_t - \nabla_{\epsilon_t} \mathcal{L}_r(\hat{z}_{t-1}, \hat{z}'_{t-1}, \epsilon_t, \hat{z}_t)$   
16:   end while  
17:   return  $\epsilon_t$   
18: end function  


---

19: function GOLDEN-SECTION-SEARCH( $F, a, b$ )  
20:    $\phi \leftarrow \frac{1+\sqrt{5}}{2}$  ▷ Golden ratio  
21:    $c \leftarrow b - \frac{b-a}{\phi}$   
22:    $d \leftarrow a + \frac{b-a}{\phi}$   
23:   while  $|b-a| > \epsilon$  do  
24:     if  $f(c) < f(d)$  then  
25:        $b \leftarrow d$   
26:     else  
27:        $a \leftarrow c$   
28:     end if  
29:      $c \leftarrow b - \frac{b-a}{\phi}$   
30:      $d \leftarrow a + \frac{b-a}{\phi}$   
31:   end while  
32:   return  $\frac{b+a}{2}$   
33: end function  


---

34: function ADJUST-INJECT( $z_T, \epsilon_{all}, i, O_T$ )  
35:    $ite \leftarrow 0$   
36:   while  $iter < \frac{T}{2}$  do  
37:      $I_{sim} \leftarrow \text{model}_{I_{sim}}(z_T, \epsilon_{all}, i, O_T)$   
38:     if  $I_{sim} < I_{sim}^{\min}$  then  
39:        $i \leftarrow i + 1$   
40:     else if  $I_{sim}^{\min} \leq I_{sim} \leq I_{sim}^{\max}$  then  
41:        $i \leftarrow i$   
42:       break  
43:     else  
44:        $i \leftarrow i - 1$   
45:     end if  
46:      $iter \leftarrow iter + 1$   
47:   end while  
48:   return  $i$   
49: end function  


---


```

identify an optimal or sufficiently good value for  $\alpha$  that maximizes the score function  $F(\alpha)$  in Eq. (9). This approach operates independent of the function’s derivative, enabling rapid iteration towards achieving optimal harmony. The key steps of the Golden Section Search algorithm are outlined as follows:

$$\alpha_1 = b - \frac{b-a}{\phi}, \quad \alpha_2 = a + \frac{b-a}{\phi},$$

where  $\phi$  (approximately 1.618) is the golden ratio, and  $a$  and  $b$  are the current search bounds for  $\alpha$ . During each iteration, we compare  $F(\alpha_1)$  and  $F(\alpha_2)$ , and adjust the search range accordingly:

if  $F(\alpha_1) > F(\alpha_2)$  then  $b = \alpha_2$  else  $a = \alpha_1$ .

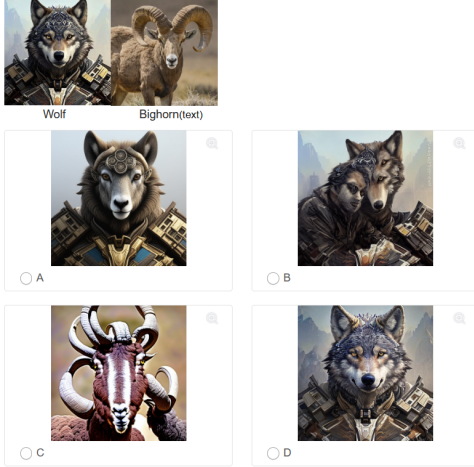
This process continues until the length of the search interval  $|b - a|$  is less than a predefined tolerance, indicating convergence to a local maximum.

## G User Study.

In this section, we delve into our two user studies in greater detail. The image results are illustrated in Figs. 6 and 7, while the outcomes of the user studies for both tasks are presented in Figs. 17 and 18. In total, we collected 570 votes from 95 participants across both studies. The specific responses for each question are detailed in Tables 7 and 8.

Notably, for the fourth question in the user study corresponding to our editing method, the example of peacock and cat fusion is shown in Fig.6, the number of votes for InfEdit [60] slightly exceeded ours. However, upon examining the image results, it becomes evident that their approach leans towards a disjointed fusion, where one half of an object is spliced with the corresponding half of another object, rather than directly generating a new object as our method does.

\* 3. Given the original image and the corresponding fusion text, please select the most remarkable fusion of two objects based on novelty, harmony, and artistic value.



\* 8. Please select your preferred fused image of the two given objects from the options below.

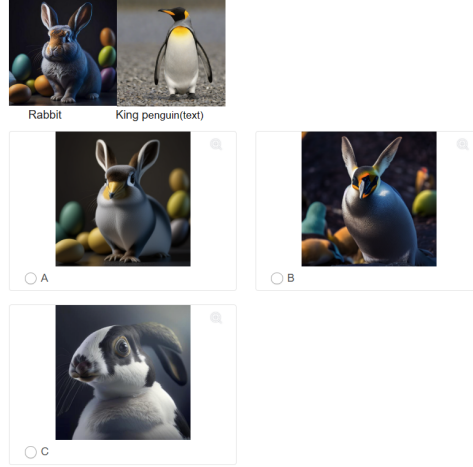


Figure 17: user study with image-editing methods. Figure 18: user study with mixing methods.

Table 7: User study with image editing methods.

image-prompt \ options(Models)	A(Our ATIH)	B(MasaCtrl)	B(InstructPix2Pix)	D(InfEdit)
glass jar-salamander	77.89 %	1.05%	16.84%	4.21%
giraffe-bowling ball	89.74 %	2.11%	6.32%	6.32%
wolf-bighorn	84.21 %	1.05%	10.53%	4.21%
cat-peacock	40 %	3.16%	5.26%	51.58%
sheep-triceraptors	78.95 %	3.16%	11.58%	6.32%
bird-African chameleon	73.68 %	6.32%	4.21%	15.79%

## H More results.

In this section, we present additional results from our model. Fig. 19 showcases further generation results using our ATIH model. We experimented with four different images, each edited with four distinct text prompts.



Table 8: User study with mixing methods.

(prompt) image-prompt	options(Models)		
	Our ATIH	B(MagicMix)	C(ConceptLab)
Dog-white shark	81.05%	2.11%	16.84%
Rabbit-king penguin	83.16%	11.58%	5.26%
horse-microwave oven	71.58%	9.47%	18.95%
camel-candelabra	86.32%	6.32%	7.37%
airship-espresso maker	71.58%	11.58%	16.84%
jeep-anemone fish	83.16%	8.42%	8.42%

570 Additionally, Fig. 20 illustrates our model’s versatility with multiple prompts, emphasizing its capability for  
571 continuous editing.

572 In Fig. 21, we compare our results with those from the state-of-the-art T2I model DALL·E3 assisted by Copilot.  
573 Our model shows superior performance when handling complex descriptive prompts for image editing. We  
574 observe that the competing model struggles to achieve results comparable to ours, particularly in maintaining the  
575 original structure and layout of images, despite adequate prompts.

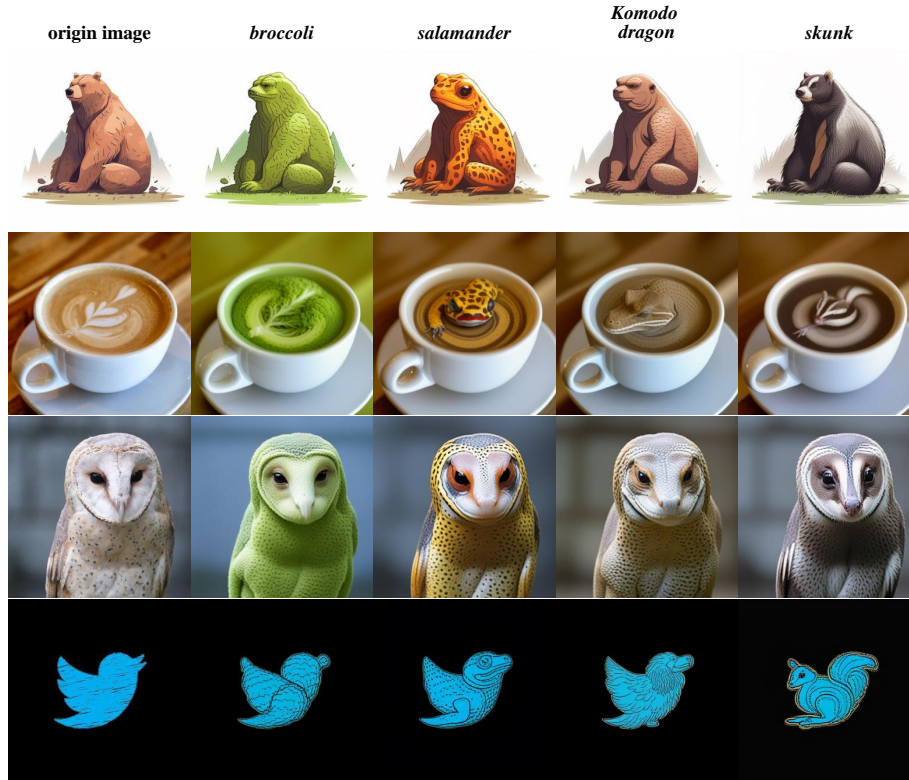


Figure 19: More visual Results.

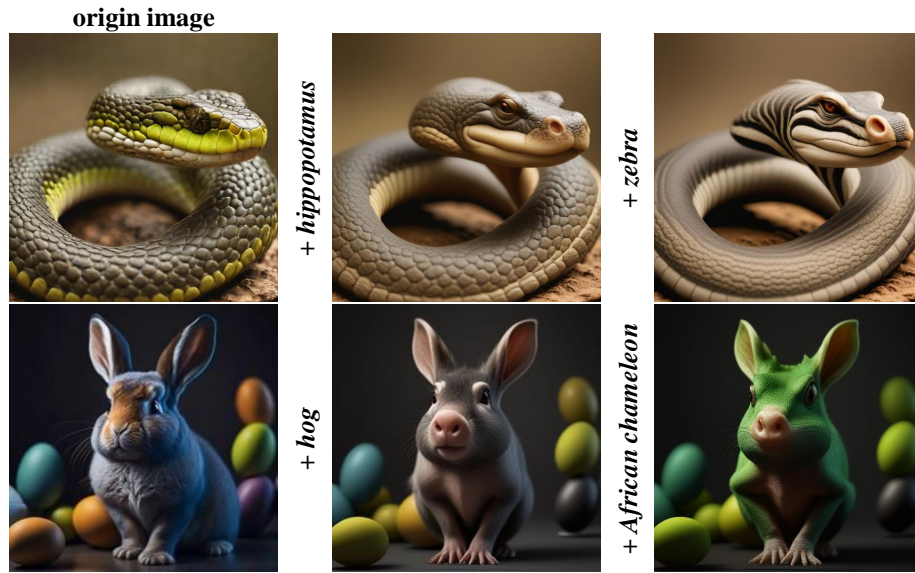


Figure 20: Fused Results With three Prompts.



Figure 21: Results comparison with complex prompt editing.