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. 11. 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.

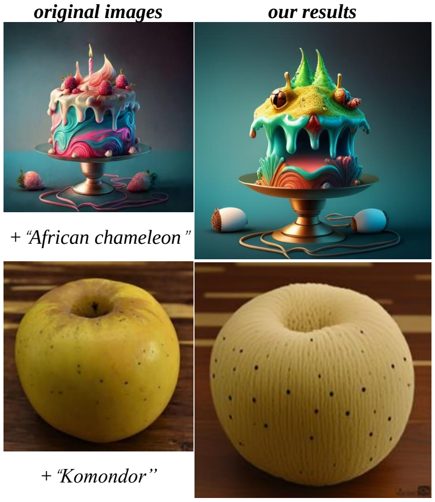


Figure 11: Failure results of our ATIH model.

C Text and Image Categories.

We selected 60 texts, as detailed in Table 5, and categorized them into 7 distinct groups. The 30 selected images are shown in Fig.12, with each image corresponding to similarly categorized texts, as outlined in Table 6. Our model is capable of fusing content between any two categories, showcasing its strong generalization ability.

Table 5: List of Text Items by Object 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
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

Table 6: Original Object 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

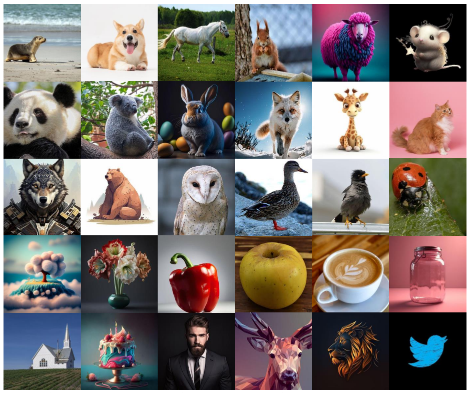


Figure 12: Original Object Image Set.

D Parameter Analysis.

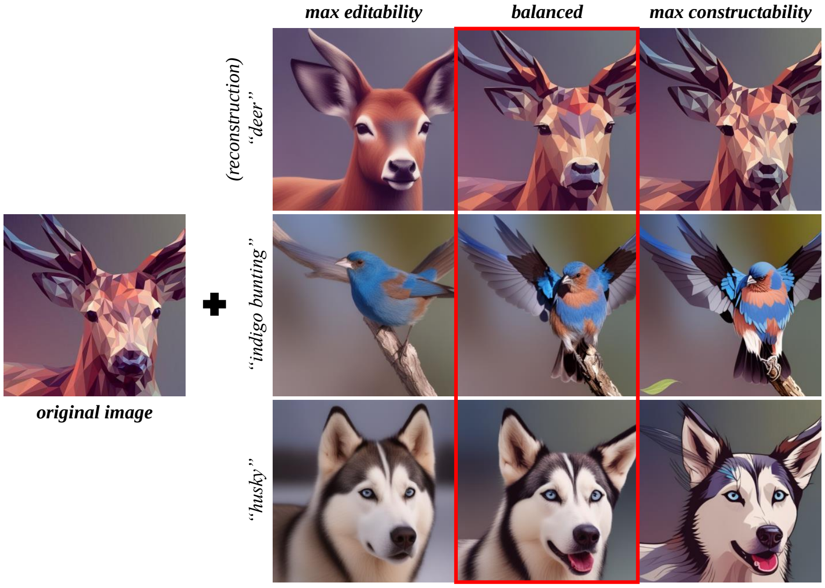


Figure 13: Image variations under different λ values. The first row displays the reconstructed images. The middle and bottom rows show the results of editing with different prompts, demonstrating variations in maximum editability, a balanced approach, and maximum constructability

Analysis of λ. Here, we provide a detailed explanation of the determination of λ. As shown in Fig. 13, we use the ratio $\lambda = \frac{L_r}{L_n}$ to balance editability and fidelity. We iteratively adjust this ratio in the range of $[0, 400]$ with intervals of 10, measuring the Dino-I score between the reconstructed and original images, as well as the CLIP-T and AES scores for images directly edited with the inverse latent values at different ratios. These experiments were conducted on the class fusion dataset, using fusion text for direct image editing. Figs. 14, 15, and 16 indicate that as the ratio increases, image editability improves, peaking at a ratio of around 260, but with a decrease in quality. At a ratio of 125, both image fidelity and the AES score achieve an optimal balance. Therefore, we set λ to 125.

Table 7: Quantitative comparison results with different λ.

λ	AES ↑	CLIP-T ↑	Dino-I ↑
0	6.116	0.413	0.927
125	**6.153**	0.417	0.902
260	6.012	0.419	0.760

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

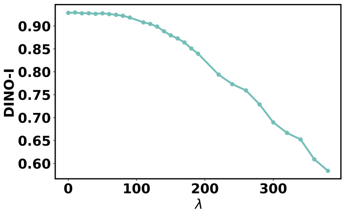


Figure 14: Dino-I changing with λ

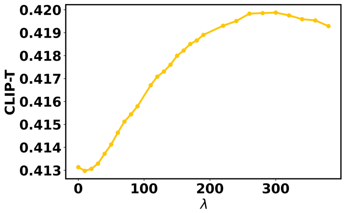


Figure 15: CLIP-T score changing with λ

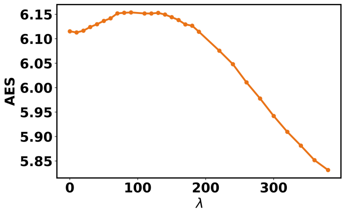


Figure 16: AES changing with λ

Analysis of $I_{\text{sim}}^{\text{min}}$ and $I_{\text{sim}}^{\text{max}}$. As shown in Fig. 17, we visualized several specific node images generated during the variation of different α factor values. When the image similarity with the original image exceeds 0.85, the 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 $I_{\text{sim}}^{\text{min}}$ to 0.45 and the maximum similarity threshold $I_{\text{sim}}^{\text{max}}$to 0.85. This range helps us achieve a good balance between retaining original image information and integrating text features.

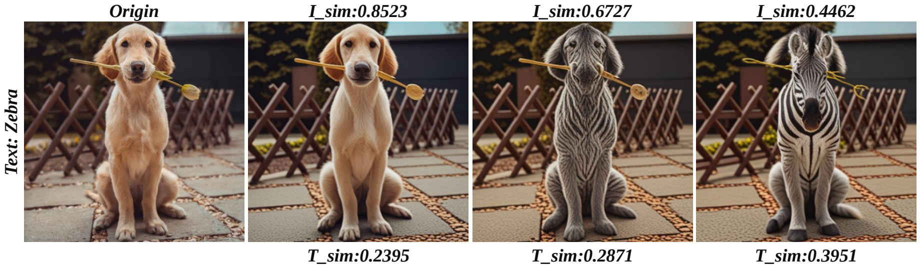


Figure 17: Illustrates the visual results of images at different similarity levels.

E Ablation Study.

We present another set of ablation study results in Fig. 18, 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 α 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.

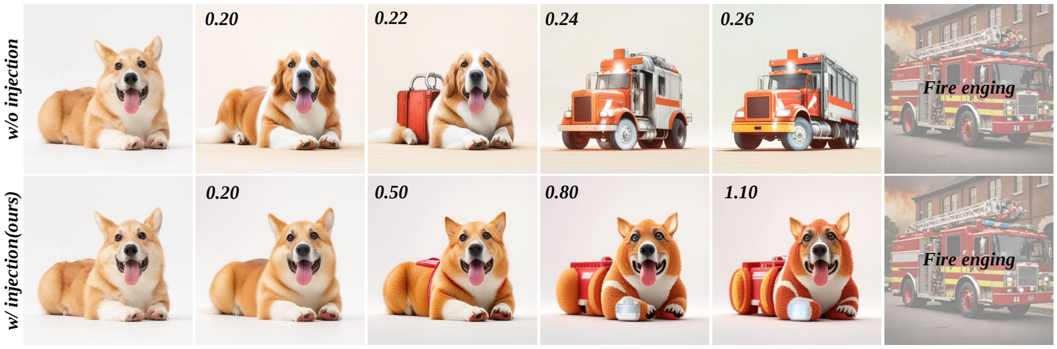


Figure 18: Results changing in Iteration w/ and w/o attention injection.

F Algorithm.

Overall, our **novel object synthesis** comprises three key components: optimizing the noise ϵ_t through a balance of fidelity and editability loss, adaptively adjusting the injection step i, and dynamically modifying the factor

Algorithm 1 Novel Object Synthesis

```
Input: An initial image latent $z_0$, 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)
Output: Object Synthesis $O$
$\{z_T, \cdots, \widehat{z}'_{t-1}, \cdots, z_0\} \leftarrow$ scheduler_inverse($z_0$)
for $t = 1$ to $T$ do
    $\widehat{z}_{t-1} \leftarrow$ step($\widehat{z}_t$)
    $\epsilon_{\text{all}}[t] \leftarrow$ Balance-fidelity-editability($\widehat{z}_{t-1}, \widehat{z}'_{t-1}, \widehat{z}_t, \epsilon_t$)
end for
$i_{\text{init}} \leftarrow T/2$
$i_{\text{final}} \leftarrow$ Adjust-Inject($z_T, \epsilon_{\text{all}}, O_T, i_{\text{init}}$)
$\alpha_{\text{good}} \leftarrow$ Golden-Section-Search($F, \alpha_{\text{min}}, \alpha_{\text{max}}$)
$O \leftarrow$ DM($z_T, \epsilon_{\text{all}}, O_T, i_{\text{final}}, \alpha_{\text{good}}$)
return $O$

function BALANCE-FIDELITY-EDITABILITY($\widehat{z}_{t-1}, \widehat{z}_{t-1}, \widehat{z}'_{t-1}, \epsilon_t$)
    while $\mathcal{L}_{\text{r}}/\mathcal{L}_{\text{n}} > \lambda$ do
        $\epsilon_t \leftarrow \epsilon_t - \nabla_{\epsilon_t}\mathcal{L}_{\text{r}}(\widehat{z}_{t-1}, \widehat{z}'_{t-1}, \epsilon_t, \widehat{z}_t)$
    end while
    return $\epsilon_t$
end function

function GOLDEN-SECTION-SEARCH($F, a, b$)
    $\phi \leftarrow \frac{1+\sqrt{5}}{2}$                                        ▷ Golden ratio
    $c \leftarrow b - \frac{b-a}{\phi}$
    $d \leftarrow a + \frac{b-a}{\phi}$
    while $|b - a| > \epsilon$ do
        if $f(c) < f(d)$ then
            $b \leftarrow d$
        else
            $a \leftarrow c$
        end if
        $c \leftarrow b - \frac{b-a}{\phi}$
        $d \leftarrow a + \frac{b-a}{\phi}$
    end while
    return $\frac{b+a}{2}$
end function

function ADJUST-INJECT($z_T, \epsilon_{all}, i, O_T$)
    $ite \leftarrow 0$
    while $iter < \frac{T}{2}$ do
        $I_{\text{sim}} \leftarrow \text{model}_{I_{\text{sim}}}(z_T, \epsilon_{all}, i, O_T)$
        if $I_{\text{sim}} < I_{\text{sim}}^{\min}$ then
            $i \leftarrow i + 1$
        else if $I_{\text{sim}}^{\min} \leq I_{\text{sim}} \leq I_{\text{sim}}^{\max}$ then
            $i \leftarrow i$
            break
        else
            $i \leftarrow i - 1$
        end if
        $iter \leftarrow iter + 1$
    end while
    return $i$
end function
```

α. These processes are detailed in Algorithm 1. Additionally, we utilize the Golden Section Search method to identify an optimal or sufficiently good value for α 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 ϕ (approximately 1.618) is the golden ratio, and a and b are the current search bounds for α. During each iteration, we compare $F(\alpha_1)$ and $F(\alpha_2)$, and adjust the search range accordingly:

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

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 8, while the outcomes of the user studies for both tasks are presented in Figs. 19 and 20. In total, we collected 570 votes from 95 participants across both studies. The specific responses for each question are detailed in Tables 8 and 9.

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 [69] 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.

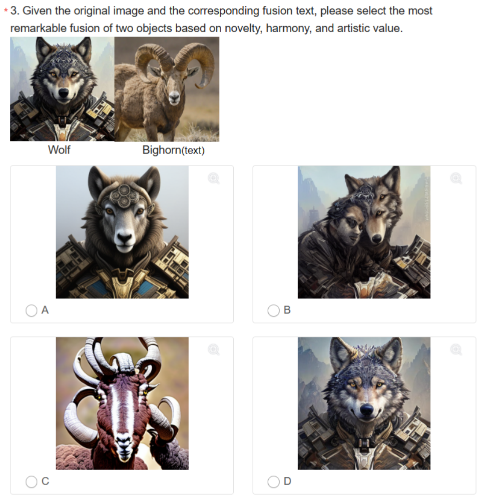

* 3. Given the original image and the corresponding fusion text, please select the most remarkable fusion of two objects based on novelty, harmony, and artistic value.

Figure 19: An example of a user study comparing various image-editing methods.

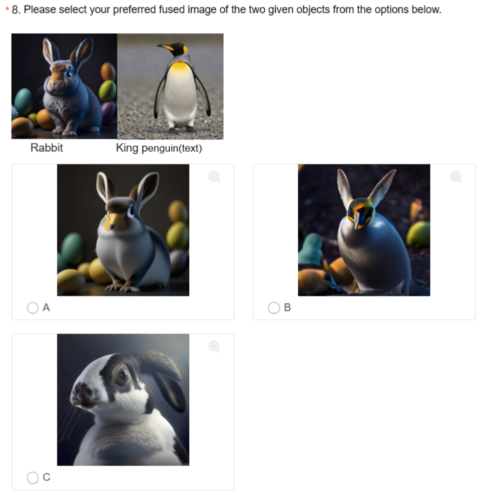

* 8. Please select your preferred fused image of the two given objects from the options below.

Figure 20: An example of a user study comparing various mixing methods.

Table 8: User study with image editing methods.

image-prompt \ options(Models)	A(Our ATIH)	B(MasaCtrl)	C(InstructPix2Pix)	D(InfEdit)
glass jar-salamander	77.89 %	1.05%	16.84%	4.21%
giraffe-bowling ball	89.74 %	2.11%	2.11%	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%

Table 9: User study with mixing methods.

options(Models) (prompt) image-prompt	A(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%

H More results.

In this section, we present additional results from our model. Fig. 21 showcases further generation results using our ATIH model. We experimented with four different images, each edited with four distinct text prompts. Fig. 22 provides further examples showcasing the effectiveness of our method in complex text-driven fusion tasks. Specifically, our approach excels in extreme cases by accurately extracting prominent features, such as color and basic object forms, from detailed textual descriptions. For instance, Fig. 22 shows a well-defined edge structure for the fawn image and the text 'Green triceratops with rough, scaly skin and massive frilled head.' Additionally, Fig. 23 illustrates our model's versatility with multiple prompts, emphasizing its capability for continuous editing.

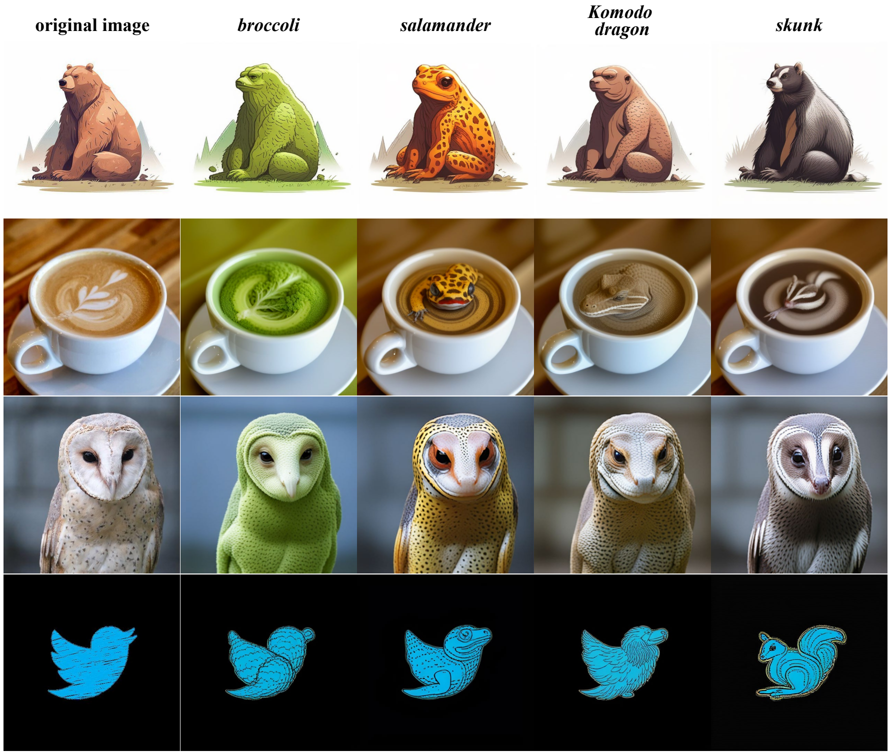


Figure 21: More visual Results.

I More Comparisons

In this section, we present additional results from our model and compare its performance against other methods.

In Fig. 24, we compare our results with those from the state-of-the-art T2I model DALL· E·3 assisted by Copilot. Our model shows superior performance when handling complex descriptive prompts for image editing. We observe that the competing model struggles to achieve results comparable to ours, particularly in maintaining the original structure and layout of images, despite adequate prompts.

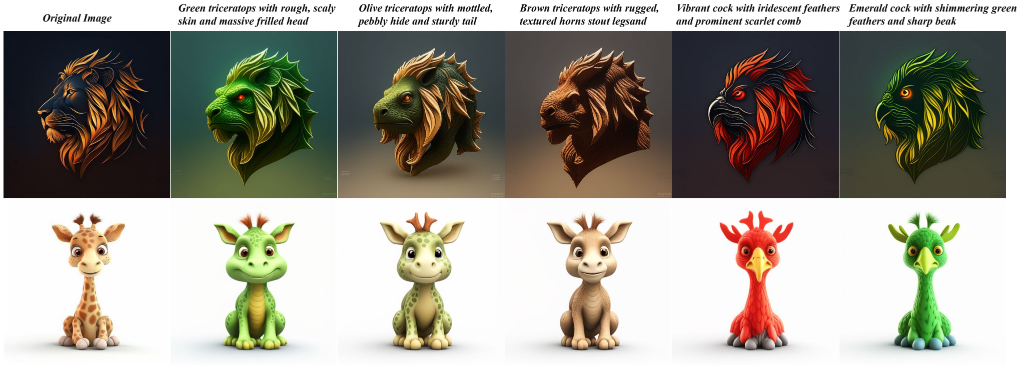

Original Image
Green triceratops with rough, scaly skin and massive frilled head
Olive triceratops with mottled, pebbly hide and sturdy tail
Brown triceratops with rugged, textured horns stout legsand
Vibrant cock with iridescent feathers and prominent scarlet comb
Emerald cock with shimmering green feathers and sharp beak

Figure 22: More visual results using complex prompt fusion.

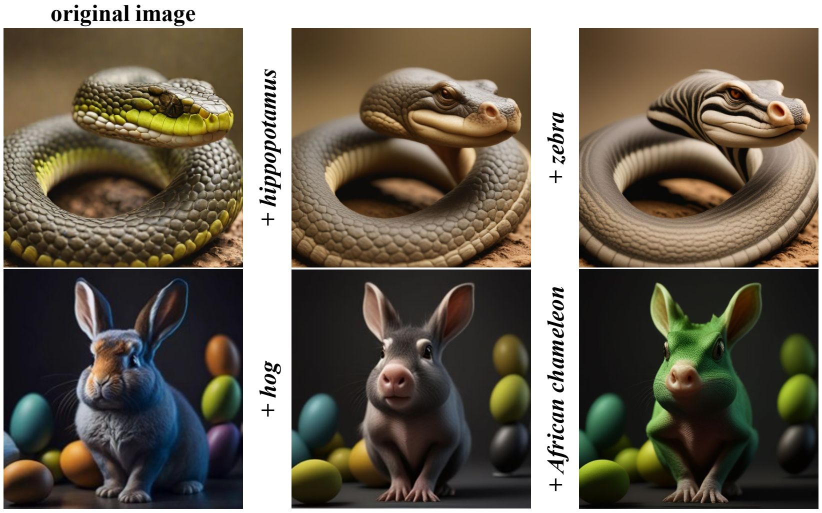
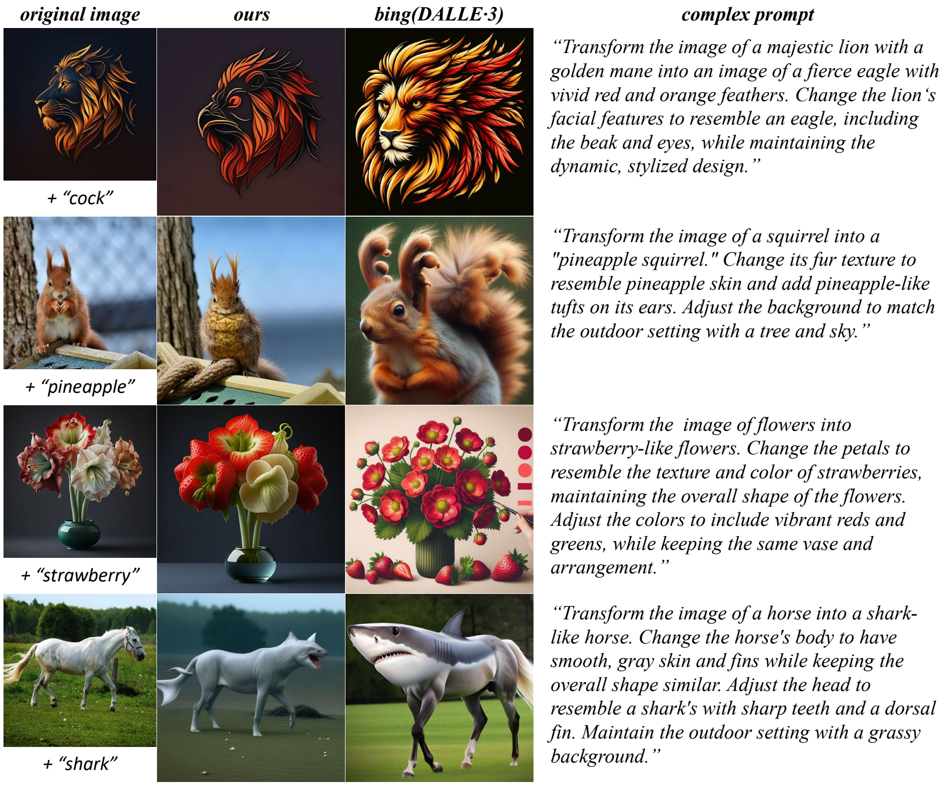


Figure 23: Fused results using three prompts.

In Figs. 25 and 26, we present additional comparison results with mixing methods. We observed that both MagicMix and ConceptLab tend to overly favor one category, as seen in examples like *Triceratops-Teddy Bear Toy* and *Anemone fish-Car*. Their generated images often lean more towards a single category.

Recently, subject-driven text-to-image generation focuses on creating highly customized images tailored to a target subject [18; 52; 9; 74]. These methods often address the task, such as multiple concept composition, style transfer and action editing [38; 8; 45]. In contrast, our approach aims to generate novel and surprising object images by combining object text with object images. Kosmos-G [45] utilize a single image input and a creative prompt to merge with specified text objects. The prompt is structured as "<i> creatively fuse with object text," guiding the synthesis to innovatively blend image and text elements. Our findings indicate that Kosmos-G can sometimes struggle to maintain a balanced integration of original image features and text-driven attributes. In Fig. 27, the images generated by Kosmos-G often exhibit a disparity in feature integration.



“Transform the image of a majestic lion with a golden mane into an image of a fierce eagle with vivid red and orange feathers. Change the lion‘s facial features to resemble an eagle, including the beak and eyes, while maintaining the dynamic, stylized design.”

“Transform the image of a squirrel into a "pineapple squirrel." Change its fur texture to resemble pineapple skin and add pineapple-like tufts on its ears. Adjust the background to match the outdoor setting with a tree and sky.”

“Transform the image of flowers into strawberry-like flowers. Change the petals to resemble the texture and color of strawberries, maintaining the overall shape of the flowers. Adjust the colors to include vibrant reds and greens, while keeping the same vase and arrangement.”

“Transform the image of a horse into a shark-like horse. Change the horse's body to have smooth, gray skin and fins while keeping the overall shape similar. Adjust the head to resemble a shark's with sharp teeth and a dorsal fin. Maintain the outdoor setting with a grassy background.”

Figure 24: Comparisons with complex prompt editing.

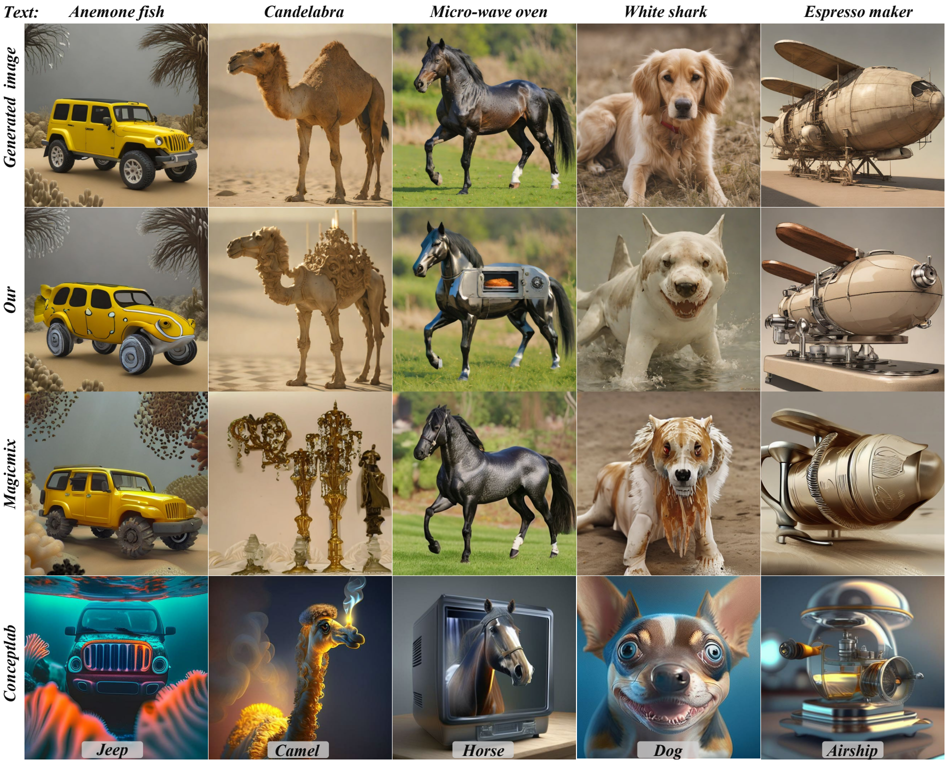


Figure 25: Comparison results of mixing methods using text-generated images.

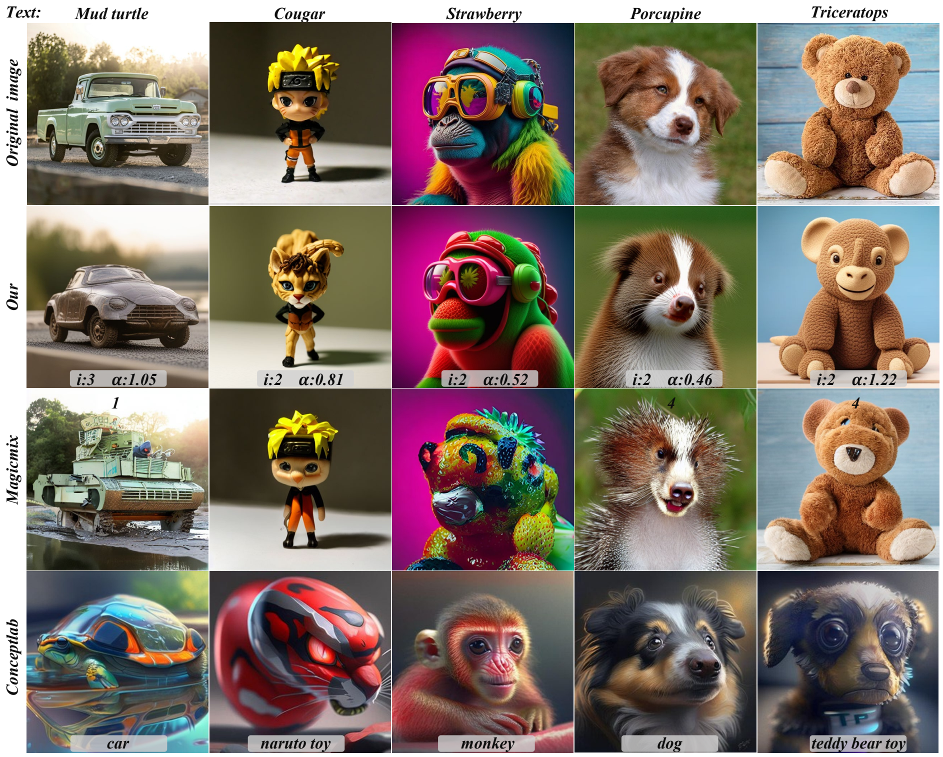


Figure 26: Further comparisons with mixing methods.

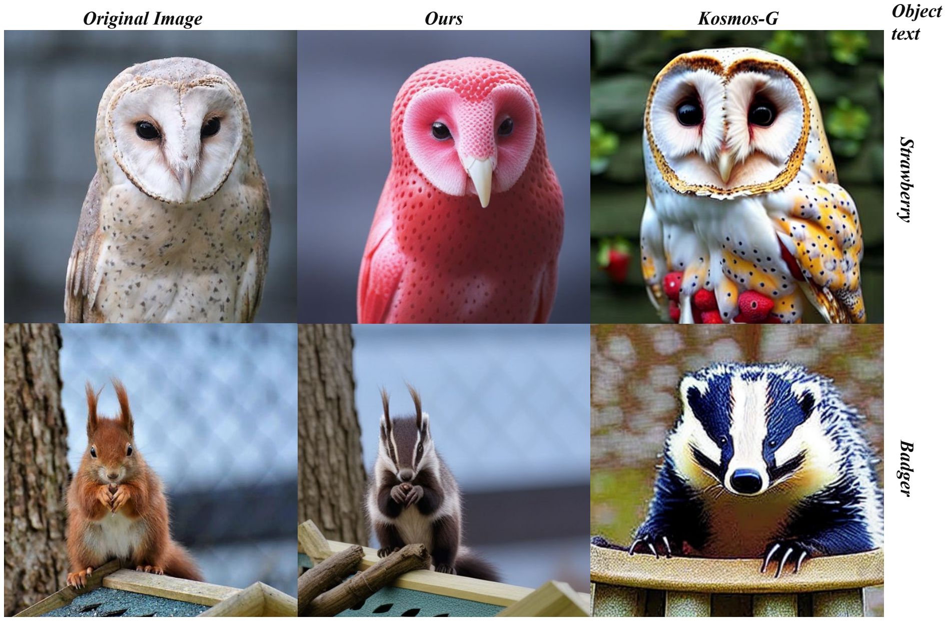


Figure 27: Comparisons with Subject-driven method.