VMetaphor: Enhancing Creativity through Automated Generation of Hybrid Visual Metaphors

Zixuan Wu
Wellesley College/MIT
aryawu@mit.edu

Meng Lu
Wellesley College/MIT
mlu108@mit.edu

Abstract

Visual metaphors are a powerful medium of expression in advertising, merging symbols and concepts to elicit rich emotional responses. Despite their effectiveness, machine learning (ML)-based generation of visual metaphors remains largely unexplored. This project bridges this gap by developing VMetaphor, a system that generates visual metaphors from user-provided object images and context. We curated a dataset of 300 images representing ideal hybrid metaphors and constructed VMetaphor with features including segmentation, mask selection, and an image-editing module based on a Stable Diffusion inpainting model fine-tuned on our dataset. Our ablation study and subsequent qualitative and quantitative evaluation confirm that the outputs from VMetaphor not only retain the creativity and semantic meanings of the original concepts but are also contextually grounded and visually appealing, showcasing its potential as a convenient and inspirational tool for designers.

1. Introduction

Metaphorical thinking is widely acknowledged as a crucial and potent driver of creativity [6, 8, 14]. In particular, visual metaphors are a form of communication that evokes emotions by combining symbols and concepts, which are used in mass media communications such as journalism and advertising [5, 7, 10, 11]. Despite their effectiveness, there has been limited research on generating visual metaphors using computer vision techniques.

Our project addresses this gap by exploring the generation of visual metaphors based on user-defined contexts. By inputting concepts such as "Seaside," "Supermarket," or "Health," "Education," the system aims to automatically create visual metaphors that combine the user's product object with an object from the given context. This project seeks to inspire new ways of perceiving the world through visual metaphors.

Figure 1. Six examples of visual metaphor from our dataset

2. Related Work

Currently, a few research investigated into the capability of SoTA generative models in generating visual metaphors and made efforts in improving the performances. [1] constructed benchmark of 5061 visual metaphor advertisement image with concept and relationship annotations, and evaluated state-of-the-art generation models, including Stable Diffusion and Imagen, identified room for improvement.

Focusing on the 'hybrid' metaphor type where the primary and secondary concepts are "visually conflated" [1], we explore the task of image collage editing tool to generate visual metaphor. [4] introduced a human-centered workflow for brainstorming possible visual blends by simple shape match between two concepts and users iteratively learn to make visual metaphors. [3] builds a photo from the sketch by searching for candidate images that match the provided text labels and performing synthesis. [9] proposed an approach to segment real life natural images into foregrounds and backgrounds, create a mask for the closest foreground object, and find a similar shape object based on local features similarity to paste onto the inpainted background.

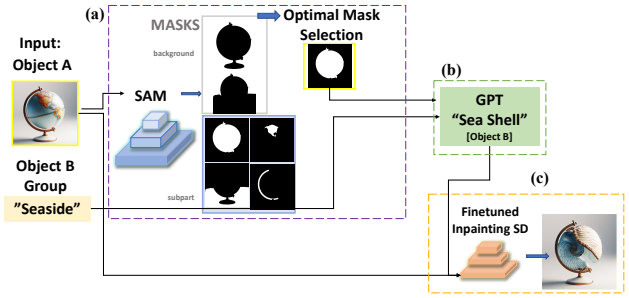

Figure 2. VMetaphor Diagram

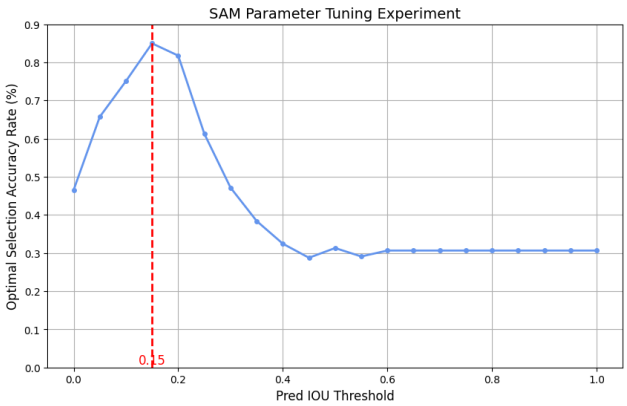

Figure 3. Example of SAM Parameter Tuning Experiment

We evaluate the generated visual metaphors by measuring their interpretation using natural language. There are several computational models [13] [12] regarding conceptual Metaphor generation based on context mapping. Leveraging them in visual context, [2] introduced the task of generating visual metaphors from linguistic metaphors, leveraging DALL_E to visualize the textual interpretation formed by GPT-3 from linguistic metaphors.

3. Method

Our Visual Metaphor Generation Pipeline creates metaphorical advertisement images tailored to user-defined concepts. It takes an input image of a product and a textual concept, and produces an output image that visually represents the metaphor. The pipeline generates masks for object parts in the input image, selects the best mask, prompts GPT-4 for an object based on the mask and concept, and uses a fine-tuned Stable Diffusion model for inpainting. The final output is an advertisement image that uses a visual metaphor to highlight a feature of the product relevant to the given context.

3.1. Segment Object By Parts

We use the Segment Anything Model (SAM) to identify areas on an object where another object can be integrated. SAM segments input images into distinct parts based on prompts. By sampling single-point input prompts in a grid across the image, SAM generates masks for the entire image. These masks can be filtered for quality and duplicates can be removed using non-maximal suppression. For a single ambiguous point prompt, SAM generates three valid masks: on the object level, on the part level, and on the subpart level. Since we are working with input images containing only one object on a white background, we exclude the mask generated on the object level and save the masks on the part level.

To optimize SAM's parameters for our specific problem of merging two objects, we aimed to provide the model with more flexibility in reconstructing missing regions. We achieved this by reducing the predicted IOU (Intersection over Union) threshold to 0.15 and the stability score threshold to 0.78. This change allows the model to include segmentation masks that it might consider low quality due to an imperfect fit but with more relaxed shapes. Additionally, we experimented with masks of varying sizes and selected the optimal masks by sorting all generated masks at the part level by size and choosing the largest masks as subsequent inpainting inputs. This decision was based on our observation that when masks were too small relative to the entire object in the image, the object's information dominated the model's output. As an example, in Fig. 3, we demonstrate our experiment for determining the predicted IOU threshold. We calculated the accuracy rate of the largest mask at the part level as the optimal mask identified in online visual metaphor based on different IOU values.

3.2. Stable Diffusion Inpainting

We leveraged Stable Diffusion Inpainting Model to generate the visual metaphor guided by masks. Stable diffusion model was picked to be used and evaluated in prior works [2] [1]. It offer a balance between training stability and image quality, making them well-suited for generating visually meaningful metaphoric images.

3.2.1 Dataset

To finetuned the stable diffusion model, we collected a dataset comprising 300 advertisement images sourced from the Internet, specifically selected for their use of "visual metaphors." We focused our selection on images categorized as "hybrid" metaphor types [1], the fusion of two distinct objects into a single composite entity. Our objective is for the model to discern and learn the underlying strategies of merging and the compositional patterns.

3.2.2 Inpainting Generation

The first step of inpainting is to identify the text input of the inpainting object that can potentially compose a visual metaphor within the user-defined field of concept. We provide GPT4 with the original object, the image of the selected mask and input the following prompt: "Here is a mask. Give me a object that (1) has very similar shape as the mask AND (2) belong to the group: {group}. Respond in this format: 'OBJECT: [FILL IN]"

With the mask-object pair as input, we run our finetuned model to generate the visual metaphor image output.

4. Results

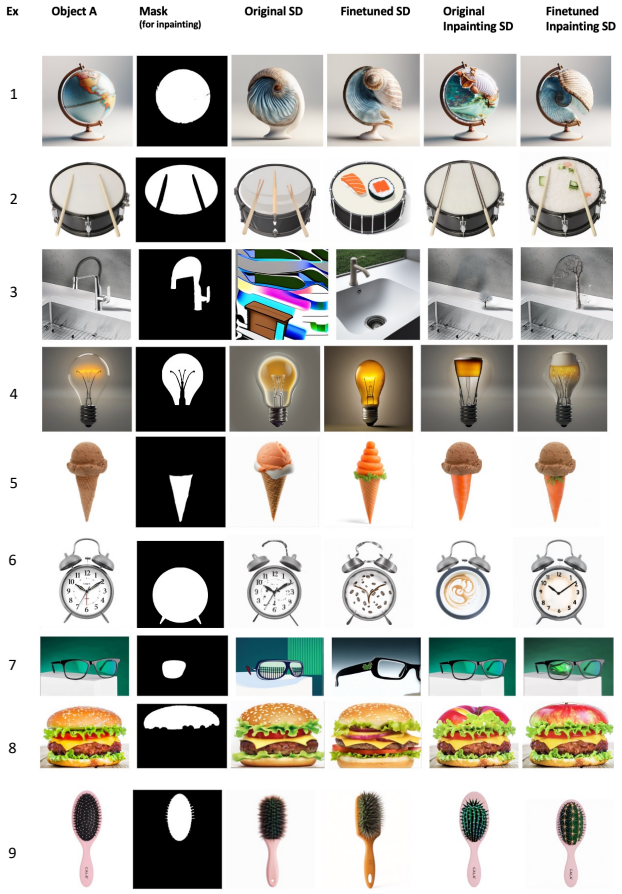

Figure 4. For condition 1 and condition 2, the model is given input of Object A and a text input "Create a visual metaphor of this image [Object A] and [Object B]"; For condition 3 and 4, the model is given input of Object A, the mask, and the text input "[object B]"

In order to understand the effect of finetuning and inpainting module of VMetaphor, we conduct ablation study.

We define two general of metrics to evaluate the generation outputs: (A) False Generation, which includes situations where at least of the objects is forgetten or other irrelevant objects are generated in the output (B) Nonoptimal Arrangement, which includes the situation where both objects are successfully generated but they have bad arrangement or placement such as direct side-by-side placement (Figure 6.1) or direct top-down placement (Figure 4.2), (C) No Error, which includes the generations when neither False Generation or Nonoptimal Arrangement occurs in the generation output. In particular, the finetuned inpainting SD achieves the highest No Error Rate, indicating that the determining effect of finetuning and inpainting mode's impact.

In the baseline model without finetuning or inpainting, the No Error rate was 4%, implying that most outputs included either FG or NA errors. The FG rate was notably

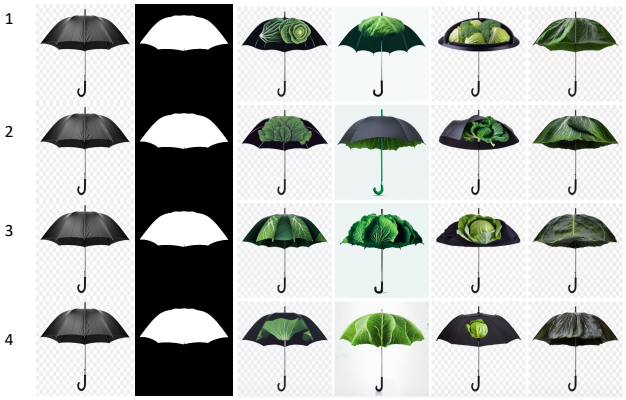

Figure 5. Umbrella-Cabbage. We generate more images for the same input pairs to ensure the analysis is not based on one random seed

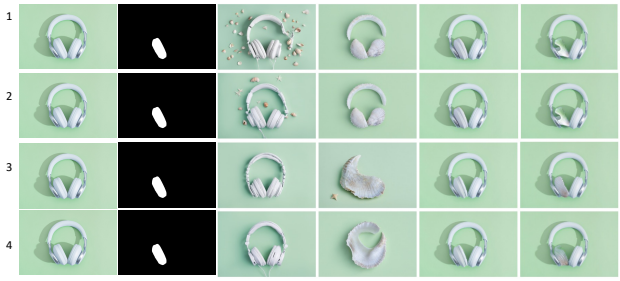

Figure 6. Headphone-Seashell.

Conditions	NAR	FGR	NER
1 Original img-to-img SD	0.313	0.647	0.04
2 Finetuned img-to-img SD	0.294	0.412	0.30
3 Original Inpainting SD	0.176	0.706	0.18
4 Finetuned Inpainting SD	0.005	0.109	0.885

Table 1. NAR, FGR and NER stands for False Generation Rate, Nonoptimal Arrangement Rate, No Error Rate correspondingly

high at 64.7%, with examples like Fig 2, 7, 4, 5, 6, and 8 showing no presence of object B and merely replicating object A. Figures 3, 1, and 9 illustrate failures such as random generation and unrecognizable renderings of object A. Additionally, the NA rate stood at 4%, with Fig 6.1, 6.2 and Fig5.1,5.2,5.3,5.4 exemplifying nonoptimal placement strategies.

Finetuning based on our dataset resulted in a 26% improvement in the No Error rate for Condition 2. It largely resolved issues related to side-by-side placement and the repetition of only one object, indicating that the model was beginning to understand that the goal of a "visual metaphor" equates composing these two abstractly related objects together. However, new compositional challenges emerged, such as in Fig 6.1, 6.2, 6.3, 6.4, Fig 4.1, where the model inappropriately utilized the entire space of object A for object B, compromising its recognizability.

The unfinetuned inpainting model demonstrated the ability to reduce the Nonoptimal Arrangement rate by 13%, as more successful integrations were noted in outputs like Fig 4.5, 4.8, 4.9, 4.1, Fig 5.1, 5.2, which strictly positioned object B within specified subparts. However, this condition also exhibited the highest False Generation rate. Notably, the strict enforcement of placing object B, as seen in Fig 5.1, 5.2, led to the introduction of unrelated objects like pots and containers, detracting from the intended visual metaphor. Large mask areas prompted the model to fill the entirety with object B, often resulting in a top-down arrangement rather than a coherent composition and merging. These two significant errors seem to indicate that the model attempts to make the output reasonable, thus always trying to place the whole object B into the mask if possible without adopting strategies of merging it such as compromising the the shape of Object B to fit the mask's edge or filling the mask with the texture.

Condition 4, integrating both finetuning and inpainting, showcased a significant elevation in the No Error rate to 89%. This indicates that VMetaphor effectively addressed the issues identified in earlier conditions, demonstrating a robust understanding and implementation of the visual metaphor concept.

In summary, the ablation study highlights the critical roles of finetuning and inpainting in enhancing the model's ability to generate and arrange visual metaphors correctly. Each adjustment contributed distinctively, with the combined approach proving most effective in achieving the desired outcomes.

5. Evaluation

In this section, we outline the methods we used to evaluate the quality and effectiveness of the generated images.

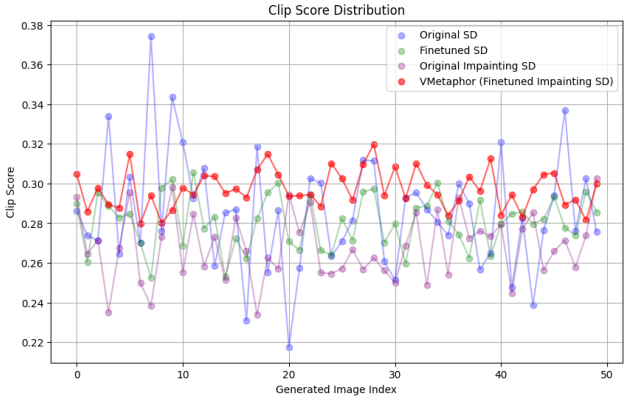

Figure 7. Clip Distribution for Four Conditions: Original SD, Finetuned SD, Original Impainting SD, Finetuned Impainting SD(VMetaphor)

Model	Mean	Median	std
0 Groundtruth	0.30	0.30	0.01
1 Original SD	0.29	0.28	0.03
2 Finetuned SD	0.28	0.28	0.01
3 Original Inpainting SD	0.27	0.27	0.02
4 Finetuned Inpainting SD*	0.30	0.29	0.01

Table 2. Distribution of Model Scores

Models	meandiff	p-adj	lower	upper	reject
1&2	-0.0046	0.619	-0.0144	0.0052	False
1&3	-0.0171	0.0001	-0.0269	-0.0074	True
1&4	0.0119	0.0102	0.0021	0.0217	True
2&3	-0.0126	0.0057	-0.0223	-0.0028	True
2&4	0.0165	0.0001	0.0067	0.0262	True
3&4	0.029	0.0	0.0193	0.0388	True

Table 3. Results of Multiple Comparison of Means - Tukey HSD

5.1. Quantitative Analysis

To quantitatively assess the effectiveness of the generated results from our system and validate the findings of our ablation study, Table 1, we employed the CLIP score metric. CLIP score measures the degree of integration between two objects by evaluating the similarity between the visual metaphor image and the text "object A and object B." A higher CLIP score indicates that objects A and B are more detectable by the CLIP model, while a lower score suggests lesser detectability. For each visual metaphor in our dataset, we calculated the CLIP score and considered these as ground truth. Our goal is to ensure that the distribution of scores from our system's generated images corroborates with the result in Table 1 statistically.

We calculated CLIP scores for all generated images (50 images) across the four model structures involved in the ablation study, as illustrated in Figure 7. By analyzing the mean, median, and standard deviation of these scores and using the Tukey HSD test for further validation, we compared them with the ground truth distribution. This comparison aimed to verify whether our experimental results align with our initial hypotheses and confirm the effectiveness of the system's components.

The statistical analysis presented in Tables 2 and 3 supports our conclusions from the ablation study. Notably, Model 4 (VMetaphor: finetuned + inpainting) shows the closest mean CLIP score to the ground truth, indicating a high level of object fusion similar to that observed in the finetuned dataset. The results from the Tukey HSD test, which rejects the null hypothesis of no significant differences between Model 4 and all other models, further demonstrate that both finetuning and inpainting contribute to the model's ability to generate a hybrid composition that are significantly different than the generation with ablated components, and also to the model's ability to generate a hybrid composition closely resembling the ground truth.

Moreover, Model 3, with the lowest mean CLIP score, is distinguishable from all other models according to the HSD results. This finding aligns with our analysis, as Model 3 exhibited the highest False Generation rate (70.6% in Table 1), theoretically justifying the lowest CLIP scores. This is because a high False Generation rate typically involves missing or incorrectly added objects, thus reducing the accurate representation of the two intended objects.

Models 1 (Original SD) and 2 (Finetuned SD) do not show significant differences in their distributions as per the HSD test results. Given their similar proportions of False Generation Rate and Nonoptimal Arrangement Rate compared to other models, their CLIP score distributions are more challenging to differentiate. However, it is important to note that Model 1 displays a higher standard deviation than Model 2, likely due to its higher False Generation Rate, which includes instances of random generation where both objects are omitted, leading to significantly lower CLIP score outliers.

5.2. Qualitative Evaluation

Given that CLIP score is not tailored towards metaphoric compositional images, we conduct two rounds of human evaluation to measure the 'hybrid' composition quality of the generated images, and the preservation of semantic meaning of visual metaphor after the composition.

To analyze the quality of hybrid composition, we perform human studies comparing two different models at a time. Specifically, human participants are given the definition of hybrid composition as the primary and secondary concepts visually conflated. [1] Human participants are then presented with four examples generated by two our finetuned models and two baseline stable diffusion models. Participants are asked to judge, for each image, if hybrid composition are achieved.

Model	**1**	**2**	**3**	**4**
Composition	9.09%	36.36%	54.55%	100%
Metaphor	9.09%	27.27%	54.55%	90.91%

Table 4. Qualitative Evaluation

To evaluate the quality of visual metaphors, we adopt the annotation definition from the paper MetaClue [1] as the linguistic interpretation of the visual metaphor. The human participants are given the generated image, and are asked if they are inspired and are able to come up with a linguistic metaphor in the form of primary concept's relation with secondary concept.

Table 4 illustrates that Model 4, VMetaphor, significantly outperforms all ablated models, with 91% improvement from original img-to-img SD and 45.5% improvement from the unfintuned inpainting SD, highlighting VMetaphor's enhanced capability in hybrid composition and visual metaphor expression, as it allows the generation of understandable and meaningful visual metaphors.

6. Conclusion

The ablation study with the analysis on Nonoptimal Arrangement Rate, False Generation Rate and No Error Rate demonstrating VMetaphor's superior performance over ablated models. CLIP score evaluation corroborates with the analysis, and human evaluation demonstrates that VMetaphor excels, with significant improvements for compositition strategies and coherent metaphor generation with semantic meaning. By enabling intuitive and effective visual metaphor generation, VMetaphor can assist with advertising strategies, enhance educational materials, and empower artists to explore new realms of digital expression.

7. Contribution

As two Wellesley students, we mostly worked together in a pair-programming setting. We both contributed equally to the construction of the dataset and the finetuning of the stable diffusion model using this dataset. Zixuan worked on writing the code pipeline for SAM mask generation and inpainting using stable diffusion. Meng worked on the experiment of mask selection for inpainting and the code for object selection using GPT. The running of experiments was done individually to generate a large sample of results in a short amount of time. With the generated results, we worked together on the ablation studies, statistical evaluation, and explaination together.

References

[1] Arjun R. Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T. Freeman, Yuanzhen Li, and Varun Jampani. Metaclue: Towards comprehensive visual metaphors research, 2023. 1, 2, 5

[2] Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. I spy a metaphor: Large language models and diffusion models co-create visual metaphors, 2023. 2

[3] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. Sketch2photo: Internet image montage. *ACM Transactions on Graphics (TOG)*, 28(5):124, 2009. 1

[4] Lydia B Chilton, Savvas Petridis, and Maneesh Agrawala. Visiblends: A flexible workflow for visual blends. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019. 1

[5] Charles Forceville. *Pictorial metaphor in advertising*. Routledge, 1996. 1

[6] George Lakoff. *The Contemporary Theory of Metaphor*, chapter The Contemporary Theory of Metaphor. Publisher Name, 1993. 1

[7] George Lakoff and Mark Johnson. *Metaphors We Live By*. University of Chicago Press, 2008. 1

[8] Todd I. Lubart and Isaac Getz. Emotion, metaphor, and the creative process. *Creativity Research Journal*, 10(4):285–301, 1997. 1

[9] Othman Sbai, Camille Couprie, and Mathieu Aubry. Surprising image compositions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3921–3925, 2021. 1

[10] Linda M. Scott. Images in advertising: The need for a theory of visual rhetoric. *Journal of Consumer Research*, 21(2):252–273, 1994. 1

[11] Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. Metaphor generation with conceptual mappings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6724–6736, 2021. 1

[12] Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. Metaphor generation with conceptual mappings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6724–6736, 2021. 2

[13] Asuka Terai and Masanori Nakagawa. A computational system of metaphor generation with evaluation mechanism. In *International Conference on Artificial Neural Networks*, pages 142–147. Springer, 2010. 2

[14] Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. *Metaphor: A Computational Perspective*, volume 9 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers, 2016. 1