Analyzing Multimodal Objectives Through the Lens of Generative Diffusion Guidance

Chaerin Kong, Nojun Kwak
Seoul National University
{veztylord,nojunk}@snu.ac.kr

ABSTRACT

Recent years have witnessed astonishing advances in the field of multimodal representation learning, with contrastive learning being the cornerstone for major breakthroughs. Latest works delivered further improvements by incorporating different objectives such as masked modeling and captioning into the frameworks, but our understanding on how these objectives facilitate learning remains vastly incomplete. In this paper, we leverage the fact that classifier-guided diffusion models generate images that reflect the semantic signals provided by the classifier to study the characteristics of multimodal learning objectives. Specifically, we compare contrastive, matching and captioning loss in terms of their semantic signals, and introduce a simple baseline that not only supports our analyses but also improves the quality of generative guidance in a straightforward manner.

1 INTRODUCTION

Vision-Language Pretraining (VLP) has attracted great attention from the community for its wide and robust applications in different downstream tasks. The seminal work of CLIP (Radford et al., 2021) employs the image-text contrastive objective to successfully embed images and text descriptions in a common feature space, inspiring numerous subsequent works that explore different objectives (Li et al., 2022a; Yu et al., 2022; Yang et al., 2022; Jang et al., 2023) and architectures (Li et al., 2021; Jang et al., 2022; Wang et al., 2022). Recently, cross-modal generative models (Ramesh et al., 2021; Saharia et al., 2022; Rombach et al., 2022; Kong et al., 2022) are also gaining wide popularity thanks to the powerful capacity of diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020) and readily available guidance of vision-language foundation models. These models aim to synthesize or edit images so that the outputs are both realistic and condition-aligned.

Conditional diffusion models embody the conditioning information in two ways: classifier guidance (Dhariwal & Nichol, 2021) and classifier-free guidance (Ho & Salimans, 2022). While classifier-guided models (Nichol et al., 2021; Dhariwal & Nichol, 2021; Kong et al., 2023) typically leverage multimodal embeddings of a joint embedding network (e.g., CLIP) using cosine similarities, it is unclear whether this approach is optimal. For example, recent works (Zhong et al., 2022; Li et al., 2022b) have pointed out that global representations of CLIP learned by the contrastive objective are not suitable for handling fine-grained correspondences between image and text as they represent the image and the text as a whole. This indicates that features learned from sequence level contrastive learning may have certain blind spots that can be made visually apparent when directly applied to the diffusion process as the guidance signal.

In this paper, we aim to obtain a better understanding of different multimodal learning objectives (e.g., image-text contrastive, image-text matching, image captioning) by analyzing their semantic signals as generative guidance. That is, we utilize the classifier-guided diffusion process to study the characteristics of different objectives by carefully inspecting the samples they produce, and further present a straightforward modification to the previous method that both supports our findings and improves the generation quality. We note that the aim of this paper is not to present a high-performing generative model. Rather, we are simply leveraging the diffusion process to visually analyze different objectives and hypothesize about the properties of the accordingly learned representations.
2 Multimodal Objective as Generative Guidance

2.1 Classifier-guided Diffusion

Dhariwal & Nichol (2021) have introduced classifier guidance as a means to steer the generative diffusion process towards the conditioning class. This requires a noise-aware classifier whose gradient can be used to guide the diffusion process. Formally, denoting the predicted parameters of timestep $t$ as $\mu_\theta(x_t), \Sigma_\theta(x_t)$, the next step diffusion sampling becomes

$$x_{t-1} \sim \mathcal{N}(\mu_\theta(x_t) + s\Sigma_\theta(x_t)\nabla_x, \log p_\phi(y|x_t), \Sigma_\theta(x_t)), \quad (1)$$

where $s$ is the step size, $y$ indicates the class label and $p_\phi$ refers to the classifier. This formulation was altered by Nichol et al. (2021) to suit text-to-image generation as follows:

$$x_{t-1} \sim \mathcal{N}(\mu_\theta(x_t) + s\Sigma_\theta(x_t)\nabla_x (f(x_t), g(c)), \Sigma_\theta(x_t)), \quad (2)$$

where $\langle \cdot, \cdot \rangle$ indicates the inner product and $f, g, c$ are the image encoder, text encoder, and the text condition, respectively.

Overall, the classifier guidance encourages the model to generate samples that are well-aligned with the condition according to a predefined metric. Obviously, the choice of this metric affects the final output sample, revealing how each metric (objective) connects the two modalities (image and text, in our case) in multiple levels.

2.2 Pretrained Models

As our goal is to study the semantic signals encoded in different objectives, we employ a pretrained diffusion backbone and a pretrained vision-language guidance model for our analysis.

**Generative Model**

For the image generator, we use a 256×256 unconditional diffusion model pretrained on Imagenet (Russakovsky et al., 2015). We note that this is not a state-of-the-art text-to-image diffusion model. The unconditional nature of this model renders it well-suited for our purpose, as it solely relies on the classifier signal for condition-aware synthesis, making it possible to analyze the encoded semantic information in a disentangled manner. Employing an excessively powerful generator can similarly obfuscate our analysis, as its generative capacity can compensate for the weakness in the guidance signal and mask its blind spots.

**Guidance Model**

Among many available candidates, we choose BLIP (Li et al., 2022a) as our main guidance mode$^1$. This model is trained on 129M image-text pairs simultaneously optimizing for three objectives: image-text contrastive (ITC), image-text matching (ITM) and image captioning (CAP). The fact that a single model can evaluate these three scores makes it an excellent guidance model for our analysis, as we can safely minimize the compounding effects coming from using different models trained with different dataset, architecture and optimization scheme. We adopt the idea from Avrahami et al. (2022) to first predict the denoised version for guidance signal computation.

2.3 Objectives and Benchmarks

We analyze three commonly used multimodal objectives: ITC, ITM and CAP. We simply replace the classifier logit in eq. (1) with the corresponding loss terms of BLIP (with the sign reversed). For details about the loss computation, please refer to Li et al. (2022a). For systematic evaluation, we mainly follow the benchmark of Saharia et al. (2022), namely DrawBench. We further add prompts from COCO (Lin et al., 2014) captions. Qualitative evaluations as well as quantitative comparisons based on user study are performed to draw insights.

3 Analyses

In this section, we first present our empirical findings and propose a straightforward modification called SHIFT that reflects the insights by employing the coarse-to-fine guidance of CAP and ITC. We acknowledge that using a more powerful model would relieve some of the issues demonstrated in this section as the data scale can compensate for the blind spots, but we believe the findings we present here nevertheless hold true and will come into play in an increasingly challenging problem setting.

---

$^1$https://github.com/openai/guided-diffusion

$^2$https://github.com/salesforce/BLIP
3.1 Findings

Figure 1: Text-to-image samples using each objective as the classifier-guidance. We only present preliminary examples for illustrative purpose. Please refer to the Appendix for more cases.

1. While IT wikipedia focus on the fine details of the salient object, CAP tends to reason about the global scene composition.

Looking at Figure 7 row 1, we can clearly see that contrastive loss is more effective in forming fine details of the main object while it often leaves out less important objects or attributes in the given prompt. We hypothesize the former is due to the core dynamic of contrastive learning that aims to learn relative distances by distinguishing objects. At the same time, as the objective only compares fully abstracted representations (i.e., [cls] tokens) that dominate each entity as a whole, it fails to densely parse the scene. On the other hand, captioning objective forces to understand the scene structure in a deeper level, making captioning-guided samples more faithful to complex texts.

2. IT wikipedia commonly fuses visual semantics together to forcefully form a global semantic.

2 in Figure 7 illustrates a more extreme case where IT wikipedia not only omits semantic components but arbitrarily mixes them. From these examples, we can diagnose that the contrastive objective does reflect semantic attributes (e.g., blue) but fails to relate them to the correct object (e.g., dog). This can be another side effect of simplified distance learning. Captioning, in contrast, requires the model to reason about both objects and their relations, deepening the scene-level understanding.

3. Patch-token cross-attention plays a key role in fine-grained visual understanding.

We now widen the scope of our analysis by further looking at ITM. In contrary to IT wikipedia that simply compares two [cls] tokens, ITM involves lower-level cross-attention between image patch tokens and text tokens (as in CAP) to output a matching score between 0 and 1. We discovered that this operation plays an important role in fine-grained visual understanding and representation robustness. To our surprise, ITM, more or less an auxiliary loss to polish multimodal representations, encodes strong semantic signals that involves dense scene understanding. Looking at 1, 2, 3, cross-attention-based objectives (ITM, CAP and SHIFT) demonstrate capacity for fine-grained visual reasoning, and ITM signal successfully materializes objects with corresponding attributes and relations, though at a lower visual quality.

4. Dense supervision makes the representations more robust to noise perturbations.

Last two rows of Figure 7 show the impact of noise in the text prompt. 4 depicts that as opposed to IT wikipedia that generates random objects under mild typo, other losses render relatively consistent outputs.
delivers a similar insight, where the phrase ‘camel in the dessert’ not ‘desert’ is likely mistaken by the text provider. These cases are very probable in the typical setting where massive noisy image-text data are crawled from the web, and we observe that dense supervision that involves low-level patch-token cross-attention shows better robustness against textual perturbations, as perturbed text inputs attend to not only themselves but also visual information to form a more robust representation.

5. **CAP** is a more indirect if not challenging form of supervision than **ITC** or **ITM**.

![Generated samples for each objective and the number of diffusion sampling steps.](image)

Lastly, we inspect the optimization complexity of each objective by differing diffusion sampling steps. As each diffusion step corresponds to an update using the loss gradient, we regard an objective that generates reasonable sample with fewer steps to have lower optimization complexity. Referring to Figure 2, we see that **ITC** and **ITM** clearly take less steps to output realistic samples compared to caption-based losses. This observation coincides with [Radford et al., 2021] and [Yu et al., 2022], where the former explicitly chose contrastive loss for training efficiency and the latter has been reported to take much more resources to converge due to captioning. We conclude that as captioning demands a more semantic visual understanding, learning becomes trickier compared to the simple distance learning.

### 3.2 New Baseline: Guidance Shift

Based on the above findings, we propose a simple yet effective baseline that takes advantage from both ends, *i.e.*, contrastive learning and captioning. To leverage the strengths from both, we introduce guidance shift, where we start with captioning loss and gradually shift to contrastive loss for the generative guidance. Formally, our **SHIFT** loss can be written as:

$$L_{SHIFT} = tL_{ITC} + (1 - t)L_{CAP},$$

where $t$ is the normalized time step, progressing from 0 to 1. The idea is to first outline the overall structure with **CAP** and then refine the details with **ITC**. To study its effectiveness, we conduct quantitative user study as well as qualitative evaluations presented in the Appendix.

![Human evaluation for photo-realism and condition-alignment.](image)

Fig 3 delivers the result. Compared to simple **ITC** baseline, **SHIFT** outperforms in both fidelity and alignment. Although **CAP**-only shows better condition alignment, **SHIFT** clearly outputs better quality samples, which is apparent from qualitative results as well. **BLEND**, a naive baseline that simply mixes **CAP** and **ITC** without gradual transition, performs significantly worse as these two signals can often be conflicting and difficult to optimize simultaneously.

Fig 4 further visually verifies our claim that shifting from **CAP** to **ITC** maximizes the positive synergy between the two dissimilar optimization goals to strike the best balance between image quality and condition alignment. Although **ITM** involves lower-level cross-attention, neither combining it with **CAP** nor **ITC** achieves the desired effect.
Figure 4: Qualitative comparison between baselines that combine multiple objectives. We observe that gradually shifting from CAP to ITC enjoys advantages from both sides, i.e., faithful scene composition and realistic details.

4 CONCLUSION

In this paper, we have studied the semantic information encoded in different multimodal objectives by visually analyzing their properties as generative diffusion guidance. We hope it provides useful insights for ensuing works and sparks further advances in the field.

Acknowledgements This work was supported by NRF grants (2021R1A2C3006659, 2022R1A5A7026673) and IITP grants (2021-0-01343, 2022-0-00320) funded by Korean Government. The authors would also like to thank Jiho Jang and Seungwoo Lee for their constructive feedback and discussions.

REFERENCES


## Appendix

| Prompt: “A picture of birds flying over a lake at sunset.” |
|---|---|---|---|
| ![ITC](image1) | ![ITM](image2) | ![CAP](image3) | ![SHIFT](image4) |

| Prompt: “A picture of a gray rabbit eating carrot.” |
|---|---|---|---|
| ![ITC](image5) | ![ITM](image6) | ![CAP](image7) | ![SHIFT](image8) |

| Prompt: “A large white cat is sitting on a table in front of a TV.” |
|---|---|---|---|
| ![ITC](image9) | ![ITM](image10) | ![CAP](image11) | ![SHIFT](image12) |

| Prompt: “A woman and her dog at the beach.” |
|---|---|---|---|
| ![ITC](image13) | ![ITM](image14) | ![CAP](image15) | ![SHIFT](image16) |

| Prompt: “A picture of a sweet shop having a colorful front entrance.” |
|---|---|---|---|
| ![ITC](image17) | ![ITM](image18) | ![CAP](image19) | ![SHIFT](image20) |

| Prompt: “A picture of a colorful train painted in blue, yellow and orange moving on tracks.” |
|---|---|---|---|
| ![ITC](image21) | ![ITM](image22) | ![CAP](image23) | ![SHIFT](image24) |

| Prompt: “A cube made of brick.” |
|---|---|---|---|
| ![ITC](image25) | ![ITM](image26) | ![CAP](image27) | ![SHIFT](image28) |

| Prompt: “A couple of lawn chairs sitting next to the ocean.” |
|---|---|---|---|
| ![ITC](image29) | ![ITM](image30) | ![CAP](image31) | ![SHIFT](image32) |

| Prompt: “A picture of Mississippi paddle boats making their way on the river.” |
|---|---|---|---|
| ![ITC](image33) | ![ITM](image34) | ![CAP](image35) | ![SHIFT](image36) |

Figure 5: Additional text-to-image generation examples. We can consistently observe that while **ITC** focuses on detailed formulation of the salient object, **CAP** and its variant **SHIFT** understand the prompt in a finer level and output more faithful visualizations. It is also apparent that **ITC** alone often leaves out certain objects or mixes different visual semantics (e.g., *colorful entrance*).
Figure 6: Additional experiments for noise robustness. Although ITC produces realistic images with clean prompts, minor typos can completely ruin their semantic signals. In contrast, losses that provide denser supervisions generally output consistent results despite textual noise, showing better robustness.

Figure 7: Additional results for optimization complexity. Captioning-based losses require more diffusion steps to generate realistic images, while ITC and ITM quickly forms reasonable shapes and appearances.