ANALYZING MULTIMODAL OBJECTIVES THROUGH THE LENS OF GENERATIVE DIFFUSION GUIDANCE

Anonymous authors
Paper under double-blind review

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) 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) 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) has 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

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

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

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

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[1] (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 model[2]. 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

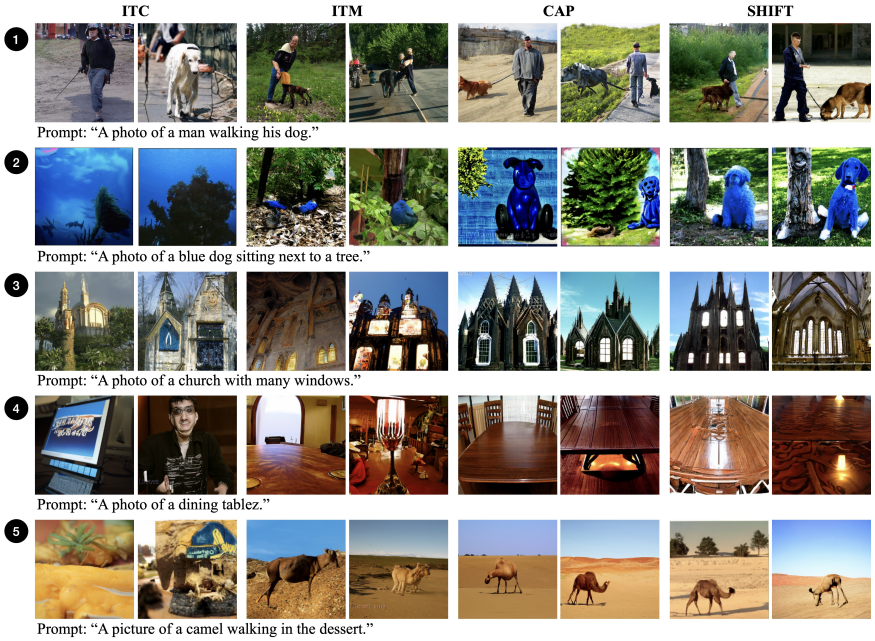

Figure 1: Text-to-image samples using each objective as the classifier-guidance. We only present preliminary examples for illustrative purpose. Please refer to the Appendix for more cases.

*1. While **ITC** focuses on the fine details of the salient object, **CAP** tends to reason about the global scene composition.*

Looking at Figure 7 row ❶, 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. **ITC** commonly fuses visual semantics together to forcefully form a global semantic.*

❷ in Figure 7 illustrates a more extreme case where **ITC** 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 **ITC** 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 ❶, ❷, ❸, 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. ❹ depicts that as opposed to **ITC** that generates random objects under mild typo, other losses render relatively consistent outputs.

❺ 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**.*

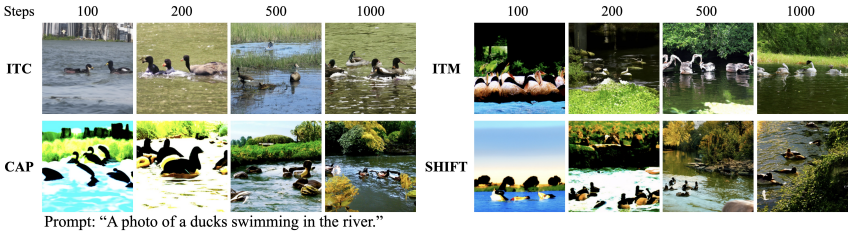

Figure 2: Generated samples for each objective and the number of diffusion sampling steps.

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:

$$\mathcal{L}_{SHIFT} = t\mathcal{L}_{ITC} + (1-t)\mathcal{L}_{CAP}, \tag{3}$$

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.

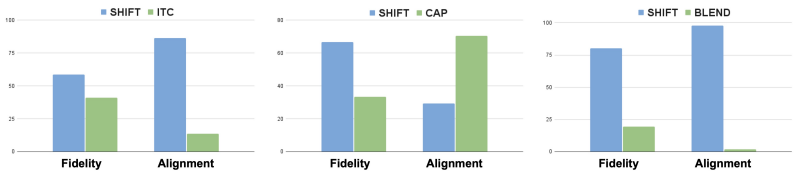

Figure 3: Human evaluation for photo-realism and condition-alignment.

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.

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.

REFERENCES

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18208–18218, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Jiho Jang, Chaerin Kong, Donghyeon Jeon, Seonhoon Kim, and Nojun Kwak. Unifying vision-language representation space with single-tower transformer. *arXiv preprint arXiv:2211.11153*, 2022.

Chaerin Kong, DongHyeon Jeon, Ohjoon Kwon, and Nojun Kwak. Leveraging off-the-shelf diffusion model for multi-attribute fashion image manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 848–857, 2023.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022a.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15671–15680, 2022.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16793–16803, 2022.