Distilling Knowledge from Text-to-Image Generative Models Improves Visio-Linguistic Reasoning in CLIP

Anonymous ACL submission

Abstract

Image-text contrastive models like CLIP have wide applications in zero-shot classification, image-text retrieval, and transfer learning. However, they often struggle on compositional visio-linguistic tasks (e.g., attribute-binding or object-relationships) where their performance is no better than random chance. To address this, we introduce SDS-CLIP, a lightweight and sample-efficient distillation method to enhance CLIP's compositional visio-linguistic reasoning. Our approach fine-tunes CLIP using a distillation objective borrowed from large text-to-image generative models like Stable-Diffusion, which are known for their strong visio-linguistic reasoning abilities. On the challenging Winoground benchmark, SDS-CLIP improves the visio-linguistic performance of various CLIP models by up to 7%, while on the ARO dataset, it boosts performance by up to 3%. This work underscores the potential of well-designed distillation objectives from generative models to enhance contrastive image-text models with improved visio-linguistic reasoning capabilities.

1 Introduction

In recent years, multimodal models like CLIP (Radford et al., 2021a) have excelled in tasks such as zero-shot classification, image-text retrieval, and image-captioning (Mu et al., 2021; Yu et al., 2022; Li et al., 2022; Mokady et al., 2021). These models are also crucial components in various state-of-the-art pipelines for tasks like segmentation and object detection (Wang et al., 2021; Lüddecke and Ecker, 2021; Minderer et al., 2022; Zhong et al., 2021). However, they struggle with visio-linguistic reasoning tasks, such as determining the spatial relationships between objects in an image (Yuksekgonul et al., 2023; Huang et al., 2023). Notably, CLIP's performance on the challenging Winoground (Thrush et al., 2022; Diwan et al., 2022), a benchmark designed to assess visio-linguistic reasoning, is close to random chance. This shortcoming is attributed to CLIP's contrastive objective which prioritizes shortcuts for retrieval, and thus impacts its ability to understand fine-grained object details and their positions (Diwan et al., 2022; Thrush et al., 2022).

In contrast, text-to-image models like Stable Diffusion (Rombach et al., 2021) excel in visio-linguistic tasks, likely due to their text conditioning enhanceing semantic consistency in its cross-attention maps (Li et al., 2023; Clark and Jaini, 2023). Li et al. (2023) recently demonstrated this on the Winoground benchmark, reliably matching captions to images with fine-grained spatial differences using denoising diffusion scores (see Fig 1). Similar results have been shown for other text-to-image models, including Imagen (Clark and Jaini, 2023), with almost all of these methods outperforming CLIP variants on the same tasks.

While these works have shown the potential of using generative text-to-image models for visio-linguistic tasks, it remains computationally intensive. For instance, computing the denoising diffusion score for image-text matching involves multiple passes through a UNet model (approximately 892M parameters) with varying noise levels and time-steps. On an entry-level GPU, this can take up to a minute for a single image-text matching task, making it impractical for real-world and real-time applications. In contrast, CLIP models can classify images up to 18 times faster (see Fig 1), requiring only one pass through both image and text encoders. A promising research direction, therefore, lies in finding methods that combine the strong visio-linguistic capabilities of text-to-image models with the rapid inference of CLIP.

To this end, we introduce SDS-CLIP, a lightweight and sample-efficient fine-tuning approach for CLIP which distills knowledge from Stable Diffusion, and enhances CLIP's visio-reasoning capabilities. Specifically, we add a regularization term to CLIP's standard contrastive loss based

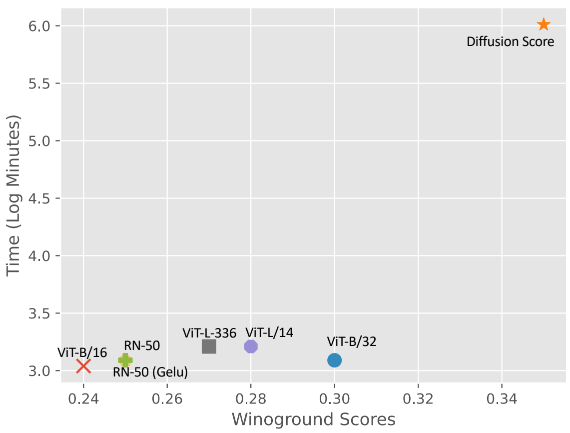


Figure 1: **CLIP variants underperform on Winoground, a visio-linguistic reasoning benchmark, compared to Diffusion Score from Stable Diffusion.** The diffusion score is computed from Stable Diffusion's loss function. Note that Diffusion Score takes $18\times$ more time than CLIP variants for inference (using 50 samplings during diffusion score computation).

on score-distillation sampling (SDS) (Poole et al., 2022). This regularization encourages CLIP's embeddings to be aligned with the denoising diffusion loss from a text-to-image model. By fine-tuning CLIP with this regularized objective on a small paired image-text dataset, specifically 118k image-text pairs from MS-COCO, we demonstrate an 1.5-7% performance gain compared to vanilla CLIP on Winoground and ARO, two highly challenging visio-linguistic reasoning benchmarks. Notably, this is achieved by only updating CLIP's Layer-Norm parameters. Furthermore, we show that SDS-CLIP's zero-shot performance is not impacted on a wide range of downstream datasets.

In summary, our contributions are as follows:

- We introduce SDS-CLIP, a novel sample-efficient and parameter-efficient fine-tuning method that integrates a distillation-based regularization term from text-to-image models.
- We empirically validate our approach on challenging benchmarks and demonstrate an improvement in CLIP's visio-linguistic reasoning, without harming its zero-shot capabilities.

2 Denoising Diffusion Score for Visio-Linguistic Reasoning

The Winoground benchmark establishes a challenging image-text matching task to measure a model's visio-linugistic reasoning abilities: given an image x, the model must match it with the correct caption c^* from a set of captions $C = \{c_i\}_{i=1}^n$, where all caption contains the same words but each describes a different spatial arrangement of the objects, with only one being correct. Concurrent works (Clark and Jaini, 2023; Li et al., 2023; Krojer et al., 2023) to this paper have showed that it is possible to use the denoising diffusion score from text-to-image generative models to perform such an image-matching task. This can be formalized as follows: for an image x and caption c, the denoising diffusion score, denoted by $d(x, c)$, is defined as:

$$d(x, c) = \mathbb{E}_{t \sim T, \epsilon \sim \mathcal{N}(0,I)}[\|\epsilon_\theta(v_\alpha(x), t, c) - \epsilon\|^2] \quad (1)$$

This denoising diffusion score can then be used to select a correct caption c^* from C as:

$$c^* = \arg\min_{c \in C} \mathbb{E}_{t \sim T, \epsilon \sim \mathcal{N}(0,I)}[\|\epsilon_\theta(v_\alpha(x), t, c) - \epsilon\|^2] \quad (2)$$

where t is the sampled time-step, ϵ_θ is the noise prediction UNet, v_α is an encoder (e.g., VQ-VAE) which maps the image x to a latent code and ϵ is the sampled Gaussian noise. Previous works (Krojer et al., 2023) have demonstrated that by adopting this approach, text-to-image models performing strongly on visio-linguistic reasoning benchmarks like Winoground, outperforming contrastive models like CLIP by a significant margin (see Fig 1). For ARO, we obtain an accuracy of 0.63 with the diffusion score which is better than CLIP models.

3 SDS-CLIP: Our Method

The core idea of our approach is to regularize the contrastive objective in CLIP with the denoising diffusion score from Stable Diffusion (see Eq.(1)). Our method builds on the recent work of (Poole et al., 2022) which maps the output of a 3D NeRF model into the input space of Stable Diffusion's UNet and optimizes its parameteres with the denoising diffusion loss, also known as the score-distillation sampling (SDS). In a similar vein, we fine-tune the parameters of CLIP using SDS. Intuitively, our set-up can be viewed as a form of knowledge distillation where the teacher is the text-to-image model and the student is CLIP. As a result, in inference, CLIP can benefit from the visio-linguistic reasoning capabilities that are already learned by text-to-image diffusion models.

Formally, we map the output of CLIP's image encoder to the input space of Stable Diffusion's UNet. Specifically, we pass a given image x through

Model	**Wino-Overall**	Object	Relation	Both	1 Main Pred	2 Main Preds	**ARO-Overall**	ARO-Relation	ARO-Attribution
ViT-B/16(CLIP)	0.24	0.28	0.18	0.57	0.29	0.11	0.57	0.52	0.62
FT with L_{CLIP}	0.23	0.27	0.19	0.56	0.30	0.11	0.56	0.51	0.62
FT with $L_{CLIP} + L_{SDS}$	**0.31**	**0.35**	**0.25**	**0.69**	**0.36**	**0.16**	**0.58**	**0.535**	**0.63**
ViT-B/32(CLIP)	0.30	0.35	0.22	0.80	0.34	0.18	0.55	0.50	0.61
FT with L_{CLIP}	0.28	0.31	0.20	0.76	0.31	0.16	0.55	0.50	0.60
FT with $L_{CLIP} + L_{SDS}$	**0.32**	**0.38**	**0.23**	0.69	**0.36**	**0.20**	**0.575**	**0.53**	**0.62**
ViT-L/14(CLIP)	0.28	0.27	0.25	0.57	0.29	0.24	0.57	0.53	0.61
FT with L_{CLIP}	0.26	0.27	0.25	0.56	0.30	0.23	0.57	0.53	0.61
FT with $L_{CLIP} + L_{SDS}$	**0.295**	**0.32**	0.25	0.53	**0.32**	0.18	**0.595**	**0.55**	**0.64**
ViT-L/14-336(CLIP)	0.27	0.32	0.21	0.57	0.30	0.19	0.57	0.53	0.61
FT with L_{CLIP}	0.23	0.28	0.19	0.53	0.26	0.17	0.57	0.53	0.61
FT with $L_{CLIP} + L_{SDS}$	**0.285**	**0.34**	**0.23**	0.56	**0.31**	**0.21**	**0.585**	**0.54**	**0.63**
ResNet-50(CLIP)	0.25	0.29	0.19	0.5	0.27	0.18	0.58	0.53	0.63
FT with L_{CLIP}	0.24	0.27	0.20	0.49	0.27	0.16	0.575	0.52	0.63
FT with $L_{CLIP} + L_{SDS}$	**0.265**	**0.30**	**0.21**	0.42	**0.29**	**0.19**	**0.60**	**0.55**	**0.66**

Table 1: **Our fine-tuning method SDS-CLIP improves CLIP performance on the Winoground benchmark by 1.5% to 7% and upto 3% for the ARO-Relation and Attribution tasks across various CLIP variants**. Specifically, we find that our method improves on the sub-categories involving *object-swap* and *relational* understanding which comprise of the majority of the tasks in Winoground. Note that *only* fine-tuning with image-text pairs from MS-COCO without the distillation loss does not lead to any improvements. OpenCLIP results in Appendix I.

CLIP's image encoder f_ϕ and map its <CLS> embedding through a linear map $h_w \in \mathcal{R}^{d\times4\times64\times64}$ into the input space of Stable Diffusion's UNet ϵ_θ. This can be formalized as $\epsilon_\theta(h_w(f_\phi(x)), t, c)$ where t is the time step and c is the corresponding text caption for the given image. We then use this term in place of $\epsilon_\theta(v_\alpha(x), t, c)$ in Eq. (2) to arrive as a denoising diffusion loss L_{SDS} which encourages image-text binding with feedback from the diffusion loss:

$$L_{SDS} = \mathbb{E}_{t\sim T,\epsilon\sim\mathcal{N}(0,I)}[||\epsilon_\theta(h_w(f_\phi(x)), t, c) - \epsilon||^2 \quad (3)$$

We practically implement this by adding this L_{SDS} loss to the original contrastive objective of CLIP such that it acts as a regularizer:

$$L_{total} = L_{CLIP} + \lambda L_{SDS} \quad (4)$$

where L_{CLIP} is defined in Appendix C.1 and λ is a hyper-parameter that can be set with a grid search. We note that there are multiple ways to incorporate a diffusion loss into CLIP's objective. We found that as an additional loss term led to the best results, however, we include the full set of design choices we considered in the Appendix.

Similar to differentiable image parameterizations (Mordvintsev et al., 2018) where a given function is optimized by backpropogation through the image generation process, the UNet parameters θ are kept frozen during the optimization process. Specifically, given $L_{total}(\phi, \gamma, w, \theta)$:

$$\phi*, \gamma*, w* = \min_{\phi,\gamma,w} L_{total}(\phi, \gamma, w, \theta) \quad (5)$$

where ϕ, γ, w are the learnable parameters of CLIP's image encoder, text encoder and the linear map between CLIP and Stable Diffusion's UNet.

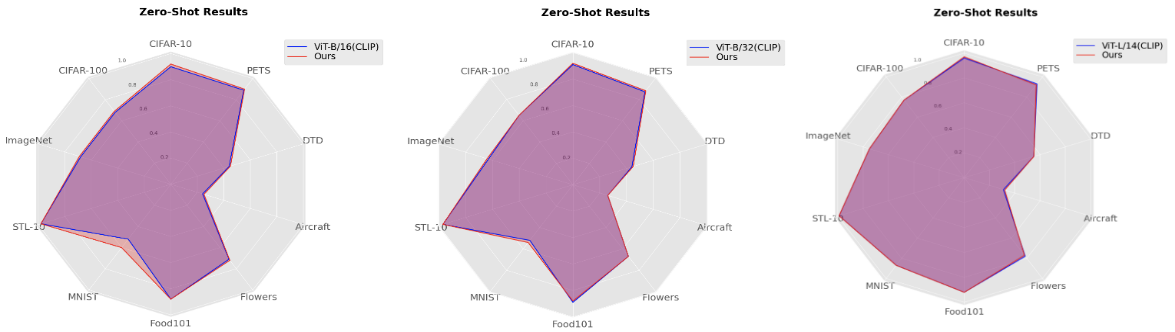
4 Experiments

In this section, we empirically validate our proposed method SDS-CLIP on two types of tasks: i) visio-linguistic reasoning using two challenging benchmarks (Winoground, ARO) and ii) zero-shot image classification using a suite of downstream datasets (ImageNet, CIFAR-100, and others). Overall, we show that our method improves CLIP's performance significantly on Winoground and some key tasks in ARO, while also marginally improving downstream zero-shot classification performance.

4.1 Experimental Setup

CLIP Models. We consider the following CLIP variants in our experiments: (i) CLIP ViT-B/16; (ii) CLIP ViT-B/32; (iii) CLIP-ViT-L-14; (iv) CLIP-ViT-L-14 336px; (v) CLIP-ResNet-50.

Implementation Details. Due to computational limit, we fine-tune CLIP from a publicly available checkpoint instead of training from scratch. Notably, we only fine-tune CLIP's LayerNorm parameters following (Basu et al., 2023) along with the linear transformation h_w – accounting for only $\approx 8M$ trainable parameters. We fine-tune these parameters using image-text pairs from MSCOCO (Lin et al., 2014). In particular, we choose MSCOCO as it is relatively small and less noisy than other image-text datasets such as CC-12M (Sharma et al., 2018). Both these factors make our fine-tuning method extremely sample- and parameter-efficient.



(a) ViT-B/16 (b) ViT-B/32 (c) ViT-L/14

Figure 2: **Our fine-tuning method does not harm the zero-shot abilities of CLIP.** In fact for certain downstream datasets (e.g., ImageNet, CIFAR-10, MNIST, Aircraft) – we observe an improvement in the zero-shot performance between $1\% - 8\%$ for ViT-B/16. For other CLIP models, we find no drop in zero-shot performance.

Baselines. We compare our method with two different baselines: (i) pre-trained (vanilla) CLIP checkpoints; and (ii) CLIP fine-tuned on MS-COCO with the standard contrastive loss without the regularization term.

4.2 Results

Winoground. We evaluate SDS-CLIP on the challenging visio-linguistic reasoning benchmark, Winoground (Thrush et al., 2022). In Table (1), we find that our approach consistently improves performance across all Winoground sub-categories and CLIP variants, yielding absolute improvements ranging from 1.5% to 7%. The largest gain of 7% is observed in ViT-B/16 (CLIP), with other CLIP variants showing consistent improvements of 1.5% to 2%. In the Appendix(Table 2), we provide results for CLIP variants pre-trained on public data, where similar improvements are observed. On further inspection of the Winoground sub-categories, we find that SDS-CLIP shows consistent improvements in “object-swap" and “relation". It is worth noting that the “both” sub-category, which combines both “object-swap" and “relation" tags, makes up only $\tilde{5}\%$ of all tasks, thus are potentially not fully representative of all scenarios involving both object swaps and relational understanding. We also analyse SDS-CLIP’s robustness to the number of predicates in captions and find that overall, it enhances performance in tasks where there are both one and two predicates.

ARO. The ARO dataset (Yuksekgonul et al., 2023) comprises tasks for (i) attribute-understanding and (ii) relational-understanding. In Table 1, we find that SDS-CLIP enhances performance by 1%-3% in the "attribute-binding" and "relational understanding" tasks.

Impact on CLIP’s zero-shot performance. From Fig 2, we find that SDS-CLIP’s zero-shot classification capbilities are not impacted, relative to vanilla CLIP. In fact, we find that ViT-B/16’s zero-shot performance improves across a range of downstream datasets (with up to 8% improvement for MNIST).

While Stable-Diffusion is pre-trained on a much larger set of image-text pairs than CLIP, in Appendix K, we show that the CLIP variants pre-trained on LAION-2B still suffer on Winoground. In fact, we show that using SDS-CLIP can improve compositional reasoning of such CLIP variants. In Appendix H, we show results with fine-tuning on the larger CC-3M (Sharma et al., 2018).

5 Related Works

While CLIP models (Radford et al., 2021a) are renowned for their robust zero-shot classification, recent research (Thrush et al., 2022; Diwan et al., 2022) has exposed their limitations in visio-linguistic reasoning. In contrast, recent studies have demonstrated that text-to-image models (Clark and Jaini, 2023; Li et al., 2023; Krojer et al., 2023; Chen et al., 2023) outperform CLIP in reasoning tasks. These models in fact leverage scores computed from the diffusion objective. We note that while (Poole et al., 2022) use score-distillation sampling for text to 3D generation, ours is the first work to adapt the formulation as a regularizer and improve compositional abilities in CLIP.

6 Conclusion

Our paper introduces SDS-CLIP, a novel data and parameter-efficient method that effectively enhances CLIP’s visio-linguistic reasoning abilities by distilling knowledge from text-to-image models, without compromising its zero-shot abilities.

7 Limitations

The primary limitation of our method is the inability to use large batch-sizes on moderate size GPUs. This is due to the fact that the regularizer L_{SDS} requires a full backward pass through the UNet, even though its parameters are frozen. We also find that while the original diffusion score is good at *object-understanding, attribute-understanding* and *relational-understanding* tasks, it does not perform well on ordering tasks from the ARO dataset. For this reason, distillation from Stable-Diffusion potentially may not be effective in improving CLIP's performance on ordering tasks. Similar results are also observed in concurrent works such as (Krojer et al., 2023).

8 Ethical Considerations

Vision-language models such as CLIP have been known for inheriting biases (Agarwal et al., 2021) due to their training data. Our work uses a well-known widely used dataset (MS-COCO) for the fine-tuning procedure and therefore does not introduce any additional bias. In fact, our distillation method mitigates some of the inherited bias in CLIP which earlier did not lead to good reasoning capabilities.

References

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. 2021. Evaluating CLIP: towards characterization of broader capabilities and downstream implications. *CoRR*, abs/2108.02818.

Samyadeep Basu, Daniela Massiceti, Shell Xu Hu, and Soheil Feizi. 2023. Strong baselines for parameter efficient few-shot fine-tuning.

Huanran Chen, Yinpeng Dong, Zhengyi Wang, X. Yang, Chen-Dong Duan, Hang Su, and Jun Zhu. 2023. Robust classification via a single diffusion model. *ArXiv*, abs/2305.15241.

Kevin Clark and Priyank Jaini. 2023. Text-to-image diffusion models are zero-shot classifiers.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. 2022. Why is winoground hard? investigating failures in visuolinguistic compositionality.

Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Tangjie Lv, Zhipeng Hu, and Wen Zhang. 2023. Structure-clip: Enhance multi-modal language representations with structure knowledge.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2016. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *CoRR*, abs/1612.06890.

Benno Krojer, Elinor Poole-Dayan, Vikram Voleti, Christopher Pal, and Siva Reddy. 2023. Are diffusion models vision-and-language reasoners?

Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. 2023. Your diffusion model is secretly a zero-shot classifier.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. *CoRR*, abs/2201.12086.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Timo Lüddecke and Alexander S. Ecker. 2021. Prompt-based multi-modal image segmentation. *CoRR*, abs/2112.10003.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. 2022. Simple open-vocabulary object detection with vision transformers.

Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. Clipcap: Clip prefix for image captioning.

Alexander Mordvintsev, Nicola Pezzotti, Ludwig Schubert, and Chris Olah. 2018. Differentiable image parameterizations. *Distill*. Https://distill.pub/2018/differentiable-parameterizations.

Norman Mu, Alexander Kirillov, David A. Wagner, and Saining Xie. 2021. SLIP: self-supervision meets language-image pre-training. *CoRR*, abs/2112.12750.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021a. Learning transferable visual models from natural language supervision.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021b. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. 2021. CRIS: clip-driven referring image segmentation. *CoRR*, abs/2111.15174.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. 2023. Open-vocabulary panoptic segmentation with text-to-image diffusion models.

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

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. 2021. Regionclip: Region-based language-image pretraining. *CoRR*, abs/2112.09106.

A Benchmark Datasets

A.1 Benchmark datasets

Winoground (Thrush et al., 2022; Diwan et al., 2022) is a challenging vision-language dataset for evaluating the visio-linguistic characteristics of contrastively trained image-text models. The dataset consists of 400 tasks, where each task consists of two image-text pairs. The objective is to independently assign the correct text caption to each image. Each task is also annotated with meta-data corresponding to whether the task requires object-understanding, relational-understanding or both. The tasks in Winoground are challenging as the images differ in fine-grained ways and assigning the correct text captions requires inherent compositional visual reasoning.

ARO (Yuksekgonul et al., 2023) similarly tests visio-linguistic reasoning and consists of three types of tasks: (i) Visual Genome Attribution to test the understanding of object properties; (ii) Visual Genome Attribution to test for relational understanding between objects; and (iii) COCO-Order and Flickr30k-Order to test for order sensitivity of the words in a text, when performing image-text matching. We highlight that Winoground though slightly smaller in size than ARO is more challenging as it requires reasoning beyond visio-linguistic compositional knowledge (Diwan et al., 2022).

A.2 Does distilling features directly from UNet help?

Previous works such as (Xu et al., 2023) find that the frozen features of the UNet contain structural information about the image. Motivated by this, we also investigate if distilling knowledge directly from the frozen UNet features is beneficial, Given an image x and its caption c, the frozen features f from the UNet (where $I(x, c) = \epsilon_\theta(v_\alpha(x), t, c)$, similar to (Xu et al., 2023)) can be extracted. We then use these frozen internal representations from the UNet to regularize features of the image encoder in CLIP. In particular:

$$L_{total} = L_{CLIP} + \lambda \| h_w(f_\phi(x) - I(x, c)) \|_2^2 \quad (6)$$

However, we find that distillation in this way does not lead to improved performances for visio-linguistic reasoning. In fact, for ViT-B/16 (CLIP) we find the Winoground score to decrease from 0.24 to 0.23. This result shows that using score-distillation sampling which involves backpropogation through the UNet is critical to distill knowledge from diffusion models to other discriminative models.

B SDS-CLIP: Algorithm

Algorithm 1 Algorithm to fine-tune CLIP with distillation from Stable-Diffusion for improved visio-linguistic reasoning

Require: $\mathcal{D}$: image-text pairs, f_ϕ: CLIP's image-encoder, g_γ: CLIP's text-encoder, ϵ_θ: UNet; N: Number of Epochs; λ: Hyper-parameter for the regularizer; $|B|$: Batch-size.

```
while i ≠ N do
    {x_j, y_j}_{j=1}^{|B|} ← Sample a batch from D
    t ← Sample time-steps using DDPM
    ε ← Sample Gaussian noise ε ~ N(0, I)
    L_clip ← Compute contrastive loss as in eq. (7)
    L_SDS ← Compute SDS loss as in eq. (3)
    L_total ← L_clip + λ L_SDS
    L_total.backward()                          ▷ Backprop
    φ, γ, w ← Update the relevant parameters
    i ← i + 1
end while
```

C Preliminaries

C.1 CLIP

CLIP (Radford et al., 2021b) is a image-text model which is pre-trained using a contrastive objective, typically on internet-scale data. The core intuition of the training objective is to align the text and image embeddings of image-text pairs in a shared embedding space. To do this, CLIP consists of two components: (i) an image encoder f_ϕ which transforms a raw image x_i into an image embedding $e_{img}(x_i) = f_\phi(x_i) \in \mathbb{R}^d$, also denoted by the <CLS> token; and (ii) a text encoder g_γ which transforms a raw text caption c_i into a text embedding $e_{text}(c_i) = g_\gamma(c_i) \in \mathbb{R}^d$ also denoted by <EOS> token, both of which map to an embedding dimensionality d. Given a dataset $\mathcal{D} = \{(x_i, c_i)\}_{i=1}^N$ of image-text pairs, where (x_i, y_i) is the i^{th} image-text pair, CLIP uses a contrastive objective to pull the image and text embeddings of matched pairs together, while pushing those of unmatched pairs apart. Formally, the contrastive objective can be defined as:

$$L_{CLIP} = L_{image-text} + L_{text-image} \quad (7)$$

where:

$$L_{image-text} = -\frac{1}{2N}\sum_{j=1}^{N}\log\{\frac{\exp(e_{img}(x_j)^T e_{text}(c_j)/\tau)}{\sum_{k=1}^{N}\exp((e_{img}(x_j)^T e_{text}(c_k)/\tau))}\} \quad (8)$$

$$L_{text-image} = -\frac{1}{2N}\sum_{j=1}^{N}\log\{\frac{\exp(e_{img}(x_j)^T e_{text}(c_j)/\tau)}{\sum_{k=1}^{N}\exp((e_{img}(x_k)^T e_{text}(c_j)/\tau))}\} \quad (9)$$

where τ is a trainable temperature parameter. Usually $\mathcal{D}$ is an internet-scale dataset consisting of millions of image-text pairs. Furthermore, during pre-training, the embeddings $e_{img}(x_i)$ and $e_{text}(c_i)$ are normalized to have a unit-norm.

D When does distillation not help CLIP?

While we find that distilling knowledge from Stable-Diffusion to CLIP helps in *object-swap*, *relational-understanding* and *attribution-binding* visio-linguistic tasks, it does not help on tasks where the order of the text is perturbed (e.g. the COCO-Order and Flickr-Order tasks in the ARO dataset). In fact, we find that the denoising diffusion score in Equation (1) leads to accuracies of 0.24 for COCO-Order and 0.34 for Flickr-Order which is in fact lower than CLIP models. Concurrent works (Krojer et al., 2023) has shown similarly low performance for text-ordering tasks. A potential reason could be that ordering tasks only test for grammatical understanding which current text encoders cannot effectively model. Another reason could be that the denoising diffusion score is not affected by word ordering as the image semantics are not changed as a result.

E Notes on Fine-tuning Dataset

We use MS-COCO (Lin et al., 2014) which is widely used for multimodal learning. This dataset does not contain any names or uniquely identifies individual people or offensive content.

F More Experimental Details

Hyper-parameters. We perform a hyperparameter sweep for the learning rate and the regularization hyperparameter λ for ViT-B/16. We use these same hyperparameters for different CLIP variants including ViT-B/32, ViT-B/14, ViT-L/14-336px and ResNet-50. In particular, we set $\lambda = 0.001$ and set the learning rate as 5×10^{-5}. We use a batch-size of 32 for all the different CLIP models. We use Stable-Diffusion v1-4 as the teacher model in our experiments.

Note on Full Fine-tuning. All our experiments were primarily done by fine-tuning only the LayerNorm parameters. In the initial phase of the project, we also fine-tune all the parameters of the text and image encoder in CLIP, however it results in worse performances than those reported in Table. (1). Potentially, this can be due to overfitting issues when used in conjunction with the new regularizer. We therefore run all the experiments with LayerNorm tuning as it leads to the best results.

Model	**Overall**	Object	Relation	Both	1 Main Pred	2 Main Preds
ViT-B/16(LAION 400M)	0.24	0.29	0.17	0.59	0.28	0.11
COCO FT with L_{CLIP}	0.24	0.26	0.21	0.54	0.31	0.10
COCO FT with $L_{CLIP} + L_{SDS}$	**0.30**	**0.34**	**0.23**	0.55	**0.33**	**0.14**

Table 2: **Additional results on Winoground with ViT-B/16 CLIP pre-trained on public data (LAION-400M).**

Total GPU Hours. For all our experiments we use NVIDIA-A6000 and each fine-tuning experiment takes ≈6 hours.

G Additional Results with Stable-Diffusion-v2-1

In particular, with our distillation strategy with Stable-Diffusion v-2.1 as a teacher – we obtain the following results on Winoground: (i) ViT-B/16: 0.35; (ii) ViT-B/32: 0.33; (iii) ViT-L/14: 0.31; (iv) ViT-L/14-336px: 0.31; (iv) ResNet-50: 0.28; All the scores are higher than the fine-tuned model with Stable-Diffusion-v1-4 as the teacher, therefore highlighting that a teacher with better compositional generation capabilities will be a better choice.

H Fine-tuning with Conceptual Captions

We primarily use MS-COCO as : (i) It's a relatively small dataset which can keep the fine-tuning steps relatively smaller and scaling the fine-tuning dataset will increase fine-tuning time; (ii) It's a well-established, relatively diverse and well annotated image-text dataset which is used by the community. We also fine-tuned with CC-3M (Sharma et al., 2018), but found the improvements to be similar in lines to that using MS-COCO. For e.g., On Winoground with CC-3M, we find the following performance after distillation with Stable-Diffusion-v1-4: (i) ViT-B/16: 0.32; (ii) ViT-B/32: 0.32; (iii) ViT-L/14: 0.30; (iv) ViT-L/14-336px: 0.28; (iv) ResNet-50: 0.27. These scores are only marginally better than using MS-COCO, although the dataset size is more than 30 times – which shows that a high-quality dataset such as

Model	**Overall**	Object	Relation	Both	1 Main Pred	2 Main Preds
ViT-B/16(LAION 2B)	0.27	0.32	0.19	0.61	0.29	0.12
COCO FT with $L_{CLIP} + L_{SDS}$	**0.31**	**0.36**	**0.24**	0.53	**0.36**	**0.17**

Table 3: **CLIP (Pre-trained with 2B images) still underperforms on Winoground.** We show the CLIP even when trained with LAION-2B (similar scale of training data as Stable-Diffusion) still underperforms the diffusion score from Stable-Diffusion. This shows that scale of data alone cannot be useful in mitigating reasoning capabilities in CLIP.

MS-COCO is sufficient for improving compositional abilities in CLIP.

I Results with OpenCLIP

In Table 2, we show that our method is compatible with OpenCLIP. In particular, we find that distillation to OpenCLIP improves its visio-linguistic score from 0.24 to 0.30. These results highlight the generalizability of our distillation method.

J Additional Results on CLEVR

We apply our fine-tuned model on the CLEVR task (Johnson et al., 2016) – which consists of images of 3D shapes isolating phenomena such as spatial reasoning or attribute binding. We find that the diffusion-score leads to a score of 0.67, whereas the best CLIP variant in our test-bed (CLIP ViT-L/14) scored 0.63. With our distillation loss during fine-tuning – this score improved to 0.65 with a 2% gain.

K Is it the Scale of Pre-Training Data Which Helps?

In Table 3, we show that CLIP models even when trained at the same scale of pre-training data as Stable-Diffusion (LAION-2B) struggle on the Winoground dataset. We specifically highlight that CLIP (when pre-trained on 2B image-text pairs) obtain a score of 0.27, whereas the diffusion model when trained on similar pre-training corpus obtains a score of 0.35. This clearly shows that at a similar pre-training scale, diffusion models (with their diffusion objective) are better compositional learners than CLIP like models. Our distillation method from Stable-Diffusion improves the Winoground score from 0.27 to 0.31 on CLIP(pre-trained on 2B image-text pairs).

L Beyond CLIP

We find that Open-CoCa (Yu et al., 2022) pre-trained on 2B image-text pairs obtains a score of 0.30 on Winoground. With our distillation strategy, we find that the score improves to 0.33 highlighting that our distillation strategy can be used for models beyond CLIP. A full investigation of the impact of our distillation method on various vision-language models is deferred towards future work.