Supplementary material for
Text-to-Image Diffusion Models are Great Sketch-Photo Matchmakers

Subhadeep Koley[1,2] Ayan Kumar Bhunia[1] Aneeshan Sain[1] Pinaki Nath Chowdhury[1]
Tao Xiang[1,2] Yi-Zhe Song[1,2]
[1]SketchX, CVSSP, University of Surrey, United Kingdom.
[2]iFlyTek-Surrey Joint Research Centre on Artificial Intelligence.
{s.koley, a.bhunia, a.sain, p.chowdhury, t.xiang, y.song}@surrey.ac.uk

A. Time and computational complexity

Unlike the *iterative inference* of text-to-image generation via stable diffusion (SD) model [3], diffusion-based feature extraction needs a *single-step inference* (Sec. 4). Most importantly, instead of running SD model six times, we resort to an *efficient implementation* where we repeat the query sketch tensor six times along batch dimension and use a set of different random noises to extract six distinct SD features *simultaneously* in *one* step. Thus, the complexity and runtime *do not* scale *linearly*. Instead, it takes a similar running time compared to competing SOTAs for the same input size. For instance, our diffusion feature extraction (with ensembling) takes 0.85ms *vs*. ZS-LVM's [4] 0.83ms or B-Triplet+VP (VGG)'s 76ms for a 224×224 image on a single Nvidia V100 GPU. Performing feature ensembling to boost performance and stability (Sec. 7) would increase the inference time slightly ($0.82 \rightarrow 0.85$ms). However, in case of a computation bottleneck, one may avoid this with a slight dip in performance (*e.g*., *Sketchy*: mAP@200 $0.746 \rightarrow 0.725$; *TU-Berlin*: mAP@all $0.680 \rightarrow 0.671$; *Quick, Draw!*: mAP@all $0.231 \rightarrow 0.220$). Notably, even without feature ensembling, our method surpasses the next best method (*i.e*. ZS-LVM [4]) on *all* 3 benchmark datasets. Consequently, we leave the choice of utilising this gain provided by ensembling (at a slight cost of inference time) to the end-users.

B. Performance-complexity trade-off

Even *with* feature ensembling, our method takes 0.85ms to extract a query-sketch feature (for a 224×224 sketch) compared to 0.83ms of our closest competitor (ZS-LVM [4]), which is only $\sim$2.4% higher, yet boosts Acc.@1 by 11.4% (ZS-FG-SBIR on Sketchy). While ZS-LVM [4] takes 9.46G FLOPs (CLIP-ViT-B/32) to process a sketch of size 224×224, our method uses 1.29G FLOPs, which is $7.33\times$ lower, while boosting mAP@all by 14.4% on the Quick, Draw! dataset.

C. Ablating Stable Diffusion versions

We ablate multiple SD [3] versions on Sketchy [5] dataset in Tab. 1. While SD v1.x models utilise CLIP [2] text encoder during their pre-training, v2.x models resort to much larger-scale OpenCLIP [1]. Evidently, SD v2.x models perform better than v1.x ones with v2.1 achieving the highest score. This is likely due to v2.x models' adaptation of the much larger-scale OpenCLIP [1] encoder during pre-training.

Table 1. Ablating SD versions.

SD version	Sketchy [5]	
	mAP@200	Acc.@1
v1.4	0.726	28.93
v1.5	0.730	29.81
v2.0	0.738	30.21
v2.1 (Ours)	**0.746**	**31.94**

D. Result across different ensemble sizes

Fig. 1 depicts qualitative results for ZS-FG-SBIR on Sketchy across different runs with different ensemble sizes.

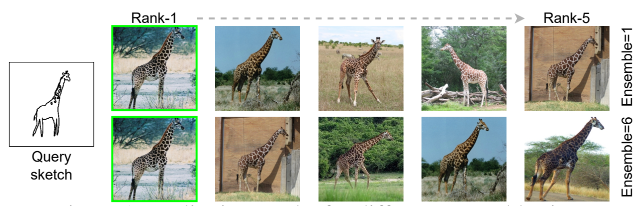

Figure 1. qualitative results for different ensemble sizes.

References

[1] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023. 1
[2] 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 *ICML*, 2021. 1
[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 2022. 1
[4] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Subhadeep Koley, Tao Xiang, and Yi-

Zhe Song. CLIP for All Things Zero-Shot Sketch-Based Image Retrieval, Fine-Grained or Not. In *CVPR*, 2023. 1

[5] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The Sketchy Database: Learning to Retrieve Badly Drawn Bunnies. *ACM TOG*, 2016. 1