A. Website with results, videos, and benchmark

Our project website at https://ajayj.com/vectorfusion includes videos of the optimization process and many more qualitative results in SVG format. The benchmark used for evaluation consists of 128 diverse prompts sourced from prior work and is available at https://ajayj.com/vectorfusion/svg_bench_prompts.txt.

B. Ablation: Reinitializing paths

We reinitialize paths below an opacity threshold or area threshold periodically, every 50 iterations. The purpose of reinitializing small, faint paths is to encourage the usage of all paths. Path are not reinitialized for the final 200-500 iterations of optimization, so reinitialized paths have enough time to converge. Reinitialization is only applied during image-text loss computation stages. Note that for SD + LIVE + SDS, we only reinitialize for SDS finetuning, not LIVE image vectorization. For CLIPDraw, we only reinitialize paths for our closed Bézier path version, not for the original CLIPDraw version, which consists of open Bézier paths where it is difficult to measure area. We detail the hyperparameters for threshold, frequency, and number of iterations of reinitialization in Table 2.

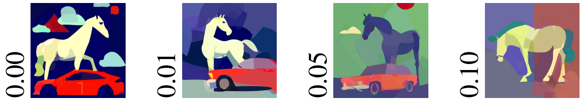
Table 2. Path reinitialization hyperparameters

Method	**Opacity Thresh.**	**Area Thresh.**	**Freq.**	**Iters**
SDS (from scratch)	0.05	0	50	1.5/2K SDS steps
SD + LIVE + SDS	0.05	64 px^2	50	0.8/1K SDS steps
CLIPDraw (icon.)	0.05	64 px^2	50	1.8/2K CLIP steps

Table 3 ablates the use of reinitialization. When optimizing random paths with SDS, reinitialization gives an absolute +3.0% increase in R-Precision according to OpenCLIP H/14 evaluation. When initialized from a LIVE traced sample, reinitialization is quite helpful (+12.5% R-Prec).

C. Ablation: Saturation Penalty

We proposed a saturation penalty for pixel art 4.4. We did not use it for icongraphy, but it greatly reduces saturation, as shown below.



D. Ablation: Number of paths

Table 3. Evaluating path reinitialization with 64 closed, colored Bézier curves, our iconographic setting. VectorFusion reinitializes paths during optimization to maximize their usage. This improves caption consistency both when training randomly initialized paths with SDS (SDS w/ reinit), and when initializing with a LIVE traced Stable Diffusion sample (SD + LIVE + SDS w/ reinit).

		Caption consistency			
		CLIP L/14		OpenCLIP H/14	
Method	K	R-Prec	Sim	R-Prec	Sim
SDS	0	75.0	24.0	75.0	28.8
w/ reinit	0	78.1	24.1	78.1	**29.3**
SD + LIVE + SDS	4	64.8	22.6	68.8	26.7
w/ reinit	4	**78.9**	**29.4**	**81.3**	24.5

VectorFusion optimizes path coordinates and colors, but the number of primitive paths is a non-differentiable hyperparameter. Vector graphics with fewer paths will be more abstract, whereas photorealism and details can be improved with many paths. In this ablation, we experiment with different number of paths. We evaluate caption consistency across path counts. For methods that use LIVE, this ablation uses a path schedule that incrementally adds 2, 4, and 10 for a total of 16 paths, and a path schedule of 8, 16, 32, and 72 for 128 total paths. We set K=4 for rejection sampling. Figure 9 and Table 4 show results. Consistency improves with more paths, but there are diminishing returns.

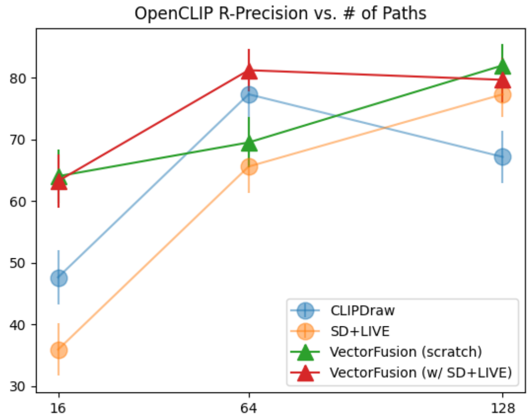


Figure 9. Increasing the number of paths generally improves our caption consistency metrics. We find 64 to be sufficient to express and optimize SVGs that are coherent with the caption.

E. Ablation: Number of rejection samples

In this section, we ablate on the number of Stable Diffusion samples used for rejection sampling. We include results in Figure 10 and Table 5. Rejection sampling greatly improves coherence of Stable Diffusion raster samples with the caption, since rejection explicitly maximizes a CLIP image-text similarity score. After converting the best raster sample

Table 4. **Caption Consistency vs. # Paths**. Increasing the number of paths allows for greater expressivity and caption coherency. However, it also increases memory and time complexity, and we opt for 64 paths to balance between performance and time constraints. We use rejection sampling K=4 for SD+LIVE and SD+LIVE+SDS, and we optimize open Bézier paths for CLIPDraw.

		Caption consistency			
		CLIP L/14		OpenCLIP H/14	
Method	# Paths	R-Prec	Sim	R-Prec	Sim
SDS (scratch)	16	68.0	23.4	64.1	27.4
SD+LIVE	16	33.6	19.7	35.9	22.9
SD+LIVE+SDS	16	63.3	22.9	63.3	27.3
CLIPDraw	16	58.6	23.9	47.7	27.4
SDS (scratch)	64	76.6	24.3	69.5	28.5
SD+LIVE	64	57.0	21.7	59.4	25.8
SD+LIVE+SDS	64	78.9	**29.4**	81.3	24.5
CLIPDraw	64	**85.2**	27.2	77.3	**31.7**
SDS (scratch)	128	83.6	24.8	**82.0**	29.7
SD+LIVE	128	77.3	23.7	77.34	28.7
SD+LIVE+SDS	128	78.1	24.8	79.7	29.7
CLIPDraw	128	73.4	25.7	67.2	30.4

to a vector graphic with LIVE (SD+LIVE), coherence is reduced 10-15% in terms of OpenCLIP H/14 R-Precision. However, using more rejection samples generally improves the SD+LIVE baseline. In contrast, VectorFusion is robust to the number of rejection samples. Initializing with the vectorized result after 1-4 Stable Diffusion samples is sufficient for high SVG-caption coherence.

F. Ablation: Classifier-Free Guidance

We compare guidance scales in Table 6 and Figure 11. This hyperparameter seems fairly robust. While high guidance (>50) can lead to cartoonish generations, it is important for coherence.

G. Ablation: SDS + CLIP Hybrid Losses

We investigate combining the SDS and CLIP losses. Adding an additional CLIP loss improves our CLIP-based metrics in Table 7. Qualitatively, the additional CLIP loss leads to very different generations in Figure 11. While generated SVGs are less saturated, there are many artifacts characteristic of optimizing the CLIP loss.

H. Pixel Art Results

We ablate saturation penalties and different loss objectives in Figure 8. We use K=4 rejection samples for the initial Stable Diffusion raster image. Simply pixelating the best of K Stable Diffusion samples (SD+L1) is a straightforward way of generating pixel art, but results are often unrealistic and not as characteristic of pixel art. For example, pixelation results in blurry results since the SD sample does not use completely regular pixel grids.

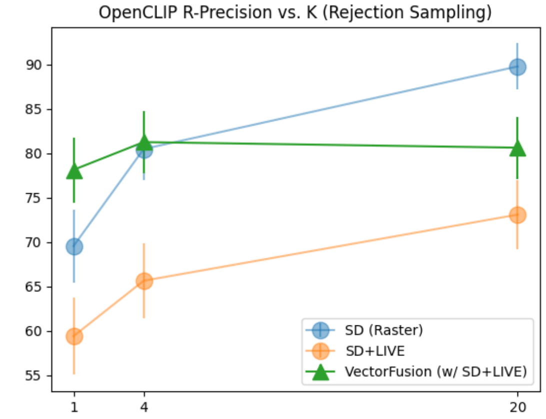


Figure 10. Coherence with the caption improves with additional rejection samples. Even with 20 rejection samples, the vectorized Stable Diffusion image baseline (SD+LIVE) still underperforms VectorFusion with no rejection. VectorFusion also slightly benefits from a better initialization, using 4 rejection samples of the SD initialized image.

Table 5. **Caption Consistency vs. K**. By increasing rejection sampling, we improve Stable Diffusion outputs. This improves both SD and SD+LIVE caption consistency. However, we find that VectorFusion matches Stable Diffusion consistency for K=4 and retains performance for K=20. This suggests that VectorFusion improves upon Stable Diffusion outputs and is robust to different initializations.

		Caption consistency			
		CLIP L/14		OpenCLIP H/14	
Method	K	R-Prec	Sim	R-Prec	Sim
SD (Raster)	1	67.2	23.0	69.5	26.7
SD+LIVE	1	57.0	21.7	59.4	24.8
SD+LIVE+SDS	1	78.1	24.1	78.1	29.3
SD (Raster)	4	81.3	24.1	80.5	28.2
SD+LIVE	4	69.5	22.9	65.6	27.6
SD+LIVE+SDS	4	78.9	**29.4**	81.3	24.5
SD (Raster)	20	**89.1**	25.4	**89.8**	30.1
SD+LIVE	20	71.5	23.6	73.1	28.5
SD+LIVE+SDS	20	79.8	25.0	80.6	**30.3**

Finetuning the result of pixelation with an SDS loss, and an additional L2 saturation penalty, improves OpenCLIP's R-Precision +10.2%. Direct CLIP optimization achieves high performance on CLIP R-Precision and CLIP Similarity, but

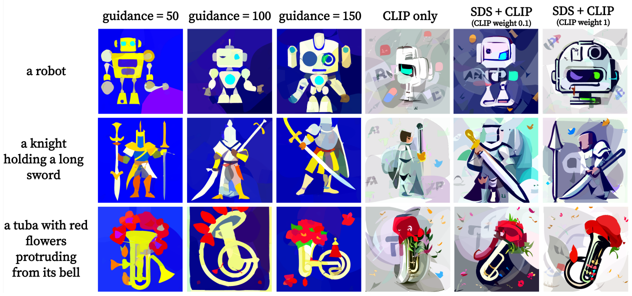

guidance = 50
guidance = 100
guidance = 150
CLIP only
SDS + CLIP (CLIP weight 0.1)
SDS + CLIP (CLIP weight 1)
a robot
a knight holding a long sword
a tuba with red flowers protruding from its bell

Figure 11. **SDS Guidance Scale and CLIP + SDS loss**. Our default guidance is 100. A hybrid CLIP + SDS loss produces samples that are more cohesive than using the CLIP loss only, and samples that are less saturated than using the SDS loss only.

Table 6. **SDS Guidance Scale.** We use SDS + rejection sampling (K=4) and LIVE with 64 paths. Our default guidance is 100, and a guidance scale of 50 leads to the best quantitative evaluation metrics.

	CLIP L/14		OpenCLIP H/14	
Guidance	R-Prec	Sim	R-Prec	Sim
5	21.1	18.0	14.8	16.5
50	**80.5**	25.0	**85.2**	**30.0**
100	78.9	**29.4**	81.3	24.5
150	75.8	24.5	81.3	29.7

Table 7. **SDS + CLIP.** SD + rejection (K=4), LIVE, and 64 paths.

	CLIP	CLIP L/14		OpenCLIP H/14	
Method	Weight	R-Prec	Sim	R-Prec	Sim
SDS	0	78.9	**29.4**	81.3	24.5
CLIP	1	70.1	23.3	74.0	28.4
SDS+CLIP	0.1	89.1	25.3	**90.6**	**32.0**
SDS+CLIP	1	**90.6**	25.3	87.5	31.8

Table 8. **Pixel Art.** We compare CLIP-based optimization and SDS-based optimizations. In addition, we ablate the saturation penalty, which makes pixel art more visually pleasing.

		Caption consistency			
	Sat	CLIP L/14		OpenCLIP H/14	
Method	Penalty	R-Prec	Sim	R-Prec	Sim
SDS (scratch)	0	57.9	21.3	43.8	23.1
SDS (scratch)	0.05	53.9	21.6	42.2	22.7
SD+L1	-	60.9	22.8	52.3	24.5
SD+L1+SDS	0	61.8	23.0	51.6	24.6
SD+L1+SDS	0.05	61.7	21.8	62.5	24.2
CLIP	-	80.5	26.6	73.4	27.5

we note that like our iconographic results, CLIP optimization often yields suboptimal samples.

I. Experimental hyperparameters

In this section, we document experimental settings to foster reproducibility. In general, we find that VectorFusion is robust to the choice of hyperparameters. Different settings can be used to control generation style.

I.1. Path initialization

Iconographic Art We initialize our closed Bézier paths with radius 20, random fill color, and opacity uniformly sampled between 0.7 and 1. Paths have 4 segments.

Pixel Art Pixel art is represented with a 32×32 grid of square polygons. The coordinates of square vertices are not optimized. We initialize each square in the grid with a random RGB fill color and an opacity uniformly sampled between 0.7 and 1.

Sketches Paths are open Bézier curves with 5 segments each. In contrast to iconography and pixel art, which have borderless paths, sketch paths have a fixed stroke width of 6 pixels and a fixed black color. Only control point coordinates can be optimized.

I.2. Data Augmentation

We do not use data augmentations for SD + LIVE + SDS. We only use data augmentations for SDS trained from scratch, and CLIP baselines following [4]. For SDS trained from scratch, we apply a perspective and crop augmentation. Our rasterizer renders images at a 600x600 resolution, and with 0.7 probability, we apply a perspective transform with distortion scale 0.5. Then, we apply a random 512x512 crop. All other hyperparameters are default for Kornia [30].

I.3. Optimization

We optimize with a batch size of 1, allowing VectorFusion to run on a single low-end GPU with at least 10 GB of memory. VectorFusion uses the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.9$, $\epsilon = 10^{-6}$. On an NVIDIA RTX 2080ti GPU, VectorFusion (SD + LIVE + SDS) takes 25 minutes per SVG.

For sketches and iconography, the learning rate is linearly warmed from 0.02 to 0.2 over 500 steps, then decayed with a cosine schedule to 0.05 at the end of optimization for control point coordinates. Fill colors use a 20× lower learning rate than control points, and the solid background color has a 200× lower learning rate. A higher learning rate for coordiantes can allow more structural changes.

For pixel art, we use a lower learning rate, warming from 0.00001 to 0.0001 over 1000 iterations. We also add a weighted L2 saturation penalty on the image scaled between [-1, 1], $1/3 * \text{mean}(I_r^2 + I_b^2 + I_g^2)$, with a loss weight of 0.05. Both the lower learning rate and the L2 penalty reduced oversaturation artifacts.

J. Perceptual Quality Metric

We measure aesthetic scores in Table 9 using the LAION aesthetics classifier, trained on frozen CLIP features of 250k images with human quality labels from 1-10 (Schuhmann et al 2022). VectorFusion with our latent SDS loss has the most aesthetic results, comparable to raster samples.

Table 9. VectorFusion from scratch produces the most aesthetic samples, even outperforming Stable Diffusion images.

Method	K	Aesthetic
CLIPDraw (scratch)	–	4.10 ± 0.81
Stable Diff (**raster**)	1	5.39 ± 0.93
+ rejection sampling	4	5.37 ± 0.84
SD init + LIVE	1	4.90 ± 0.91
+ rejection sampling	4	4.93 ± 0.89
VectorFusion (scratch)	–	**5.50** ± **0.79**
+ SD init + LIVE	1	5.45 ± 0.75
+ rejection sampling	4	5.35 ± 0.73