

StarVector: Generating Scalable Vector Graphics Code from Images and Text

Supplementary Material

In the following sections, we provide additional details on the datasets used in this paper, present further experiments, and describe our baselines in detail. We also discuss the StarVector architecture, its training process, and the method for sampling SVG code from the model. Additionally, we provide more insights into SVG-Bench, including the proposed datasets and the different baselines within the evaluation setup. Finally, qualitative results are presented to showcase the strengths and limitations of our foundational model.

8. SVG Datasets in SVG-Bench

Here we describe available SVG datasets in the recent literature. We extend our description of the datasets used for training and evaluating StarVector and other baselines. Earlier SVG datasets proposed in the literature (mainly datasets of emojis and fonts) were not easily accessible due to broken URLs and no direct entry point. Therefore, we provide them as part of SVG-Bench for easy reproducibility. We introduce splits for train, validation, and testing. The train set is used to optimize the parameter weights of the network. The validation is used for tuning sampling hyperparameters, and the test is used for evaluation. Our model can handle up to 8k context tokens. Therefore, our datasets only consider examples with up to 8,192 tokens. See Table 5 for a complete description of the datasets. See Figures 6, 7, 8 for ground truth examples of the test sets of SVG-Bench.

8.1. Datasets with Simplified SVGs.

We create simplified versions of our four main datasets, i.e. emojis, icons, fonts, and SVG-Stack. This is done because DeepSVG [14] requires a simplification of the SVG in its input. The simplification consists of eliminating complex primitives and using only vector *paths*. Also, color and shapes are abstracted only to use simple line strokes.

8.2. Creating the SVG-Fonts Dataset

To construct the SVG-Font dataset, we replicate the procedure described in SVG-VAE [47]¹, which provides a list of public URLs containing open font packages. We download these packages, excluding any with broken URLs. The TTF files are then converted to SFD format, and we further use Inkscape² to convert them into SVG code. Samples from the test set are shown in Figure 7 (bottom-left). These samples contain only path elements and represent a

¹https://github.com/magenta/magenta/tree/main/magenta/models/svg_vae

²<https://inkscape.org/>

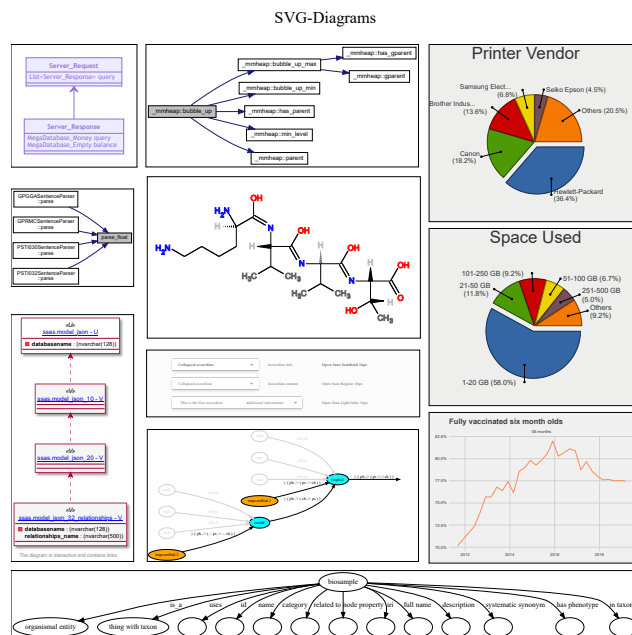


Figure 6. SVG Diagrams examples. These are ground truth SVG examples from the test set. They are presented as SVG, showing the challenge of understanding intricate structures and small texts, with images of variate aspect ratios.

narrow range of images, each consisting of a single character per image, making them an ideal test case for SVG generation [13, 14, 47].

8.3. Creating SVG-Diagrams

We introduce a novel SVG dataset, SVG-Diagrams, which focuses exclusively on diagrams, graphs, workflows, and other designs characterized by discrete elements such as boxes, arrows, and text. To construct this dataset, we filtered all SVGs containing the text element. Table 5 provides detailed statistics about the dataset, and Figure 6 illustrates test samples from SVG-Diagrams.

This new benchmark for diagram generation is highly relevant, as it addresses a use case that cannot be tackled by traditional image-processing models or methods limited to the path primitive. Only approaches capable of leveraging SVG code can fully exploit the use of primitives to generate such structured designs effectively.

8.4. Generating Synthetic Captions on SVG-Stack

To generate textual instructions for vector images, we process SVG-Stack images using visual captioning models.

Table 5. **Summary of datasets.** We offer a summary of statistics about the datasets used in our training and evaluation experiments. This datasets are included in SVG-Bench. The subscript *sim* stands for the simplified version of the dataset, as required by some baselines.

Dataset	Train	Val	Test	Source	Token Length	SVG Primitives	Annotation
SVG-Stack	2,1M	108k	5,7k	TheStack [35]	$1,822 \pm 1,808$	All	Caption
SVG-Stack_{sim}	601k	30,k	1,5k		$2k \pm 918$	Vector path	Caption
SVG-Diagrams	-	-	472		$3,486 \pm 1,918$	All	Caption
SVG-Fonts	1,8M	91,5k	4,8k	Glypazzn [47]	$2,121 \pm 1,868$	Vector path	Font type
SVG-Fonts_{sim}	1,4M	71,7k	3,7k		$1,722 \pm 723$	Vector path	Font type
SVG-Emoji	8,7k	667	668	OpenMoji, NotoEmoji, TweMoji	$2,551 \pm 1,805$	All	Class
SVG-Emoji_{sim}	580	57	96		$2,448 \pm 1,026$	Vector Path	Class
SVG-Icons	80,4k	6,2k	2,4k	DeepSVG [14]	$2,449 \pm 1,543$	Vector path	-
SVG-Icons_{sim}	80,435	2,836	1,277		$2,005 \pm 824$	Vector path	-
SVG-FIGR[‡]	270k	27k	3k	IconShop [94]	$5,342 \pm 2,345$	Vector path	Class, Caption

This approach provides a textual description for each image, enabling us to fine-tune our model to follow textual instructions. For this task, we leverage off-the-shelf AI captioners, specifically BLIP2 [41] and Llava [45].

Through prompt engineering, we guide these models in performing the captioning task with reasonable quality. The prompt we used is shared in Prompt 8.4. After automatically captioning all SVG samples in SVG-Stack, we compute the CLIP Score for the text-image pairs generated by the two models—producing two captions per image. Using a CLIP Score threshold of 30, we filter out text captions that fall below this threshold.

Prompt 1. Utilized with BLIP2 and Llava for SVG Captioning: You are a helpful assistant. Your task is to caption the input images with a concise and clear description that represents what are the contents of the image.

8.5. Data Augmentation for SVG

We introduce several data augmentation operations on SVGs that aim to perform minor modifications to the SVG code and rasterize it to get a new sample while training. We include rotation, color, and curve noise. We evaluate this setting on datasets with fewer samples, namely SVG-Emoji and SVG-Icons, as the other two datasets are large enough to do not overfit. Results are shown in Table 8. Both datasets display improvements using these augmentations. We see a substantial uplift for SVG-Emoji, which has limited training data.

We introduce several augmentation operations to SVGs to apply slight changes that help our model learn to generate more precise results — for instance, being able to capture

exact colors from the image and encode them in hexadecimal code to insert it in the `fill` attribute of the SVG element. Applying rotations or adding noise to the curve’s control points helps the model learn to precisely capture the position of the edges or thickness of the stroke.

We perform random rotations in an angle range. We perform color changes by first parsing the element’s color using the `fill` attribute and adding slight white Gaussian noise to the RGB values. We propose curve noise by injecting a small Perlin [80] noise into the control points in Bézier curves. We also experimented with adding Gaussian noise, which resulted in much less natural results. We apply this noise by uniformly sampling a scalar from the interval between 0.01 and 0.05 and use it to scale the noise.

We apply these augmentations directly on the SVG code, which involves parsing the XML code and accessing the attributes and arguments of the primitives defined. We use the libraries `BeautifulSoup`³ and `SvgPathTools`⁴. Some primitives are simplified using our augmentations.

9. SVG Methods and Baselines

Here, we describe the previous methods and baselines used to compare StarVector’s performance in Image-to-SVG and Text-to-SVG generation tasks. We consider previous deep learning-based methods and image-processing methods. We evaluate the baselines with publicly available code in our proposed setup.

9.1. Image-to-SVG Baselines

We reproduce all previous approaches on our proposed SVG-Bench benchmark, as the available results stem from an unclear version of the fonts, emojis, and icons datasets. For the Image-to-SVG task, we consider several baseline

³<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

⁴<https://github.com/mathandy/svgpathtools>



Figure 7. **Datasets in SVG-Bench.** Ground truth test examples from the test sets of **SVG-Stack**, **SVG-Emoji**, **SVG-Fonts**, and **SVG-Icons**. We show SVG images.

methods across deep learning and image processing approaches.

In the deep learning methods category, we start with

DeepSVG[14], Im2Vec[67], and LIVE[50], using the official implementations with the hyperparameters proposed by the authors, and applying their pre- and post-processing



Figure 8. **Simplified Datasets in SVG Bench.** Ground truth test examples from the simplified test sets of **SVG-Stack**, **SVG-Emoji**, **SVG-Fonts**, and **SVG-Icons**. We show SVG images.

code as required. Additionally, we incorporate the recent GPT-4 Vision[56], which is capable of processing images as input and generating corresponding SVG code as output.

For the image processing-based methods, which do not

rely on data-driven learning, we consider VTracer[87], Autotracer[51], and Potrace [59], running them on the test sets of SVG-Bench.

Table 6. **Summary of SVG Methods.** We compare SVG generation methods based on *Image Processing*, *Latent Variable*, *Differentiable Rendering*, and *Multimodal LLM*, evaluating their performance in *SVG Generalization* and *SVG Generation Tasks*. The *SVG Generalization* column shows whether a model generates diverse SVG types (e.g., icons, logos, complex shapes) with ✓ or specializes in a subtype (e.g., emojis, fonts) with ✗. The **SVG Primitive Coverage** column indicates access to all SVG primitives. The table also evaluates **Image Vectorization**, **Text to SVG**, and **Diagram Generation**, using ✓ for support and ✗ for limitations. *DeepSVG requires modifications for image input.

Method Type	Model	Input	Train Supervision	SVG Coverage and Generalization		SVG Generation Tasks		
				SVG Generalization	SVG Primitive Coverage	Image Vectorization	Text to SVG	Diagram Generation
Image Processing	Vtracer	Image	Image	✓	✗	✓	✗	✗
	Autotrace	Image	Image	✓	✗	✓	✗	✗
	Potrace	Image	Image	✓	✗	✓	✗	✗
Latent Variable	Im2Vec	Image	Image	✗	✗	✓	✗	✗
	DeepSVG	SVG	Vector	✗	✗	✓*	✗	✗
	SVGFormer	SVG	Vector	✗	✗	✓	✗	✗
Diff. Rendering	DiffVG	Image	Image	✓	✗	✓	✗	✗
	LIVE	Image	Image	✓	✗	✓	✗	✗
	SAMVG	Image, Text	Image	✓	✗	✓	✗	✗
	SVGDreamer	Image, Text	Image	✓	✗	✗	✓	✗
Multimodal LLM	GPT-4 V	Image, Text	SVG	✓	✓	✗	✓	✓
	CodeLlama	Image	SVG	✓	✓	✗	✓	✓
	IconShop	Text	SVG	✗	✗	✗	✓	✗
	StarVector	Image, Text	SVG	✓	✓	✓	✓	✓

Autotrace⁵ [51] is a tool designed for converting images to vector graphics, similar to Potrace. It supports various input formats and can output to several vector formats. Autotrace’s key feature is its ability to transform pixelated images into smooth, scalable vectors, making it ideal for upgrading images for various applications without losing detail or clarity. Our experiments leverage the Python bindings⁶ implementation of AutoTrace.

Potrace⁷ [59] is a utility designed to convert images into refined, scalable vector graphics. It accepts input in various bitmap formats and outputs to a selection of vector formats. This functionality is particularly valuable for generating SVG of scanned imagery, such as logos and handwritten documents. We employ a Python library⁸, which acts as a wrapper around the original C implementation of Potrace.

VTracer⁹ [87] is an image processing algorithm to convert images to SVGs. This 3-step pipeline algorithm relies on the hierarchical clustering of images, which are traced into vectors. First, pixels are converted into paths and then simplified into polygons. In the last step, polygons are smoothed and approximated with a Bezier curve fitter. We use the Python library¹⁰ for experiments, a wrapper over the Rust implementation. Similar to Im2Vec, we scale down

all the images to 128X128 resolution. We use all the default values for the image processing engine, which generates a multi-colored SVG.

Im2Vec [67] uses an end-to-end VAE, trained using only image supervision to produce vector graphics. The input rasterized image is encoded to a ‘global’ latent vector and passed to an RNN to produce latent code for each path. The path decoder decodes these codes into Bezier paths to generate the output SVG. We used the publicly available code¹¹ to report the results.

We scaled all the images to 128×128 resolution to be compatible with the Im2Vec model. We used a learning rate of 5×10^{-4} and a batch size of 8. We implemented a custom post-processing operation for converting the vector parameters obtained during Im2Vec inference to obtain compilable SVG code.

LIVE, (Layer-wise Image Vectorization) [50] is a method for progressively generating SVGs that closely fit a given raster image by recursively adding and optimizing closed vector *paths*. Using a differentiable renderer (based on DiffVG [43]), LIVE enables direct optimization of paths under raster image supervision while controlling shape complexity by adjusting the number of path segments. It introduces component-wise path initialization, identifying key visual components to ensure efficient topology extraction and minimize redundant shapes. LIVE achieves high-quality reconstructions with fewer paths, reducing SVG file size com-

⁵<https://potrace.sourceforge.net/>

⁶<https://github.com/lemonyte/pyautotrace>

⁷<https://potrace.sourceforge.net/>

⁸<https://github.com/tatarize/potrace>

⁹<https://github.com/visioncortex/vtracer>

¹⁰https://github.com/etjones/vtracer_py

¹¹<https://github.com/preddy5/Im2Vec>

pared to other approaches. Nevertheless, its test time optimization approach makes it time-consuming during generation. We utilized their official open-source implementation¹² with the proposed hyperparameters. This method requires to define a constant number of *paths*; the more paths defined, the more accurate. We have performed an ablation on the number of paths (see Table 3) and found that *paths*=32 is an optimal value that brings good visual results. However, it takes more than 10 minutes to generate a single SVG, which makes it slow for a professional use case.

DiffVG [43] is a landmark in vector graphics research, pioneering deep learning-based methods with the first differentiable vector graphics rasterization pipeline. By leveraging a combination of anti-aliasing techniques and gradient-based optimization, DiffVG ensures differentiability. Unlike methods relying on non-differentiable curve-to-mesh conversions, DiffVG employs a forward-backward rasterization process, where the forward pass generates anti-aliased images and the backward pass computes gradients with respect to vector graphic parameters. Using the official implementation¹³ and proposed hyperparameters, we ablate the number of *paths*, finding *paths*=60 to be optimal. DiffVG balances versatility and performance, achieving approximately 30 seconds per generation while excelling in differentiable rendering tasks.

DeepSVG [14] was introduced as a hierarchical path-based VAE encoder-decoder transformer architecture. Here, input paths are encoded separately using a path encoder and aggregated using a second encoder to produce a latent vector. The decoder uses this latent vector to output the path representations, which provide actual draw commands and arguments. We used the open-source code¹⁴ to reproduce the results on different datasets. However, since the DeepSVG framework only allows simplified SVGs, we report results on the ‘simplified’ test sets (see Table 1).

This model can only handle simplified SVGs composed of simple line strokes and splines (see examples in Figure 9). Further, it can only process SVGs with eight groups (i.e., groups of shapes or parent nodes) and vector paths of at most 30 commands. To reproduce the DeepSVG baseline, we use the original hyperparameters, including a learning rate of $1e-3$ and a number of epochs of 50. We use a batch size of 200, except for the smaller emoji dataset, where we experiment with a batch size of 50.

GPT-4 Vision. We use GPT-4V [56] by inserting an image and zero-shot prompting to generate SVG code. Here, we show how one can use prompt engineering [10, 11, 56] to condition the model to generate executable SVG code representing the given image. Prompt 9.2 was used for this endeavor. We use the OpenAI library¹⁵.

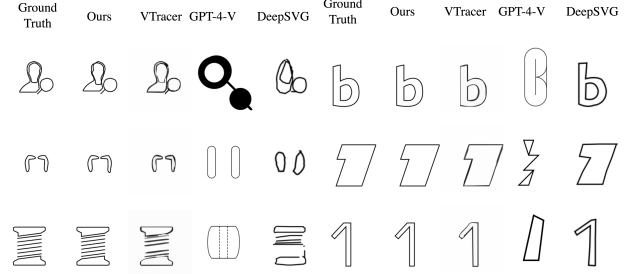


Figure 9. **Image-to-SVG** results on simplified SVG-Icons and SVG-Fonts test set.

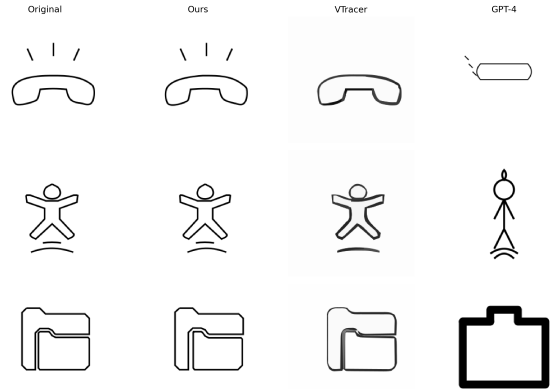


Figure 10. **Image-to-SVG** results on SVG-Icons test set.

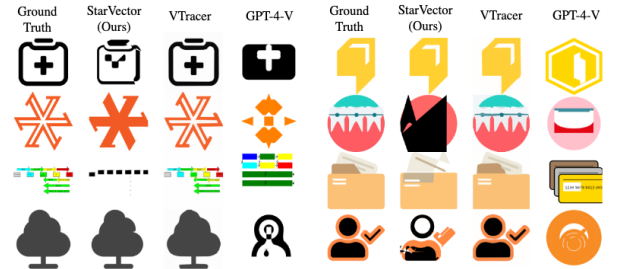


Figure 11. **Image-to-SVG** results on SVG-Stack test set. We show cherry-picked failure examples of StarVector.

¹²<https://github.com/ma-xu/LIVE>

¹³<https://github.com/BachiLi/diffvg>

¹⁴<https://github.com/alexandre01/deepsvg>

¹⁵<https://platform.openai.com/docs/libraries>

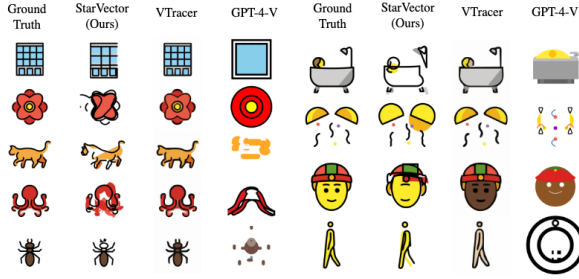


Figure 12. **Image-to-SVG** results on SVG-Emoji test set. We show cherry-picked failure examples of StarVector.

9.2. Text-to-SVG Generation Baselines

For the Text-conditioned SVG generation task, we select baselines based on works that contain reproducible methodologies in public datasets or public code repositories. We reproduce baseline models from their official repositories, respecting the proposed hyperparameters.

CodeLlama, [82, 83] has shown great success in general coding benchmarks. To the best of our knowledge, CodeLlama has seen SVGs during training. Hence, it is reasonable to consider it a strong baseline for text-conditioned SVG generation. We use Anyscale endpoints¹⁶ to generate CodeLlama results.

GPT 4, is a closed source LLM that shows state-of-the-art results in many NLP scenarios [11, 55, 71]. We evaluate GPT-4’s 0-shot ability in generating SVGs when prompted with text inputs. We use OpenAI API¹⁷ to generate results for GPT-4 in the 0-shot setting. Prompt 9.2 was used for the Text-to-SVG task.

Prompt 2. Used on GPT4-V VLM for Image-to-SVG Translation: You are a helpful assistant. Your task is to help researchers write SVG code to reconstruct the provided image as accurately as possible. You should also provide a caption for the image. You are dedicated to solving the task of Image-to-SVG conversion for a robust system. Therefore, you must always respond with the best SVG code you can create. Feel free to use multiple paths to generate a compliant SVG code within a maximum of 8000 tokens. You should present the SVG code that best reconstructs the input image enclosed in triple quotes.

¹⁶<https://app.endpoints.anyscale.com/>

¹⁷<https://platform.openai.com/docs/guides/gpt>

Prompt 3. Used on GPT4 and CodeLlama for Text-to-SVG Generation: You are a helpful assistant assisting researchers in generating SVG code from textual descriptions. You will be provided with details to guide your SVG creation. Your task is to write SVG code that accurately represents the given textual information to the fullest extent possible. You are committed to solving the task of SVG generation for a robust system, so always strive to produce the best SVG code you can. Feel free to use multiple paths and any necessary shapes, colors, or lines to generate compilable SVG code within a maximum of 9000 tokens. The goal is to ensure the resulting SVG, when rasterized, best represents the described content. Respond only with the SVG code, enclosed in triple quotes, that directly corresponds to the provided textual description. Avoid adding any explanation or commentary.

IconShop IconShop [94] uses a transformer-based architecture to encode path commands and learn to model SVG path sequences autoregressively. It has shown excellent results in simplified icon scenarios and provides a good solution to Text-to-SVG generation by extending the FIGR-SVG dataset with captions. We have access to their dataset and original splits and have trained our model on that data using a pre-trained checkpoint (trained on SVG-Stack). We have extracted the results from IconShop and included them here to compare our method.

10. Additional Experiments and Results

10.0.1. Image-to-SVG Results

We show additional Image-to-SVG results from StarVector. Figures [13 - 18] show substantial qualitative samples generated by StarVector on all the proposed datasets. All results are computed in the test sets. We can observe the weaknesses and strengths of our model. Simplified datasets (Figures 16) are near-perfectly converted to SVG. In the case of icons, in Figure 18, sometimes the model runs out of SVG code tokens, and the image is incomplete. Results on SVG-Emoji 17 show impressive performance in estimating the shape’s color and semantics. However, it lacks fine-grained and accurate positioning of objects, i.e., in some examples, the model loses track of the coherent position and form of shapes. These problems result from insufficient emoji samples, i.e., less than 10,000 training examples. This problem can be alleviated by scaling up the current model in the number of parameters (currently 1.4 billion), training data for pre-training, and computing resources.

Input Image	StarVector-8B	LIVE	VTracer	PoTrace	AutoTrace
	DinoScore: 0.9871 MSE: 0.0075	DinoScore: 0.8325 MSE: 0.0019	DinoScore: 0.8750 MSE: 0.0026	DinoScore: 0.7806 MSE: 0.0062	DinoScore: 0.8243 MSE: 0.0034
	DinoScore: 0.9941 MSE: 0.0052	DinoScore: 0.7835 MSE: 0.0026	DinoScore: 0.8687 MSE: 0.0033	DinoScore: 0.9006 MSE: 0.0100	DinoScore: 0.7814 MSE: 0.0040
	DinoScore: 0.9988 MSE: 0.0071	DinoScore: 0.8420 MSE: 0.0069	DinoScore: 0.9195 MSE: 0.0119	DinoScore: 0.9014 MSE: 0.0446	DinoScore: 0.9060 MSE: 0.0159
	DinoScore: 0.9960 MSE: 0.0150	DinoScore: 0.8670 MSE: 0.0060	DinoScore: 0.8624 MSE: 0.0066	DinoScore: 0.8014 MSE: 0.0094	DinoScore: 0.7757 MSE: 0.0082
	DinoScore: 0.9887 MSE: 0.0532	DinoScore: 0.8513 MSE: 0.0159	DinoScore: 0.7943 MSE: 0.0196	DinoScore: 0.8560 MSE: 0.0842	DinoScore: 0.7922 MSE: 0.0309
	DinoScore: 0.9974 MSE: 0.0210	DinoScore: 0.7979 MSE: 0.0083	DinoScore: 0.8366 MSE: 0.0104	DinoScore: 0.8128 MSE: 0.0147	DinoScore: 0.8001 MSE: 0.0116
	DinoScore: 0.9891 MSE: 0.0093	DinoScore: 0.8584 MSE: 0.0035	DinoScore: 0.8542 MSE: 0.0049	DinoScore: 0.8462 MSE: 0.0056	DinoScore: 0.8987 MSE: 0.0058
	DinoScore: 0.9690 MSE: 0.0316	DinoScore: 0.8433 MSE: 0.0166	DinoScore: 0.8202 MSE: 0.0164	DinoScore: 0.7061 MSE: 0.0243	DinoScore: 0.8119 MSE: 0.0209

Figure 14. We compare the results from StarVector-8B with those from the most powerful baselines. Notably, StarVector is *the only* method capable of producing acceptable results that preserve both structural integrity and textual content by utilizing a variety of SVG primitives. In contrast, other methods tend to generate blobs and curves that merely attempt to fit the structure and color of the original image. We present two metric scores for each sample: DinoScore and MSE. MSE consistently yields higher scores for other methods, as they focus on fitting vectors to the image as accurately as possible. While StarVector may not achieve perfect reconstruction, its results are preferred for their semantic fidelity. This highlights the limitations of MSE and the importance of DinoScore in capturing these aspects.

Ground Truth Generated Ground Truth Generated Ground Truth Generated Ground Truth Generated Ground Truth Generated Ground Truth Generated

5 5 a a i i b b A A I I
 B B | | 4 4 U U N N Z Z
 ì ì x x p p / / s s o o
 v v r r n n Y Y ‡ ‡ A A
 k k w w W W o o F F Q Q
 8 8 1 1 T T \ \ u u W W
 /A /A x x L L t t N N j j
 8 8 E E o o U U s s i i
 Y Y n n A A z z U U T T
 w w T T k k o o J J F F

Figure 15. **Image-to-SVG results on SVG-Fonts test set.** Results are remarkably good, obtaining perfect font reconstructions. Intricate details are preserved. This is because the dataset is very large, above 1M samples. This shows that if having access to a large dataset, StarVector can learn high-quality SVG generation.

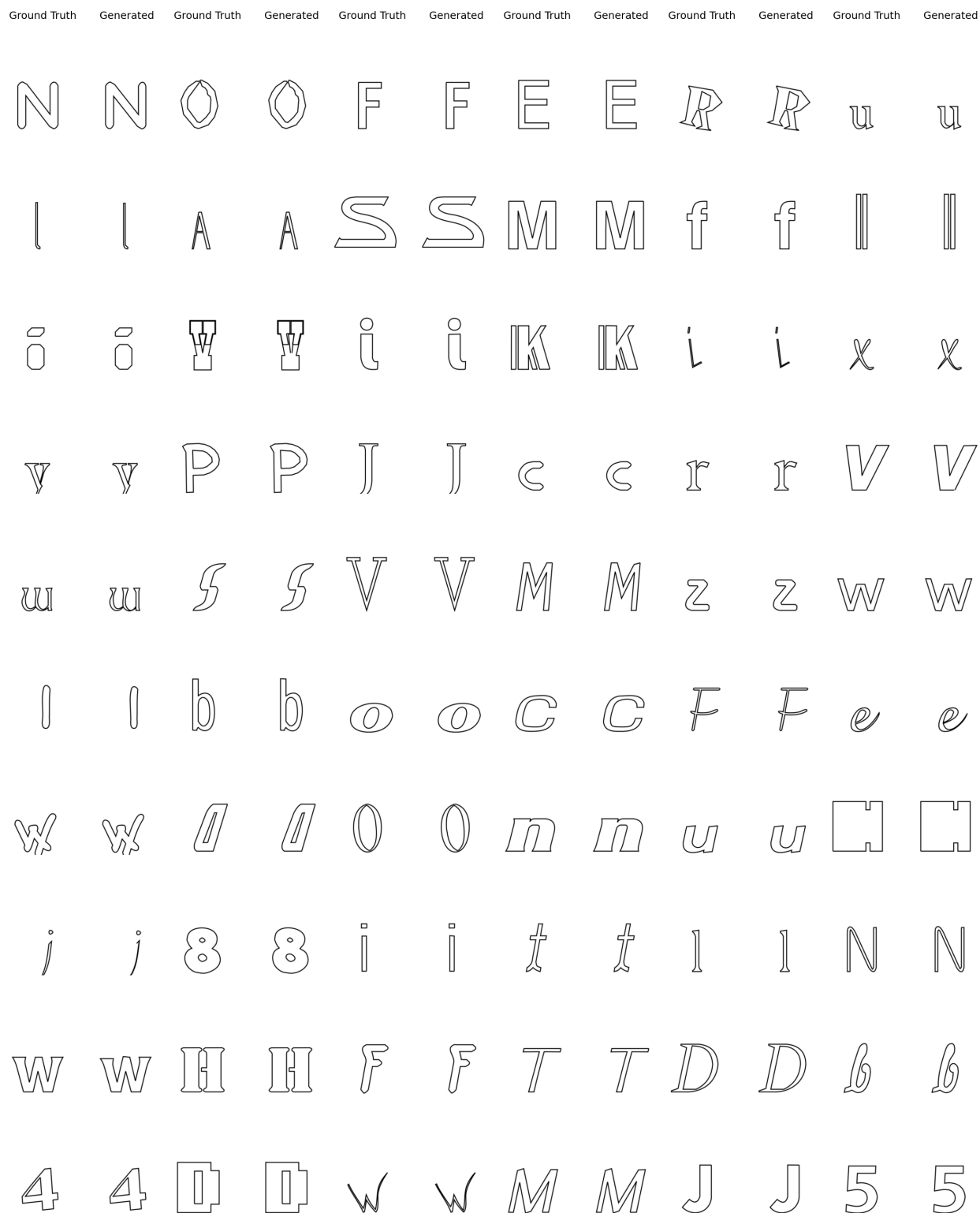


Figure 16. Image-to-SVG results on SVG-Fonts simplified test set.



Figure 17. **Image-to-SVG results on SVG-Emoji test set.** Results are mostly wrong in this benchmark, due to the small training dataset of approximately 8k examples.

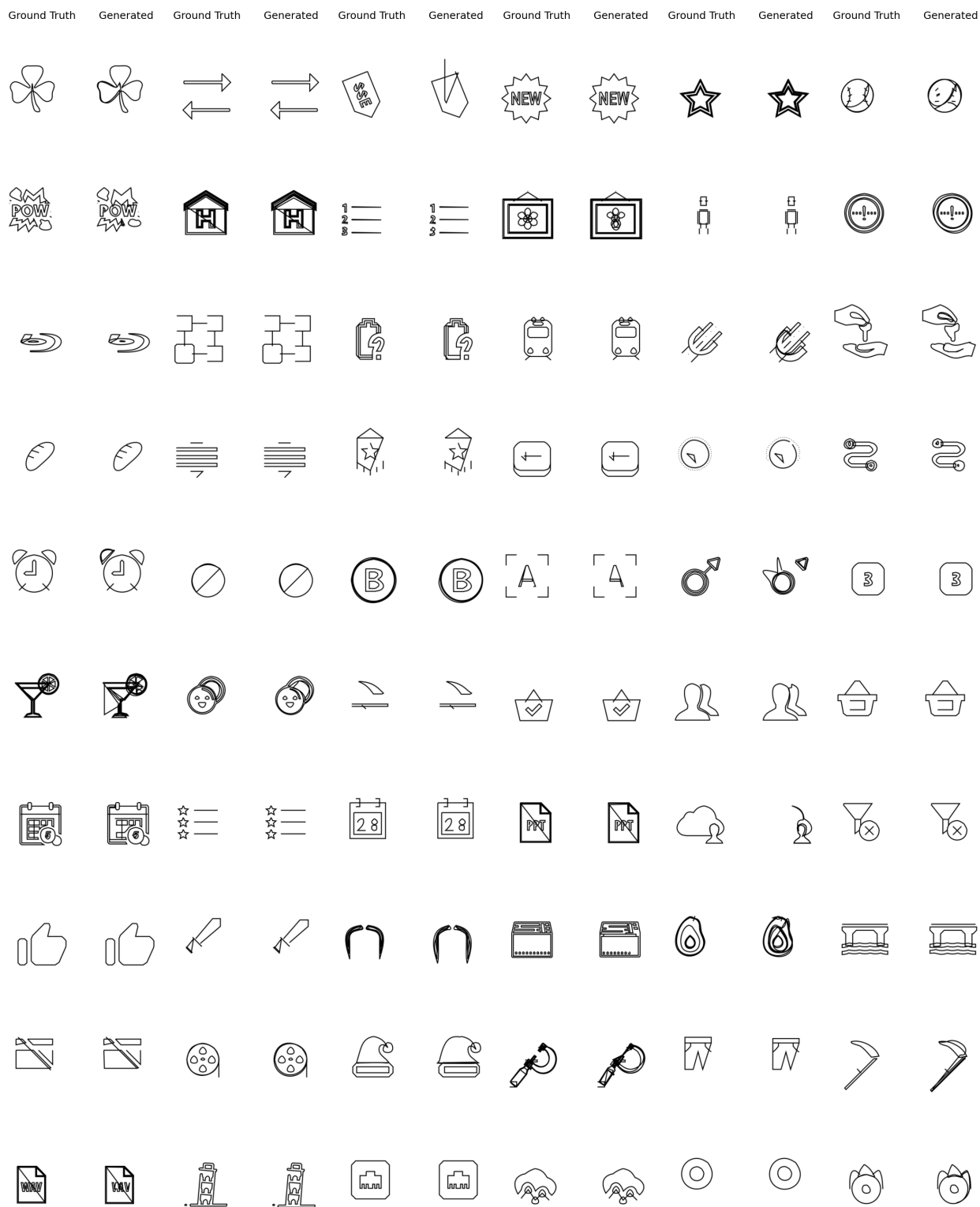


Figure 18. Image-to-SVG results on SVG-Icons test set.

Comparing parameter count of models . The number of learnable parameters in deep learning based models often correlates with performance. Pre-LLM models like Im2Vec and DeepSVG use significantly fewer parameters (up to 5M) compared to StarVector and GPT-based models, which operate in the billions. While pre-LLM models can produce accurate results, they lack the generalization ability of LLM-based approaches. Comparing StarVector with GPT models reveals that high-fidelity SVG generation is achievable with just 1B parameters, whereas GPT-4V lacks specific training for this task. Future models will likely incorporate SVG data in training, but current results already demonstrate that LLM-based approaches offer superior generalization and scalability for SVG generation, at the cost of utilizing more parameters.

Context Length Limitation. The model’s architecture imposes a clear limitation on context length, which significantly impacts training and testing data pipelines, as well as the skills the model can learn. Our experiments show that the model scales effectively with increasing context lengths, from 8k to 16k, indicating that this limitation must be addressed with techniques for handling longer contexts—an area LLMs are expected to improve. For fair comparisons, we restricted our benchmark tests to a context length of 8k and evaluated all baselines within this setting. However, the benchmarks also provide versions with longer context lengths to assess future models, as increased length generally correlates with more complex SVGs. We did not observe substantial differences in scores between the 4k and 10k token settings, primarily because the data in our benchmark can typically be represented using an average of 3k tokens.

Limitations on Complex SVG Structures. StarVector encounters challenges with complex SVG structures, intricate shapes, and detailed illustrations primarily due to limitations in its architecture. Currently, the model’s image encoder handles images by simply padding and resizing them to fixed dimensions of 224 or 384 pixels. This approach may not adequately capture the nuances of complex diagrams. A potential improvement would be to implement a dynamic image processing system akin to those found in newer Vision-Language Models (VLMs), which could enhance the model’s ability to interpret and generate intricate SVGs more effectively. Additionally, improving data cleaning processes is crucial, as the model sometimes produces hallucinated information due to noise in the input data, such as URLs or base64-encoded images. Addressing these issues through architectural enhancements and more robust data preprocessing could significantly improve StarVector’s performance on complex SVG tasks.

Generalization to Non-Standard SVGs. StarVector’s ability to generalize to non-standard SVGs—those not represented in its main training distribution—poses a significant challenge. While the model performs well on common styles and primitives encountered during training, it struggles with more unique or unconventional SVGs. This is primarily due to the model’s training data, which tends to focus on widely used shapes and designs. As a result, StarVector may exhibit a bias towards these common styles, leading to suboptimal performance when faced with SVGs that feature unusual structures or less frequent elements.

To assess StarVector’s generalization capabilities, we evaluated its performance on various datasets that include non-standard SVGs. The results indicate that while the model can produce reasonable outputs for some non-standard examples, it often falls short in accurately capturing the intricacies of less familiar styles. This limitation suggests that the model’s training set lacks sufficient diversity to encompass the full range of potential SVG designs.

To address these concerns, future work should focus on expanding the training dataset to include a wider variety of SVG styles and structures. Incorporating data from niche applications and artistic domains could enhance the model’s ability to generate SVGs across a broader spectrum of design elements. Additionally, techniques such as domain adaptation and transfer learning could be explored to improve generalization to non-standard SVGs, allowing StarVector to adapt more effectively to unfamiliar inputs.

10.1. Ablation Studies

We performed ablations on the image encoder type, the data augmentation pipeline, inference techniques, and generation parameters. Most of our ablations were performed on the StarVector-1B model for faster iteration, and we empirically find they work well on the larger StarVector-8B.

Image Encoder Ablation. The choice of image encoder for the problem of Image-to-SVG is highly impactful, as it determines how well visual information from raster images can be preserved in a representation suitable for precise reconstruction in the SVG space. We ablated the visual encoders by replacing them with VQGAN [23], ConvNext [46], and CLIP ViT-B/32 [62], in our StarVector-1B proposed architecture. This setup evaluates three commonly used approaches in visual representation learning [23, 62, 73]. In our experiments, CLIP consistently outperformed across all metrics for various datasets (see Table 9). Figures [19–22] further illustrate how VQGAN and ConvNext tend to lose local details during generation, even while maintaining semantic relevance.

Sampling technique	SVG-Fonts			SVG-Emojis			SVG-Icons			SVG-Stack		
	LPIPS ↓	SSIM ↑	MSE ↓	LPIPS	SSIM	MSE	LPIPS	SSIM	MSE	LPIPS	SSIM	MSE
Greedy	0.019	0.969	0.013	0.251	0.731	0.071	0.059	0.912	0.028	0.157	0.797	0.067
+ Beam Search (B=5)	0.018	0.970	0.012	0.250	0.732	0.070	0.058	0.913	0.027	0.156	0.798	0.066
Nucleus Sampling (T=0.5)	0.013	0.976	0.008	0.202	0.778	0.051	0.043	0.923	0.022	0.153	0.785	0.072
Nucleus Sampling (T=1.0)	0.015	0.975	0.009	0.244	0.742	0.067	0.053	0.917	0.025	0.161	0.786	0.069
+ Beam-Search (B=5)	0.034	0.948	0.027	0.244	0.742	0.068	0.065	0.913	0.027	0.195	0.766	0.089
+ Beam-Search (B=10)	0.040	0.943	0.031	0.251	0.742	0.072	0.071	0.910	0.028	0.175	0.762	0.079

Table 7. **Ablation study on sampling strategies.** We experimented using greedy decoding and added a beam search with B=5. We test nucleus sampling [31] using top p=0.9, with temperatures T=0.5 and T=1.0. The two final rows describe beam search with nucleus sampling at T=1.0. See huggingface.com/blog/how-to-generate for reference on these sampling techniques.

Pre-training on SVG-Stack. Pre-training on the SVG-Stack is highly beneficial for the downstream datasets with small data. Table 8 shows the uplift on all the metrics for different datasets. Qualitatively, we can also see that pre-training helps the model to identify the nuanced details from the images. For the case of SVG-Emoji, pre-training is a vital requirement, as it overfits without it due to limited data. Figure 17 shows that the model relies on colors and shapes to generate the SVG.

Ablation on Generation Hyperparameters. We explore the impact of different generation hyperparameters on the StarVector-1B model. After an initial exploration to empirically determine the most relevant hyperparameters, we focus our ablation on these. We find that temperature and the number of beams in beam search significantly affect performance. The model is evaluated across various configurations (see Table 7 and Figure 23). Our results show that a beam search size of 5 achieves the best outcomes, albeit with increased memory usage and runtime. Similarly, nucleus sampling with a top-p of 0.9 and a temperature of 0.5 delivers the best overall performance.

10.2. Text-to-SVG Results

Figures [25 - 28] show additional qualitative results of StarVector when performing the task of text-conditioned SVG generation, performed on SVG-Stack and FIGR-SVG test sets. Our samples show reasonable effectiveness at this task, consistently grasping features like colors, shapes, and semantic concepts. However, sometimes some details required in the prompt are lost, e.g., an exact number of circles, shapes inside other shapes, or the direction of arrows. In some cases, some vector graphics shapes lose coherence, which we attribute to our model’s current scale in terms of model parameters and context length. We suspect that these mistakes are due to the limited quality of the textual descriptions, sometimes lacking precision and grounding on the

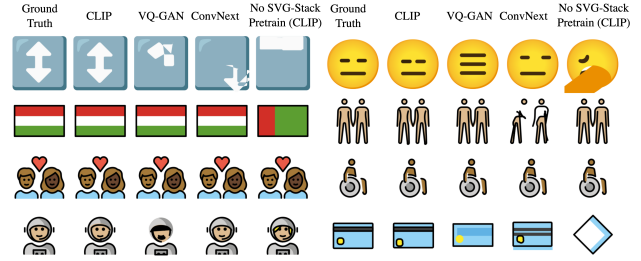


Figure 19. **Ablation of Image Encoders** Image vectorization results using different visual encoders on SVG-emoji test set. CLIP is the image encoder that delivers the best results, whereas VQ-GAN and ConvNet often miss relevant semantics of the image. No SVG-Stack Pretrain (CLIP) refers to an ablation where we use CLIP out of the box, without unfreezing its weights.

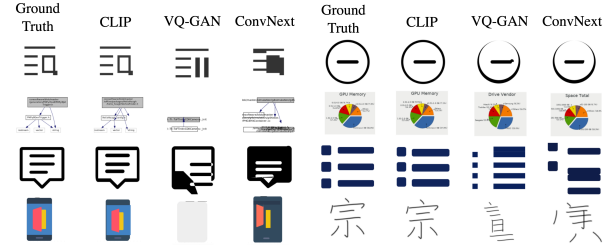


Figure 20. **Ablation of Image Encoders** Image vectorization results using different visual encoders on SVG-Stack test set. CLIP offers the best results. VQ-GAN and ConvNet often miss relevant semantics of the image.

SVG images. See Figure 24 for successful cases of Text-to-SVG generation on SVG-Stack. Figure 29 highlights some failure modes of StarVector-8B. These figures illustrate the impact of different generation temperatures. We rank the outputs generated at different temperatures based on their CLIP Score in relation to the text instruction.

Nevertheless, the StarVector approach of using LLMs for SVG code generation is the only method among baselines that allows us to create diverse vector graphics unre-

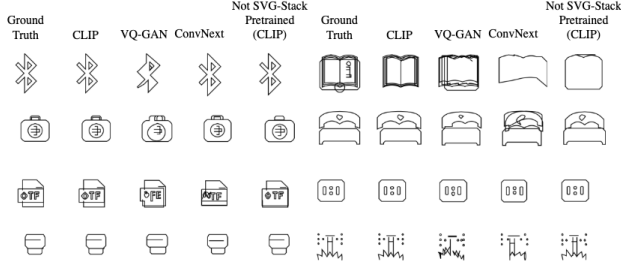


Figure 21. **Ablation of Image Encoders** Image vectorization results using different visual encoders on SVG-Icons test set. CLIP brings the best visual results. VQ-GAN and ConvNext are not able to capture correctly the details for correct vectorization. No SVG-Stack Pretrain (CLIP) refers to an ablation where we use CLIP out of the box, without unfreezing its weights. Notably, better results are obtained when training the CLIP image encoder on SVG-Stack.

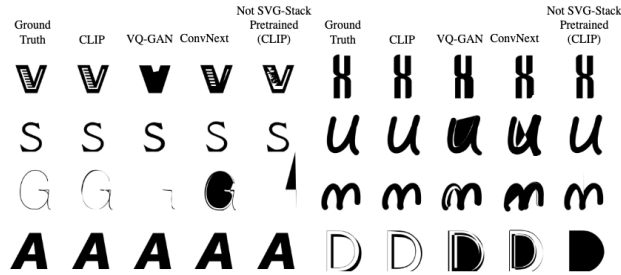


Figure 22. **Ablation of Image Encoders** Image vectorization results using different visual encoders on SVG-Fonts test set. As in the other datasets tested, CLIP brings the best visual results, as others are not able to provide perfect vector reconstruction when intricate details are present. No SVG-Stack Pretrain (CLIP) refers to an ablation where we use CLIP out of the box, without unfreezing its weights. Notably, better results are obtained when training the CLIP image encoder on SVG-Stack.

strainedly, paving the way for more challenging and intricate designs.

10.3. Results on SVG-Diagrams

Figure 2 presents the results of StarVector-8B, along with comparisons to LIVE, VTracer, Potrace, and AutoTrace. StarVector-8B is the only approach that produces plausible results, as it effectively leverages appropriate SVG primitives. DinoScore aligns well with this visual assessment, accurately reflecting the quality of StarVector-8B’s outputs.

In contrast, MSE consistently favors other baselines despite their limitations. This is because MSE prioritizes exact pixel matching, favoring models designed to fit curves and colors to the input image. However, these baselines fail to preserve the semantics of the original diagrams, resulting in outputs where the meaning and structure are completely lost.

Table 8. **Results of SVG Data Augmentation.** We ablate both our data augmentation pipeline and the use of a pretraining stage with SVG-Stack. These experiments are conducted on smaller datasets that are more susceptible to overfitting, using the StarVector-1B model. Vanilla refers to the StarVector model trained directly on the given dataset without SVG-Stack pretraining. Next, we introduce our data augmentation pipeline. Finally, we initialize training from an SVG-Stack pretrained checkpoint and fine-tune on the given dataset. The “+” symbol indicates that the methods from the previous rows are also included.

Method	SVG-Emojis			SVG-Icons		
	LPIPS ↓	SSIM ↑	MSE ↓	LPIPS ↓	SSIM ↑	MSE ↓
StarVector (vanilla)	0.355	0.683	0.108	0.104	0.845	0.047
+ Data Augmentation	<u>0.329</u>	<u>0.706</u>	<u>0.097</u>	<u>0.057</u>	<u>0.905</u>	<u>0.029</u>
+ SVG-Stack Pretrain	0.225	0.748	0.061	0.057	0.894	0.031

Table 9. **Ablation of Image Encoders.** We ablate different image encoders with StarVector-1B, namely CLIP ViT-B/32 [62], VQ-GAN [23], and ConvNext [46]. We experiment with training experiments on SVG-Fonts and SVG-Emojis datasets. CLIP gives the best results on all reconstruction metrics.

Encoder	SVG-Fonts			SVG-Emojis		
	LPIPS ↓	SSIM ↑	MSE ↓	LPIPS ↓	SSIM ↑	MSE ↓
CLIP	0.026	0.955	0.021	0.202	0.778	0.051
VQGAN	0.092	0.854	0.072	0.345	0.688	0.099
ConvNext	0.085	0.854	0.073	0.311	0.708	0.088

10.4. Analysis of SVG Primitives

This section examines how StarVector leverages SVG primitives to produce more compact and semantically accurate SVGs. In contrast to prior models constrained to using only *path* primitives, StarVector effectively utilizes the entire range of SVG primitives, including parametrically defined shapes, gradients, and text elements.

This enhanced capability stems from its ability to operate directly within the SVG code space, facilitated by its multimodal, transformer-based architecture [2, 45], which integrates visual and textual inputs. StarVector generates SVG code that closely resembles the input raster image while maintaining semantic awareness, enabling the use of symmetry, parametric shapes, and text. Prior methods, limited to first-order *path* primitives, lack this semantic understanding, resulting in less compact and less expressive SVG representations.

Qualitative Analysis of SVG Primitives Table 10 presents tests conducted on StarVector-8B and VTracer using simple designs composed of basic shapes such as circles, rectangles, and triangles, with variations in color, transparency, and levels of overlap. StarVector-8B demonstrates the ability to precisely identify the primitives that

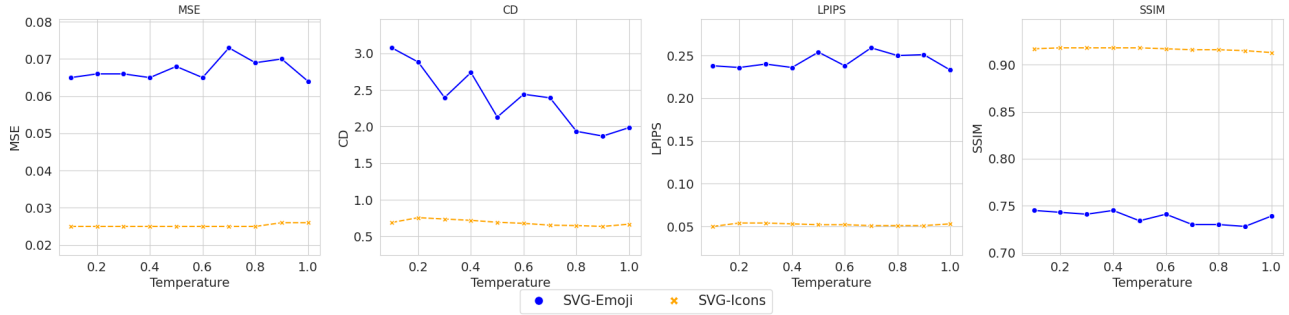


Figure 23. **Ablation study on sampling temperature.** We tested the performance impact of StarVector-1B when changing the sampling temperature. Results are computed for SVG-Emoji and SVG-Icons validation sets.

make up each design, producing visually accurate results while maintaining compact, concise, and interpretable SVG representations.

SVG Tag Distribution Here, we show how StarVector can generate complex SVGs using the full syntax of the SVG language, in contrast with most of the literature methods, which are restricted to using only the *path* command. Figure 30 displays the distribution of SVG tags in the SVGs generated by StarVector along with the distribution in the SVG-Stack dataset, showcasing the strength of our method in using SVG tags and syntax in a similar way to the original in-the-wild dataset. We have computed the exact statistics on previous methods and found that they cannot come close to StarVector in this metric, as they are limited to using *paths* and basic primitives. The effective usage of the large array of SVG tags and syntax makes our method the first SVG model to support these complexities.

10.5. Human Evaluation

We conducted a human evaluation to compare the outputs of StarVector-8B, our best model, with those of the most powerful baselines. Participants were selected from diverse backgrounds and carefully screened for conflicts of interest, with none of the key authors involved. The evaluation was performed through a web interface (shown in Figure 31) that provided anonymized outputs and randomized sample presentations to prevent pattern recognition or bias.

The results, presented in Figure 5, demonstrate a strong preference for StarVector-8B across all settings, especially in SVG-Diagrams tasks. This highlights a disconnect between pixel-based metrics (e.g., MSE, SSIM) and human visual perception of SVGs. While baseline models often prioritize pixel-perfect reconstruction, human evaluators preferred StarVector’s sharp, well-defined shapes and its effective use of primitives (Figure 30).

Spearman correlation analyses between model metrics and human evaluations further emphasize this gap. MSE shows weak correlations (0.0596 and -0.1002), indicating

its inadequacy as a predictor of human preferences. In contrast, DinoScore exhibits significantly stronger correlations, with values of -0.6193 and 0.6214. Moreover, a robust correlation of 0.7577 between differences in DinoScore and human evaluation scores highlights DinoScore as a more reliable metric for assessing SVG quality in alignment with human judgment.

10.6. Comparing StarVector with Baselines

Here, we discuss the results of each baseline individually and compare them to our proposed approach.

1. **DeepSVG** [14] is an elegant approach to learning a latent variable model for SVG. It proves effective at learning the task for the *simplified* datasets (Figure 9). It can accurately represent corners and edges. However, it only works in simplified datasets. This limitation restricts it from being a suitable solution in real applications.
2. **Im2Vec** [67] proposes a training procedure that does not require having SVG ground truth. It uses only pixel-based reconstruction loss with the input image, finding the optimal SVG parameters using a differentiable rasterizer like DiffVG [43]. This framework is appealing, as it aims to be used in images without SVG supervision. However, it requires hundreds of epochs with a reduced dataset to overfit the model to those examples, only working on modeling *training* examples, as seen in [67]. Im2Vec results on the datasets presented SVG-Bench are quite poor as it has bad generalization. Therefore, qualitative samples are not presented.
3. **GPT-4 Vision** [56] excels at capturing the semantics of images and generating captions that accurately describe them. The SVG generated from this description































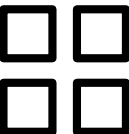





























Text Instruction	Ground Truth	Temp = 0.0	Temp = 0.125	Temp = 0.25	Temp = 0.375	Temp = 0.5
Generate an SVG of x icon in black circle on white background						
Generate an SVG of a green check mark in a flower						
Design an SVG of pencil icon vector						
Draw an SVG picture of a black and white icon of a paper document						
Design an SVG of a colorful square logo with a blue background.						
Draw an SVG image of four square black and white icons.						
Generate an SVG of a folder with a fan on it.						
Draw an SVG image of a blue and white train icon.						
Create an SVG of a large orange bell, positioned in the center of the image.						
Create an SVG for a logo of a company with a gold triangle						

Figure 24. **Text-to-SVG Results.** We show successful Text-to-SVG results using StarVector-8B. We sample 5 different temperatures as an ablation, showing the sensitivity of this parameter during generation. Results are presented in SVG (not raster images)















Caption	Svg Gen.	Caption	Svg Gen.	Caption	Svg Gen.	Caption	Svg Gen.	Caption	Svg Gen.	Caption	Svg Gen.
A black and white image of a cross		A circle with dots on it in black and white		Tv icon vector		A file icon with a white square on it		The japanese flag is shown in a white background		Chinese font design for the word 'love'	
A black and white icon of a pen and paper.		A black circle with a play button in the center.		Memory size pie chart		A knife is shown on a white background		An i symbol in a black circle		A pink light bulb icon.	
A purple and white striped logo		A white arrow on an orange background.		A white cross on a white background.		A square shape is shown in the image.		A gray circle with a question mark in the center.		A green smiley face with a smiley face	
Folder icon with a person icon		A fork and a leaf icon on a white background		An eye icon with a triangle shape		A yellow star is displayed on a white background.		Sun icon in a circle on a white background		An open book icon on a white background	
A black and white microphone icon.		A black star on a white background		A black and white icon of a no entry sign		A green button with a gray arrow pointing to it		A black and white icon of an envelope.		A black and white sun symbol.	
A black and white square with a black border.		A black and white map marker symbol.		A green and black circle with a white background		A black and white image of a game controller		A video player icon is displayed in a square.		The s logo in black and white	
A black and white headset icon.		A graph of the number of people in a class		A black and white icon of a trophy cup		A red square with a white letter A in the center.		An eye icon in a circle with a black dot		A blue and white icon of a camera.	
A pink pen on a white background vector		Standard js logo		A black and white logo of a curved shape		A blue and white paper clip icon.		A black and white illustration of a cell phone		A speech bubble with an x symbol on it	
A blue and white sign that says "continue setup".		A gray square with four squares on it		The facebook logo, with a white letter f		Apple pay logo		The flag of france and the flag of france		A yellow star with a black outline.	
A black and white triangle shape with two lines		A black and white icon of a lightning bolt		A yellow sign with the number 10 on it.		A black and white image of a japanese symbol		Chinese font for the word love		A pie chart with the words ethernet model	
A blue square with a black line in the middle.		A white and gray checkered pattern		The icon for the instagram app		A black circle is placed on a white background.		A cross symbol with four lines		The facebook logo is shown on a white background	
A black and white arrow pointing to the right		A red heart with a lightning bolt through it.		A black cross on a white background.		A person avatar icon on a white background		A black and white airplane icon.		A heart with a cross in it	
An envelope icon on a white background		The facebook logo in black and white		A white square with a black border		A black heart shape on a white background.		A blue button with the word stata on it		The flag of oman is shown in red and green	
A black cross on a white background		A blue bar is displayed on a white background.		A black and white logo with two arrows		A black square is placed on a white background.		A black circle with a white square in the middle.		A clock icon is displayed in a white background.	

Figure 25. **Text-to-SVG Generation** results using StarVector-1B on SVG-Stack test set (i).

Caption	Svg Gen.	Caption	Svg Gen.	Caption	Svg Gen.	Caption	Svg Gen.	Caption	Svg Gen.	Caption	Svg Gen.
Square icon vector		A red emotion with an angry face		The letter v is shown in black and white		A pink speech bubble on a white background		Whatsapp logo		A black and white icon of a sun	
A black and white image of a large X.		Twitter logo in a circle		A yellow and black shield with a rooster on it.		A black star with a white outline.		A grey and white image of a lock and key.		A white X is displayed on a pink background.	
A yellow triangle with a black outline.		A black and white logo with a white eye		A black and white icon of a document		Headphones icon vector		A black and white icon of a radio wave.		A black and white light bulb icon.	
A black and white cloud with a rain icon		A pixel style image of a brown box		A black and white logo for JPG.		A purple circle with an envelope icon		Pixel heart png		A black circle with a smiling face on it.	
Three stacked white boxes with green lines		The v logo with a blue arrow		Ethereum logo with a blue background		A blue cloud is displayed on a white background.		A black and white coffee cup icon.		A black and white YouTube icon.	
A letter a logo with a red and black letter		A purple circle with a down arrow in the center.		A black and white crescent on a white background		A black and white logo of a drop		A black and white pizza icon.		A phone icon in a heart shape	
The s logo in a black and white circle		A black circle with a white circle in the middle.		A black and white cloud icon.		The c logo in a green square		A black and white logo of a letter G.		A hand holding a bag with the letter r	
A blue arrow pointing up on a white background		The letter c in black and white		A phone icon with a plus sign		Youtube logo with a play button		A yellow emotion with glasses on it		Chinese character for cross	
A black and white shopping cart icon.		A black and white magnifying glass icon.		A brown and white battery icon.		A black and white clock with a blue hour hand.		Clipboard icon with a square and a square		The html5 logo with the word html5	
A purple square icon with two circles on it		A black and white image of a hamburger.		A black and white film reel icon		A green circle with an arrow pointing upwards.		A black and white sign with the letter e		Sun icon vector	
A blue circle with a white V in the center.		Credit card icon vector		A dollar sign icon on a white background		A black and white square with a white border		A black and white key symbol.		A red triangle with a white background	
Cross symbol clip art		A black and white icon of a refrigerator.		A purple and gray circle with a crescent moon		A white and purple logo of a letter C.		A purple circle with a knife icon		A black and white image of a question mark.	
A blank white card with a green border		A blue square with an x in the middle		A black and white smiley face icon		A black and white icon of a trash can.		A diagram of the heartbeat api		A red power button with an arrow pointing up	
Sun icon clip art		Google plus logo		Youtube logo with a red play button		A blue circle with a white center		A black and white mouse cursor icon		A black and white icon of a x	

Figure 26. Text-to-SVG Generation results using StarVector-1B on SVG-Stack test set(ii).













































































Caption	Svg Gen.	Caption	Svg Gen.	Caption	Svg Gen.	Caption	Svg Gen.	Caption	Svg Gen.	Caption	Svg Gen.
I love Halloween skull-and-bones emoji.		A necktie is a formal neckwear for a uniform.		A square is a quadrilateral shape in geometry.		The abstract square is distorted by a third.		Map pin.		The lorry is a delivery truck for transport.	
The rupee is currency.		Weight dumbbell.		Remove groceries from shopping basket.		Cleaning laundry with a washing machine.		The soup spoon is tableware.		A square and a cube.	
Hand-drawn sad emoticon upset and angry.		Player interface for video play.		A good idea is thinking about a light bulb.		The up arrow directs the move.		The arrow sign shows the three-way direction.		Wifi-router is an electronic device.	
A frozen sweet treat, popsicle ice-cream.		Wifi is online through w-lan.		Unlock your smartphone screen notification.		Mustache.		Pinpoint your music interests on the map.		The gear wheel's configuration is set by the cog.	
There is a signpost with a direction arrow.		Search with a magnifier to explore and view.		A square is a shape.		Bell notification.		A bar chart is a type of chart.		A tank-top is a type of shirt or undershirt.	
A necktie is an accessory for clothes.		The smiley emoji represents happy emotions.		The milk bottle is a dairy beverage.		The chef wore a toque in the kitchen.		Man has a mustache.		The arrow points down to the heart with love.	
I like the social like with a thumbs-up.		The Holy Cross is a Christian symbol.		Create new cloud server plus.		Upload bag shopping arrow.		The pin marks the location with a pointer.		A hexagon is a shape with 6 sides.	
The map pin marks the location.		Move in the direction of the sharp turn arrow.		The sun is a solar emoji emoticon.		The user's avatar is a map pin on their profile.		Right arrow.		The idea invented electricity's light bulb.	
Search using a magnifying glass.		Wifi.		Protect the privacy of a private document/file.		Letter or email message.		Data charted on a graph shows market statistics.		I drive my backpack with flash storage.	
Search and find with a magnifying glass.		A medal, award, badge, prize, or star.		Admin shields star.		The anchor stopped the heavy navy ship.		A rhombus is a symmetrical diamond-shaped figure.		Money is currency or cash, such as the euro.	
Chat using a speech bubble.		The unhappy face emoji.		I wore a necktie.		Check the checklist for a checkmark.		The audio cord connects to the microphone.		A rupee coin is currency used for payment.	
Aircraft, missile, rocket, spacecraft.		Web hosting on the cloud server.		Search with magnifying glass tool and zoom.		A beer mug.		Apple is a fruit.		The pound bulb is an idea for a light bulb.	
Usb flash drives store data in memory.		Move the arrow to expand and drag to drop.		Search with magnifier.		One dice is used in the game of luck.		Favorite star.		The bus is a vehicle for transportation.	
The arrow indicates the up direction where.		Candle.		A trapezoid is a shape.		Find location with GPS map pin.		Follow the arrow for direction.		The cassette tape contains music sound on tape.	

Figure 27. Text-to-SVG generation on FIGR-SVG test set (i).





















































































Caption	Svg Gen.	Caption	Svg Gen.	Caption	Svg Gen.	Caption	Svg Gen.	Caption	Svg Gen.	Caption	Svg Gen.
Left arrow points backward.		A people group network.		Women have a female gender symbol.		Spruce, pine, and Christmas trees are all plants.		Download the signpost arrow direction.		Flat brush used for painting flat illustration.	
Shop for a cart while shopping.		Sweep with a broom to clean the floor.		The bowl is a food container in the kitchen set.		The tool has options and settings with cogs.		The atom has a proton link in physics science.		Boat or ship.	
Map-pin marks a location on a map.		The map pin shows the location with GPS.		The erlenmeyer flask is a chemistry equipment.		Bug is an insect that can have a virus or error.		The modern thick arrow points right.		Male restroom sign.	
The pill bottle contained drugs for health.		Snowflakes fall in winter weather.		The pointer points to the arrow right.		Tie.		Service bell.		Anchor.	
Wifi is wireless connectivity to a network.		The pharmacy uses a mortar and pestle.		The ace of clubs is a poker card.		Move in the direction of the right arrow square.		Map-pin.		Minimal trophy.	
Watch TV on monitor.		Connect to the Wi-Fi.		Crown.		Lightning has electricity and high voltage.		Music sound files contain musical notes.		Find and focus by zooming with a magnifier.	
Media player plays movie, music, and video.		Buy a home online using a laptop.		Record sound with a microphone for voice audio.		Water in a plastic bottle is a beverage drink.		Up arrow indicates direction.		Cylinder.	
Boxer-briefs are a type of underwear.		Withdraw cash from ATM for euro money.		A smartwatch displays the time.		Calculate math using a calculator for accounting.		Target in crosshair.		A spoon is silverware.	
Arrow points up at junction for navigation.		The MKV file extension is a type of video file.		Search for glass with magnifying-glass view.		Christmas ball is a decoration.		Find location using GPS search and pins.		Secure euro payment with money lock.	
The up arrow indicates the top direction.		My favorite documents are starred on my computer.		Drawers in wardrobe.		Wifi is a network for internet connection.		Erlenmeyer flask and beaker.		Play videos on YouTube using the video player.	
Place a map pin.		The receiver follows the direction of the arrow.		Buy cart or basket from market for shopping.		The wise owl has evil eyes.		Chat or message when talking in a conversation.		Online video on desktop PC.	
Snowflakes have symmetry.		The mpg file is a document format.		Download the pointer arrow down using technology.		I rate a half star.		Euro.		The pitcher holds a beverage.	
Ice cream is a dessert food popped in a popsicle.		Christianity revolves around Jesus and the cross.		Sort tiles in a grid.		A file is a document, paper, or sheet with pages.		A QR code is a type of general barcode.		Search and find tools magnify and zoom.	
Online shopping trolley/cart finance.		Camera takes photos with lens for multimedia.		The magnifying glass can indicate an enlargement.		The thermometer measures temperature in Celsius.		Picture camera photography.		A Wi-Fi network signal.	

Figure 28. Text-to-SVG generation on FIGR-SVG test set (ii).

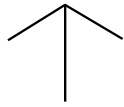










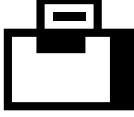

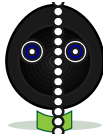






































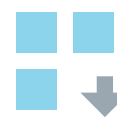







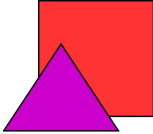

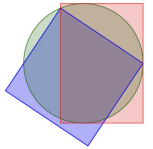

Text Instruction	Ground Truth	Temp = 0.0	Temp = 0.125	Temp = 0.25	Temp = 0.375	Temp = 0.5
Make an SVG of a black arrow pointing upwards.						
Generate an SVG of a black and white image of a box with a square shape.						
Design an SVG of a black cat in tears.						
Make an SVG of HC logo in a white circle						
Design an SVG of the tesla logo						
Design an SVG of a black and white image of three dots						
Draw an SVG image of a green recycling symbol, three arrows pointing inward.						
Make an SVG of a black and white icon of a computer program called "GIT".						
Generate an SVG of a blue and gray square with a blue arrow pointing to it.						
Design an SVG of a woman wearing a yellow shirt and blue jeans.						

Figure 29. **Text-to-SVG Results.** We show failure Text-to-SVG results using StarVector-8B (cherry picked examples that show limitations). We sample 5 different temperatures as an ablation, showing the sensitivity of this parameter during generation. Results are presented in SVG (not raster images)

Table 10. **Usage of SVG Primitives.** Image vectorization results of StarVector and VTracer applied to images containing basic shapes, such as circles, rectangles, and polygons, with varied colors and transparencies. The leftmost column shows the input images prompted for vectorization, and other columns show the output SVG code, with the SVG primitives in red color. StarVector accurately identifies and generates SVG code for each primitive, preserving their distinct characteristics. In contrast, VTracer relies on the `path` primitive, resulting in SVG code that captures the input image in terms of pixels, with less fidelity to individual shapes. Due to the length of VTracer’s SVG output, only the initial lines are shown. VTracer serves as a baseline model, representative of other baselines, which are omitted for space but exhibit similar behavior, primarily using `path` without shape recognition.

Test Example	StarVector	VTracer
	<pre> <svg width="150" height="150" xmlns="http://www.w3.org/2000/svg" xmlns:xlink="http://www.w3.org/1999/xlink"> <rect x="35" y="0" width="114" height="132" style="fill:rgb(255,50,50);stroke-width:1;stroke:rgb(0,0,0)"/> <polygon points="56.25,49.5,112.5,147.75,0,147.75" style="fill:rgb(200,0,200);stroke-width:1;stroke:rgb(0,0,0)"/> </svg> </pre>	<pre> <svg version="1.1" xmlns="http://www.w3.org/2000/svg" width="530" height="460"> <path d="M0_0_C174.9_0_349.8_0_530_0_C530_151.8_530_303.6_530_460_C355.1_ 460_180.2_460_0_460_C0_308.2_0_156.39999999999999_0_0_Z" fill="#FE3232" transform="translate(0,0)"/> <path d="M0_0_C4.06429570258797_3.6330418563035494_6.775718865918918_ 7.966461775462079_9.71484375_12.50390625_C10.273576812744153_ 13.357300872802739_10.832309875488278_14.210695459605478..." fill="#A95869" transform="translate(343,60)"/> <path d="M0_0_C40.260000000000005_0_80.52000000000001_0_122_0_ C123.16960912291928_66.0224184213402_123.16960912291928_ 66.0224184213402_123.416015625_93.876953125_ C123.47276957931317..." fill="#A95869" transform="translate(343,60)"/> ... </svg> </pre>
	<pre> <svg version="1.1" id="Layer_1" xmlns="http://www.w3.org/2000/svg" xmlns:xlink="http://www.w3.org/1999/xlink" x="0px" y="0px" width="64px" height="64px" viewBox="0 0 64 64" enable-background="new 0 0 64 64" xml:space="preserve"> <ellipse fill="none" stroke="#CD2A2A" stroke-width="2.5" stroke-miterlimit="10" cx="32.333" cy="32.083" rx="29.333" ry="31.25"/> <polygon fill="#CD2A2A" stroke="#CD2A2A" stroke-width="0.25" stroke-miterlimit="10" points="32.333,3.833,55.167,46.333, 9.5,46.333"/> <line fill="none" stroke="#000000" stroke-width="0.25" stroke-miterlimit="10" x1="0" y1="0.167"x2="64"y2="0.167"/> </svg> </pre>	<pre> <svg version="1.1" xmlns="http://www.w3.org/2000/svg" width="1196" height="1140"> <path d="M0_0_C394.68_0_789.36_0_1196_0_C1196_376.20000000000005_1196_ 752.40000000000001_1196_1140_C801.3199999999999_1140_406.64_1140_0_ 1140_C0_763.8_0_387.5999999999999_0_0_Z" fill="#FEFEFE" transform="translate(0,0)"/> <path d="M0_0_C1.8671875_0.24609375_1.8671875_0.24609375_4_1_C5.1328125_ 2.722656250000071_5.1328125_2.722656250000071_6.125_5.0625_ C8.342476699064719_9.859072675207557_10.955923466443437_ 14.196311705833779_13.8125_18.625_C17.369173047863796..." fill="#A95869" transform="translate(343,60)"/> <path d="M0_0_C1.3488281421188617_2.6976562842377234_0.8033826481469646_ 4.15164333837942_0_7_C-0.6599999999999682_7.329999999999984_ -1.3199999999999932_7.660000000000025_-2.8_C-2.6555511917654258_ 10.527332350510733_-2.6555511917654258_10.527332350510733..." fill="#A95869" transform="translate(343,60)"/> ... </svg> </pre>
	<pre> <svg width="300" height="300" xmlns="http://www.w3.org/2000/svg" version="1.1"> <rect x="110" y="10" width="160" height="230" fill="pink" stroke="red"/> <circle cx="160" cy="120" r="120" fill="tan" stroke="green"/> <fill="blue" fill-opacity="0.25" stroke="blue" stroke-width="2" /> </svg> </pre>	<pre> <svg version="1.1" xmlns="http://www.w3.org/2000/svg" width="910" height="934"> <path d="M0_0_C300.3_0_600.6_0_910_0_C910_308.22_910_616.44_910_934_ C609.7_934_309.4_934_0_934_C0_625.78_0_317.55999999999995_0_0_Z" fill="#928299" transform="translate(0,0)"/> <path d="M0_0_C1.484999999999994_0.9900000000000091_1.484999999999994_ 0.9900000000000091_3_2_C3_2.659999999999968_3_3.320000000000005_3_4_ C3.729609374999995_4.103124999999977..." fill="#A95869" transform="translate(343,60)"/> <path d="M0_0_C1.320000000000005_0.659999999999966_2.639999999999986_ 1.320000000000003_4_2_C3.009999999999991_2_2.019999999999982_2_1_ 2_C1_2.659999999999966_1_3.320000000000003" fill="#A95869" transform="translate(343,60)"/> ... </svg> </pre>
	<pre> <svg viewBox="0 0 100 100" xmlns="http://www.w3.org/2000/svg"> <circle cx="50.5" cy="39.5" r="40" fill="#60f0ff" /> <ellipse cx="50" cy="39" rx="50" ry="20" fill="#2810ff" /> <polygon points="10,100,50,42,90,100" fill="green" /> </svg> </pre>	<pre> <svg version="1.1" xmlns="http://www.w3.org/2000/svg" width="510" height="520"> <path d="M0_0_C168.3_0_336.6_0_510_0_C510_171.6_510_343.2_510_520_C341.7_ 520_173.39999999999998_520_0_520_C0_348.4_0_176.8_0_0_Z..." fill="#2E17FE" transform="translate(0,0)"/> <path d="M0_0_C0.6599999999999682_0_1.319999999999932_0_2_0_ C3.1463165283203125_1.3810577392578693_3.1463165283203125_ 1.3810577392578693_4_2_8955078125_3.356689453125_ C4.916970977783194_3 fill="#2E17FE" transform="translate(0,0)"/> <path d="M0_0_C1.5393356364577242_2.844848138263643_2.426629556693727_ 5.489600992013067_3.1875_8.625_C8.714349606195654_ 27.37779319117132_28.666518702768336_42.668449172855844_45_52_ C77.0038243523029..." fill="#2E17FE" transform="translate(0,0)"/> <path d="M0_0_C0.3300000000000409_0_6599999999999682_0_1_0_C1_101.31_1_ 202.62_1_307_C-16.159999999999968_307_33.31999999999999_307_-51_ 307_C-69.10137393182612_285.06539394143425_-69.10137393182612_ 285.06539394143425_-73.55859375_275.48046875..." fill="#2E17FE" transform="translate(0,0)"/> ... </svg> </pre>

is valid and effectively incorporates semantic concepts along with the accurate colors of the input image into the SVG code (see Figures [9 - 12]). However, it falls

short in terms of reconstruction fidelity, as GPT-4V was not specifically trained for reconstruction tasks, making these results predictable.

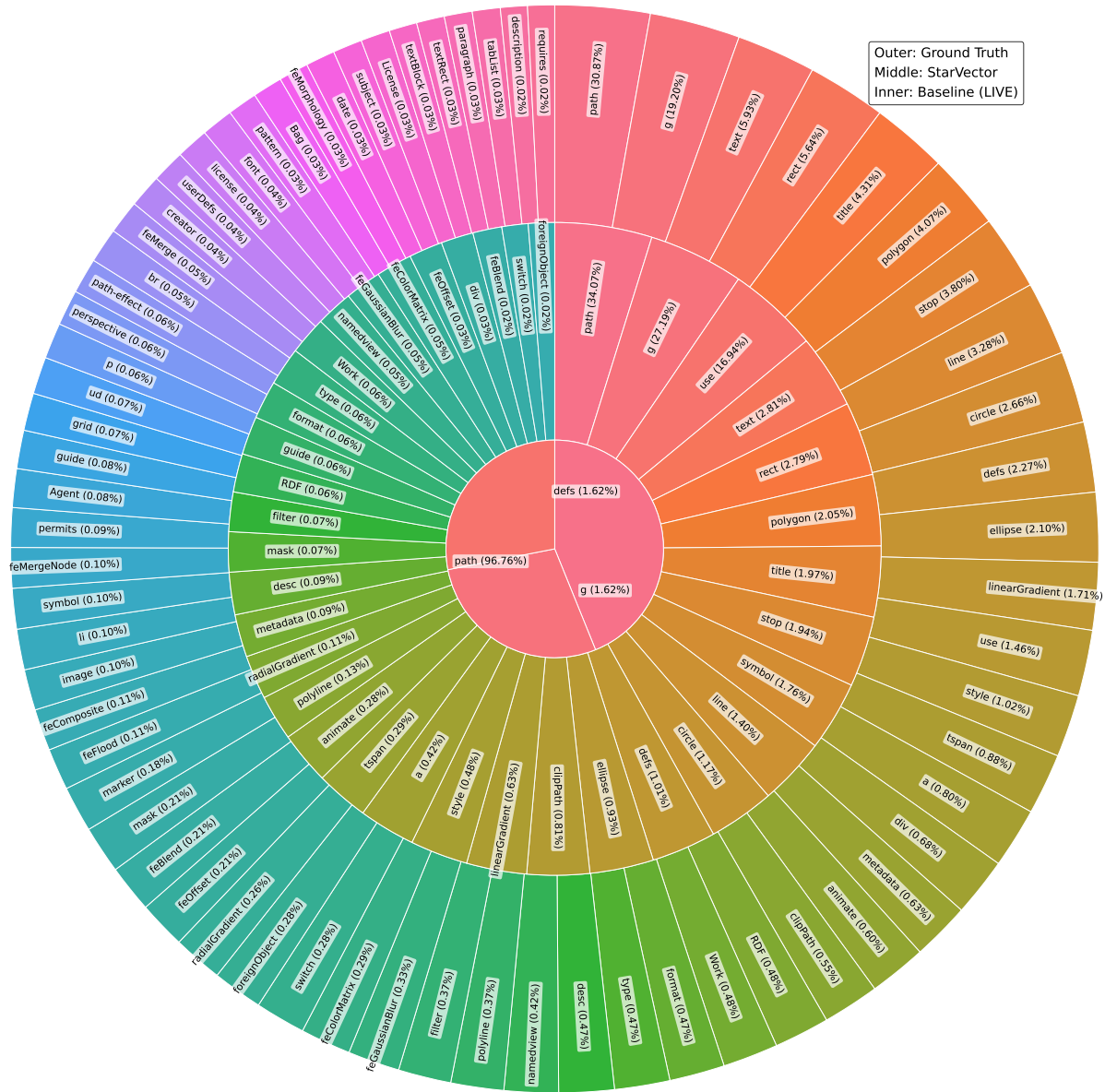


Figure 30. **Distribution of SVG Primitives and SVG Tags.** We show the frequency of SVG tags that appear in the ground truth of SVG-Stack dataset (outer ring), compared to the frequency of tags generated by StarVector (middle ring), and compared to our most performant baseline (LIVE). For visualization purposes, we apply a logarithmic scaling over the counts and show the base percentage in parentheses. StarVector generates SVGs that contain tags with a similar distribution to the ground truth. The baseline is limited to `paths` or basic primitives.

4. **LIVE** achieves the best results in terms of pixel-based metrics like MSE, LPIPS, and SSIM (see Tables [1, 2, 3]). Using 32 paths, it effectively represents a wide range of images, making it highly versatile for vectorization tasks, including natural images, which StarVector cannot

handle due to its specialized training. However, LIVE has notable limitations. It relies on a slow test-time optimization process (approximately 10 minutes per sample, using 32 paths, Table 3) to refine SVG outputs for good MSE scores, often introducing unwanted visual artifacts.

Additionally, its exclusive use of *path* primitives results in significantly larger SVG files (19k tokens), as shown in Table 1. In contrast, StarVector produces compact (3k tokens) and professional-grade SVGs by leveraging a variety of SVG primitives beyond paths, achieving higher precision and efficiency.

Notably, we find that this method, and the broader family of differentiable techniques it belongs to, is unsuitable for generating images that require specific primitives, such as those in diagrams. Its performance on the SVG-Diagrams benchmark is poor, as illustrated in Figure 14.

5. **DiffVG** offers comparable results as the other baselines in terms of reconstruction metrics, and it can produce suitable SVG image vectorization with 32 paths (same as LIVE), as seen in Tables [1, 2, 3]. These tables also show that DiffVG produces large SVGs as seen in the number of tokens (Tokens column), approximately 19k tokens, similar to LIVE. This means that files are extremely large compared to the ones of StarVector. As mentioned before, this is due to StarVector leveraging understanding and SVG primitives. Nevertheless, this method is substantially faster than LIVE, requiring 30 seconds per sample.

6. **Image Processing Methods: VTracer, Potrace, AutoTrace** Previous image processing methods for Image-to-SVG are powerful, excelling at fitting vector images to raster inputs with near-zero MSE while reliably capturing the shapes and colors of the original image. However, we identify several shared limitations across these methods: (1) lack of SVG file compression, as they often generate excessively long paths (see the Tokens column in Tables 1 and 2); (2) susceptibility to visual artifacts, especially with challenging patterns; and (3) poor performance in vectorizing diagrams, as illustrated in Figure 14. On the SVG-Diagrams benchmark, **VTracer, Potrace, and AutoTrace** struggle with producing high-quality results.

These methods perform best on images that can be segmented into distinct regions by color or texture but fail with complex patterns, such as small, closely spaced shapes or fine details. For example, in Figures 9 and 4, small polygons and intricate text are inadequately vectorized, with details often lost. All three methods are restricted to *path* primitives, limiting their ability to support features like optical character recognition or rendering text with the `<text>` tag. Although **Potrace** sometimes better preserves text compared to others, it still

cannot recognize or encode it semantically.

Despite these limitations, these methods excel in generation speed, making them highly efficient for many tasks. As shown in Table 3, **VTracer** and **AutoTrace** can generate SVGs in under a second, while **Potrace** typically takes around 10 seconds. In comparison, StarVector requires over a minute, and other baselines can take anywhere from 10 to 20 minutes. While StarVector’s semantic richness and compact outputs make it better suited for certain applications, these image processing methods demonstrate a clear advantage in scenarios where speed is critical.

7. **IconShop** achieves remarkable results on the SVG-FIGR dataset, as demonstrated in Table 4 and Figures of their original work [94]. However, it is not designed to handle SVG-Stack due to its restriction to modeling only the *path* primitive. StarVector outperforms IconShop, as shown in Tables 4, across metrics such as FID, FID CLIP, and CLIP Score. Qualitative examples further highlight StarVector’s superior performance in Text-to-SVG generation within the FIGR-SVG dataset.

8. **LLMs for Code Generation.** Methods utilizing LLMs to directly generate SVG code present appealing advantages. In our evaluation, we assessed GPT-4, GPT-4V, CodeLlama, and our proposed StarVector approach. By leveraging the code space, these models can utilize various SVG primitives based on their understanding of raster images, including path shapes and higher-order primitives like circles and text. This capability enables applications in new domains, such as diagram generation, as demonstrated by the results in Figure 14. However, most LLMs have not been specifically trained for the Image-to-SVG task, which limits their performance. In contrast, StarVector outperforms other LLMs due to its dedicated architecture and tailored training methodology, excelling in both image understanding and SVG generation.

Upon reviewing the complete set of results, we conclude that the StarVector approach is the only deep learning-based Image-to-SVG model capable of achieving results comparable to those of VTracer, Potrace, and AutoTrace. Furthermore, StarVector paves the way for novel research avenues in vector graphics generation, enabling applications such as diagram generation, Text-to-SVG generation, and potentially enhanced editing and understanding of vector images.

10.7. Human Evaluation

We conducted a human evaluation to compare the outputs of StarVector-8B, our best model, with those of the most powerful baselines. Participants were selected from diverse backgrounds and carefully screened for conflicts of interest, with none of the key authors involved. The evaluation was performed through a web interface (shown in Figure 31) that provided anonymized outputs and randomized sample presentations to prevent pattern recognition or bias.

The results, presented in Figure 5, demonstrate a strong preference for StarVector-8B across all settings, especially in SVG-Diagrams tasks. This highlights a disconnect between pixel-based metrics (e.g., MSE, SSIM) and human visual perception of SVGs. While baseline models often prioritize pixel-perfect reconstruction, human evaluators preferred StarVector’s sharp, well-defined shapes and its effective use of primitives (Figure 30).

Spearman correlation analyses between model metrics and human evaluations further emphasize this gap. MSE shows weak correlations (0.0596 and -0.1002), indicating its inadequacy as a predictor of human preferences. In contrast, DinoScore exhibits significantly stronger correlations, with values of -0.6193 and 0.6214. Moreover, a robust correlation of 0.7577 between differences in DinoScore and human evaluation scores highlights DinoScore as a more reliable metric for assessing SVG quality in alignment with human judgment.

11. StarVector Method

Here, we provide details on the StarVector architecture, training recipe, and generation process.

11.1. Architecture

11.1.1. Large Language Model

We consider several aspects when choosing the LLM to handle the SVG code generation. First, we require an LLM that can handle large token contexts during training, as SVG code samples are typically of long lengths (i.e., between 1,000-4,000 tokens for the most common SVG datasets but growing arbitrarily for much more complex vector graphics). Second, we need fast decoding during the generation of these large contexts. Finally, we would benefit from models that have been extensively pre-trained on general coding tasks to avoid early training costs.

Some prior works offer open-source models that fit these requirements. We explored the open-source families of models CodeGen [53], StarCoder [42] and StarCoder2 [49].

We empirically find the StarCoder family to be the most suitable choice for our requirements. StarCoder offers a pre-trained model with 1B parameters

(`starcoderbase-1b`) and a context length of 8k tokens, making it ideal for smaller-scale experiments while maintaining efficient generation speeds. The model employs Multi-Query Attention and a Fill-in-the-Middle objective, trained on 1 trillion tokens, with a context window of 8192 tokens. Its compact size ensures compatibility with GPUs, facilitating data parallelism during training—a crucial benefit when fine-tuning all network parameters, including the memory-intensive image encoder.

The second generation, StarCoder2, extends the context length to 16,384 tokens, presenting an exciting avenue for exploring training on SVGs with longer context requirements. For this, we leverage the 7B parameter version (`starcoder2-7b`), which incorporates a sliding window attention mechanism of 4,096 tokens and was trained on over 3.5 trillion tokens of code using the same Fill-in-the-Middle objective [49]. This enhanced context capacity and training scale make it a promising candidate for scaling SVG-based experiments.

These two types of LLMs define the backbones of our two StarVector variants. StarVector-1B is based on the `starcoderbase-1b` architecture and weights, while StarVector-8B is based on the `starcoder2-7b` architecture and weights.

11.1.2. Image Encoder

Our image encoding pipeline computes the images’ feature representations using a backbone image encoder and aligns them to the LLM via the Adapter module (see Figure 3). State-of-the-art image encoders are typically focused on natural images. However, our data contains images of logos, icons, fonts, or emojis, which usually contain no background (which we set to white) and mostly constant colors.

Note that the image encoder is used exclusively for the Image-to-SVG task and is not employed during the Text-to-SVG task. For Image-to-SVG, images must be projected into a representation with the same dimensionality as that of the LLM. We train the model to enable the LLM to ingest these representations and generate SVG code sequentially.

To choose the best encoder, we draw inspiration from the success of pre-trained encoder backbones in downstream computer vision tasks such as classification [62], retrieval, and generation [23], including both convolutional and transformer-based models. Specifically, we experiment with CLIP ViT-B/32 [62], ConvNext [46] (pre-trained on LAION-2B [75]), and VQGAN [23], which we pre-train on an image reconstruction task using raster images from SVG-Stack. For CLIP, we have $L_v = 257$ embeddings, including the CLS token. For VQGAN, we use the pre-quantization layers and flatten them to obtain $L_v = 196$ embeddings. For ConvNext, we flatten the last activation map to get $L_v = 49$ embeddings.

We explore several image encoders based on different

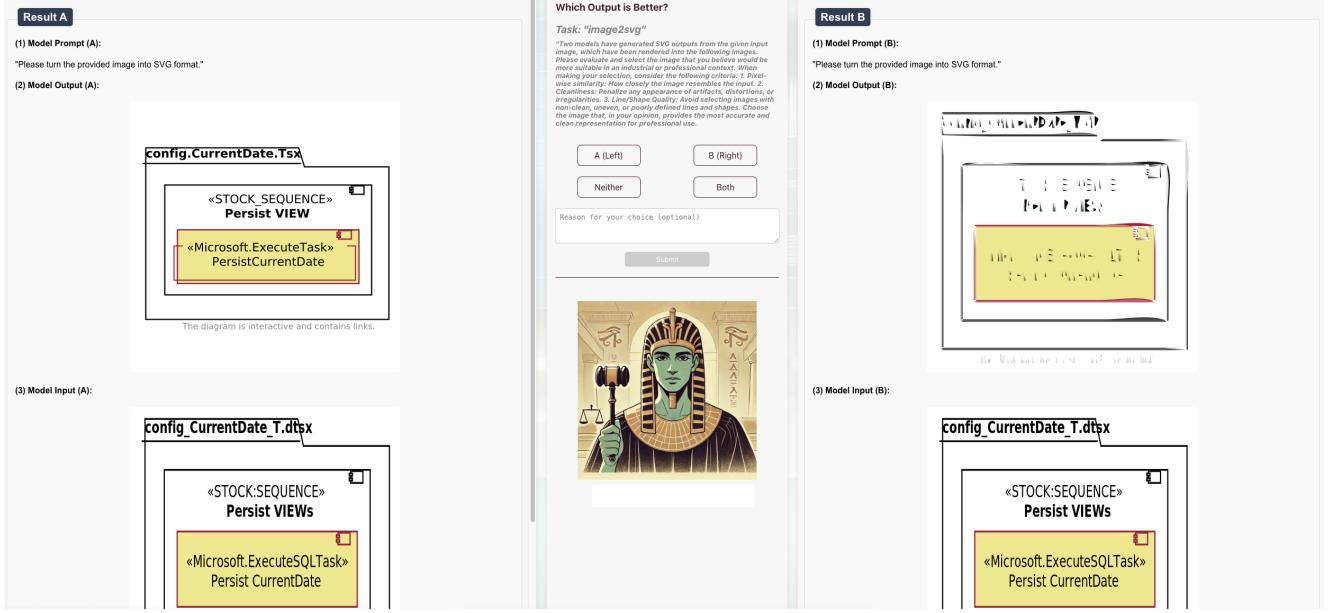


Figure 31. The web interface used during the human evaluation.

paradigms. VQGAN [23] is based on learning to project images to discrete tokens. First, we fine-tune an Imagenet [19]-pretrained VQGAN on the SVG-Stack dataset with the VQ-adversarial reconstruction task. We find that using the features before the quantization yields better results. ConvNext [46] is a convolutional backbone, which we extract features before pooling. We start from a LAION-2B [75]-pretrained checkpoint. Finally, ViT CLIP [62] is based on the Visual Transformer (ViT) [22] and is well prepared for autoregressive tasks. We extract all output representations. We use a LAION-2B pre-trained model. During the training of StarVector, all the parameters of the image encoders are updated. We find that the best choice is using CLIP. We consider that the gains in performance come from CLIP using more visual tokens (257) than the other image encoders.

The adapter first projects the features from the original dimensionality D_v to a dimensionality $D_v \times 2$, followed by a Swish non-linear activation function and a linear projection to the LLM dimensionality D_l . Finally, we apply a layer normalization [5]. We initialize the adapter parameters with Glorot [27]. Dropout [79] of 0.1 is applied at the beginning. These hyperparameters were found using a random search on SVG-Fonts.

Our results show that image resolution is essential to capture fine-grained details like texts or high-frequency patterns. As seen in the SVG-Diagrams dataset in Figure 6), diagrams and figures are part of the SVG-Stack dataset and present challenging horizontal or vertical aspect ratios. When images have these aspect ratios, we make the image fit in the 224×224 resolution, losing much detail, especially

for the OCR capabilities of reading rendered texts and accurately displaying them.

Additional results comparing image encoders can be found in Figures 19 and 22. These results show the boost in precision obtained when using CLIP. VQGAN and ConvNext often fail to capture the image’s shape and the *path*’s trajectory. We note that ConvNext performs better than VQGAN. These differences are also due to the differences in the number of parameters. The CLIP ViT-L/14 model that we use consists of 290M parameters, VQGAN consists of 29M, and ConvNext consists of 179M parameters.

Generating SVGs from natural images is out of the scope of this project. However, future work should focus on adapting this approach to natural images, drawing from [50] and [12] to create a dataset of natural images and SVG pairs.

The selected image encoder architectures for StarVector include two variants: one with fewer parameters and reduced image resolution, based on the CLIP ViT-B/32 model, which processes images at 224×224 pixels and is utilized in StarVector-1B. The second variant, SigLip (siglip-so400m302 patch14-384), has a larger number of parameters and processes images at a higher resolution of 384×384 pixels, and is employed in StarVector-8B. Given the positive results from the ViT architecture, we chose the SigLip variant due to its demonstrated effectiveness [97] and the enhanced resolution it provides.

11.2. Training

For training the StarVector model, we define the task of Image-to-SVG as an inverse rendering problem that converts a raster image (represented with visual tokens) into a

sequence of SVG code. This can be viewed as a sequence-to-sequence problem that models the translation between the image and SVG code domains. As detailed in Section 11, we utilize a CLIP ViT-B/32 for StarVector-1B and SigLip for StarVector-8B as image encoders, along with a non-linear adapter to generate a sequence of visual tokens.

The training process consists of two stages. In the first stage, the Image-to-SVG training phase, we construct sequences of visual tokens (produced by the image encoder and adapter) and SVG tokens, separated by a trigger token, `<svg-start>`. We train the LLM to learn these sequences on a large SVG-Stack dataset using a basic language modeling loss that calculates the cross-entropy loss in predicting the next token in a sequence based on the previous tokens.

This task enables the model to learn the concept of *drawing* with SVG vectors that resemble the input image. Importantly, this training can occur without supervision in the image domain (i.e., without pixel loss), relying solely on categorical cross-entropy loss for the LLM vocabulary introduced by the next-token prediction task.

In the second stage, we fine-tune the checkpoint from the first stage, which has learned SVG syntax through the Image-to-SVG task, on the Text-to-SVG task. During this phase, the image encoder is disregarded, as it becomes a Text-to-text task where the text instructions and SVG codes can be tokenized and processed directly by the LLM.

Training Details. We trained StarVector-1B on 1 node of 8 A100 80GB GPUs using Accelerate with DeepSpeed stage 2 and StarVector-8B on 8 nodes of 8 H100 80GB GPUs with Fully Shared Data Parallel (FSDP). For Image-to-SVG, we used total batch sizes of 128 and 512 for StarVector-1B and -8B, respectively, a learning rate of $1e-5$, and the AdamW optimizer. To optimize memory and computation, we employed bf16 precision, FlashAttention2, and gradient checkpointing. StarVector-1B took 7 days to train, while StarVector-8B took 10 days, with both models completing 2 epochs.

We use HuggingFace Transformers [93] and PyTorch [58] for the implementation. We use a batch size of 2. Images are processed with a resolution of 224x224, as defined by the pre-trained CLIP image encoder, and process a maximum of 8192 tokens, considering the 257 visual tokens and the rest for the SVG tokens. We use gradient batch accumulation of 8 and train on a data parallel setup with 4 A100 80GB GPUs, having an effective batch size of 64. The learning rate is set to 5×10^{-4} for training, using AdamW optimizer [48] for approximately five days of training on the SVG-Stack dataset.

11.3. Generation

Here, we describe how to sample SVG code from our model. As a decoder-only LLM [42], StarVector first computes the key-value (KV) cache using the visual tokens from the image and then produces the initial set of output logits. This stage is often quick because the model can process the entire visual token sequence simultaneously [78]. The selected token from the output logits is then input back into the model, which generates logits for the subsequent token. This process is iteratively repeated until the model produces the desired quantity of tokens. Our approach uses architectural improvements for fast decoding, such as FlashAttention [18] and Multi-Query Attention [77]. We leverage vLLM to improve inference speed.

We perform a grid search on SVG-Emoji and SVG-Icons validation sets to select the correct sampling temperature. The choice of temperature does not strongly impact the results. However, a 1-point increase in performance is observed on CD for SVG-Emoji using temperatures close to 1.0.

We also present an ablation study of StarVector-1B popular decoding techniques [31, 52, 76, 85]. Specifically, we experiment with greedy decoding, beam search, and nucleus sampling with top- p . Results are shown in Table 7. The use of nucleus sampling with top- $p=0.9$ and temperature $T=0.5$ (no beam search) shows to be the best option. The beam search improves the greedy decoding baseline but does not work well when combined with nucleus sampling, increasing the inference time. In sum, we recommend nucleus sampling [31] with top $p=0.9$ and temperature between 0.5 and 0.9 for the best performance.

Are the SVGs valid and compilable? A common issue when generating SVGs with our approach is that the maximum token length of the LLM might not be sufficient to complete the SVG code, leading to compilation errors. We find that 85% of the generated SVG fit within the context length and compile successfully. The remaining incomplete samples are post-processed with `cairosvg` to produce a complete and compilable SVG. However, in some cases, parts of the image may be lost during this process. *With this technique, 100% of the generated SVGs are valid and compilable.*

Improving SVG Quality Through Sampling. The generation process is stochastic, meaning the outputs may sometimes take an incorrect *path*, leading to failed generations or repetitive patterns. To address this, we propose a simple baseline approach: generate k SVG outputs with varying sampling parameters (e.g., by adjusting the temperature),

then compare the outputs with the ground truth using a visual metric (we propose DinoScore) to select the most accurate result. In an ablation study conducted on SVG-Stack using StarVector-8B with $k = 1$ and $k = 5$, we observe a boost in DinoScore by 0.12. Empirically, after sampling 100 test samples, we find that 32% of the SVGs are more accurate when using $k = 5$, though this increases the generation time by a factor of k . The use of vLLM helps mitigate the slower sampling process, as it operates much faster. For further improvement in code generation, previous work has used MCTS techniques [7], which leverage visual feedback more effectively to guide the generation and enhance the stochastic sampling process.