Evaluating Constrained Iterative Refinement for SVG Generation with Off-the-Shelf VLMs

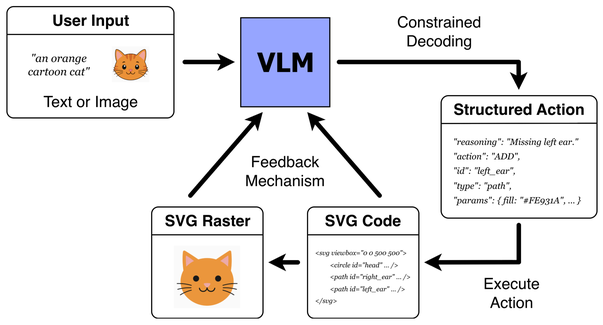
UMassAmherst | Manning College of Information & Computer Sciences

UCF | Institute of Artificial Intelligence
UNIVERSITY OF CENTRAL FLORIDA

Matthew Perlman[1], James Beetham[2], Niels da Vitoria Lobo[2], Amrit Singh Bedi[2], Mubarak Shah[2]
[1] University of Massachusetts Amherst, [2] University of Central Florida

ABSTRACT

SVGs power much of the modern visual ecosystem, yet the recent wave of generative models focus almost entirely on rasterized images. We explore whether inference-time scaffolding can unlock SVG generation capabilities in off-the-shelf vision-language models (VLMs). We introduce a constrained iterative refinement framework that combines visual feedback, structured editing, and grammar-constrained decoding to address the challenges many VLMs face when generating SVGs. Across multiple VLMs and generation settings, we find that constrained decoding improves the rate of successful generation, while iterative refinement reveals a deficit in visual reasoning and self-correction. Our results highlight both the promise and current limitations of using inference-time scaffolding to adapt general-purpose VLMs for vector graphics generation.



Our Constrained Iterative Refinement Harness.

INTRODUCTION

Scalable Vector Graphics (SVGs) are widely used for web, interface, and illustration content because they are resolution-independent, compact, and directly editable. Despite the practical benefits of SVGs, generating them remains challenging for general-purpose vision-language models (VLMs): the model must jointly produce valid code, reason about spatial relationships, and translate precise visual details into a new medium. Existing SVG generation methods address these challenges through specialized architectures or task-specific training, decoupling them from the rapidly improving ecosystem of general-purpose multimodal models.

We investigate whether inference-time scaffolding is sufficient to make off-the-shelf VLMs effective SVG generators. In particular, we systematically evaluate iterative visual refinement, three levels of editing structure, and grammar-constrained decoding for both text-to-SVG and image-to-SVG generation across multiple model scales. Our study characterizes when these mechanisms improve syntactic reliability, how they affect visual quality, and how well current VLMs can effectively translate rendered feedback into changes to the underlying vector representation.

SYSTEM OVERVIEW

We introduce a harness that equips off-the-shelf VLMs with two complementary, inference-time mechanisms: iterative refinement based on rendered visual feedback and grammar-constrained decoding that restricts outputs to appropriate SVG and JSON syntax. Together, these mechanisms target the weaknesses of general-purpose VLMs, allowing the model to iteratively refine and correct its output while remaining bounded to a syntactically valid generation space.

QUANTITATIVE RESULTS

We evaluate our framework on MMSVG-Bench across multiple model scales (Qwen3.5 9B and 27B), with and without grammar-constrained decoding, and across all four generation pipelines. Metrics are computed over successfully compiled images, then averaged equally across the two models.

		Text→SVG				Image→SVG			
Decoding	**Pipeline**	**Compile↑**	**CLIP↑**	**Aesthetic.↑**	**#Tokens**	**Compile↑**	**DINO↑**	**LPIPS↓**	**#Tokens**
Unconstrained	Single Turn	**0.837**	**0.255**	**4.655**	1.4k	0.630	**0.704**	0.396	1.7k
	Unstructured	0.475	0.246	4.560	4.8k	0.113	**0.704**	**0.372**	10.7k
	Semi-Structured	0.180	0.205	4.415	7.9k	0.500	0.617	0.418	8.8k
	Structured	0.785	0.226	4.480	3.3k	**0.650**	0.633	0.413	8.1k
Constrained	Single Turn	**0.830**	0.220	**4.530**	0.6k	**0.830**	0.411	0.579	0.4k
	Unstructured	0.599	**0.242**	4.485	5.2k	0.647	**0.617**	0.430	5.1k
	Semi-Structured	0.512	0.206	4.445	2.7k	0.375	0.464	0.506	3.2k
	Structured	0.790	0.218	4.440	2.4k	0.695	0.603	**0.420**	3.3k

Results averaged across Qwen3.5 9B and 27B on MMSVG-Bench.

QUALITATIVE RESULTS

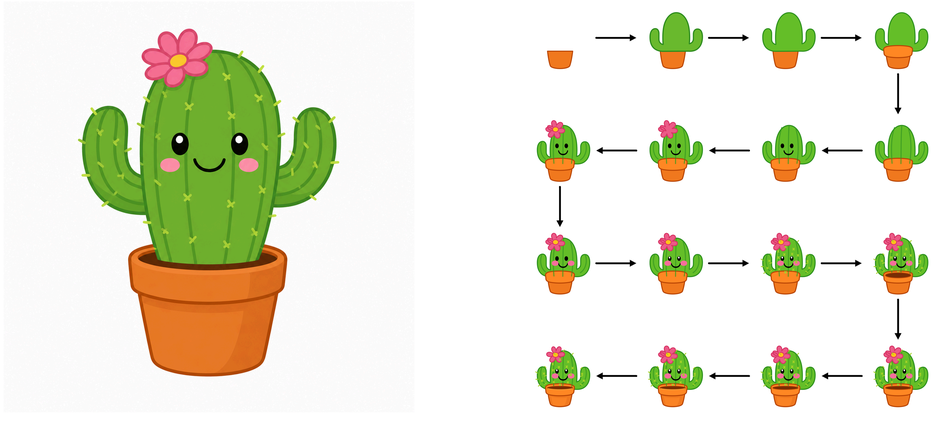
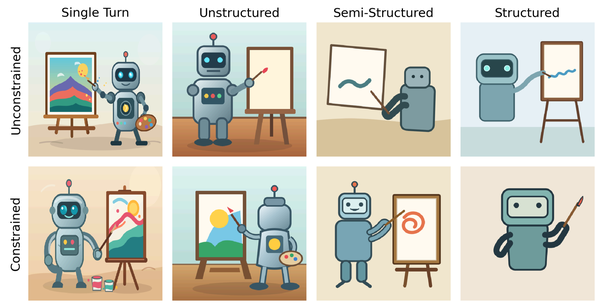


Image-to-SVG generation trajectory produced by GPT-5.6 with unconstrained structured actions. Left: reference image. Right: intermediate SVG renders across 16 iterative refinement steps.



Images generated by GPT-5.6 using each pipeline with the Text-to-SVG prompt "A robot painting".

DISCUSSION

Constrained Decoding Can Improve Compilation. We find that constrained decoding substantially improves compilation rates for most pipelines, a result consistent with our initial hypothesis.

Constrained Decoding Hurts Visual Quality. Across visual metrics and generation settings, constrained decoding consistently lowered image quality. We hypothesize that enforcing our limited grammar steers token generation toward unlikely and visually flawed SVGs.

Small Models Struggle With Visual Feedback. We find that for the models tested, iterative refinement has an inconsistent effect on visual quality, and often lowers compilation rates by introducing more opportunities to fail. Based on empirical results, we hypothesize these small models struggle to connect visual inconsistencies with the underlying SVG code.

FUTURE WORK

Several promising directions remain for improving inference-time scaffolding. Richer action spaces, alternative visual feedback mechanisms, stronger constrained decoding strategies, and prompt meta-optimization all have the possibility of improving generation quality and success rate. We also believe different evaluation metrics may better capture image quality and prompt adherence.

Beyond extending our harness, a deeper mechanistic analysis of these models' failure to incorporate visual feedback may inform improved general-purpose VLM spatial-reasoning capabilities.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation through the University of Central Florida Computer Vision REU Program. We thank the UCF Institute for Artificial Intelligence for their computing resources and guidance throughout this project.