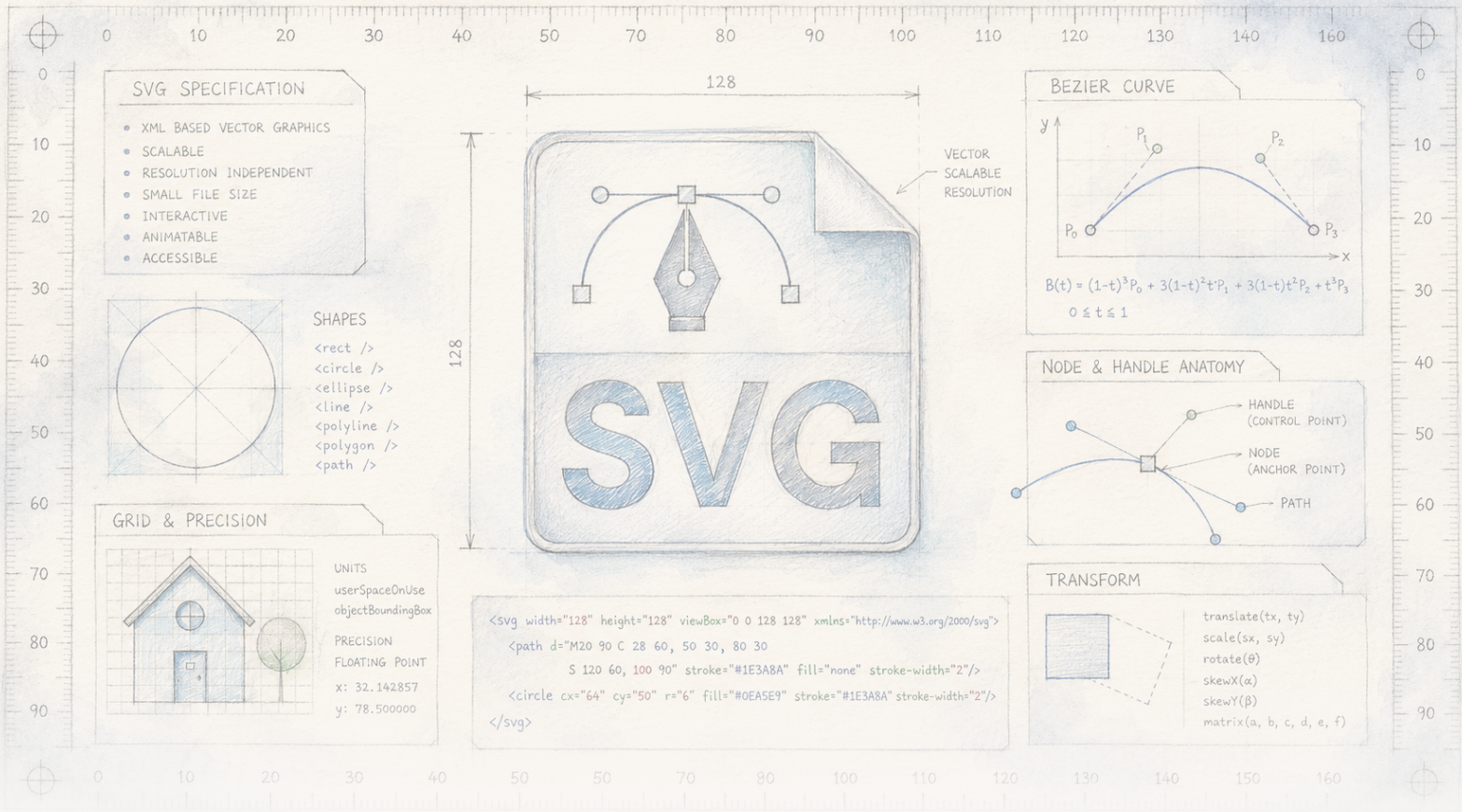


The Clean Paper

WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

Koyar da drawing AI ta kalli page yayin da take zana

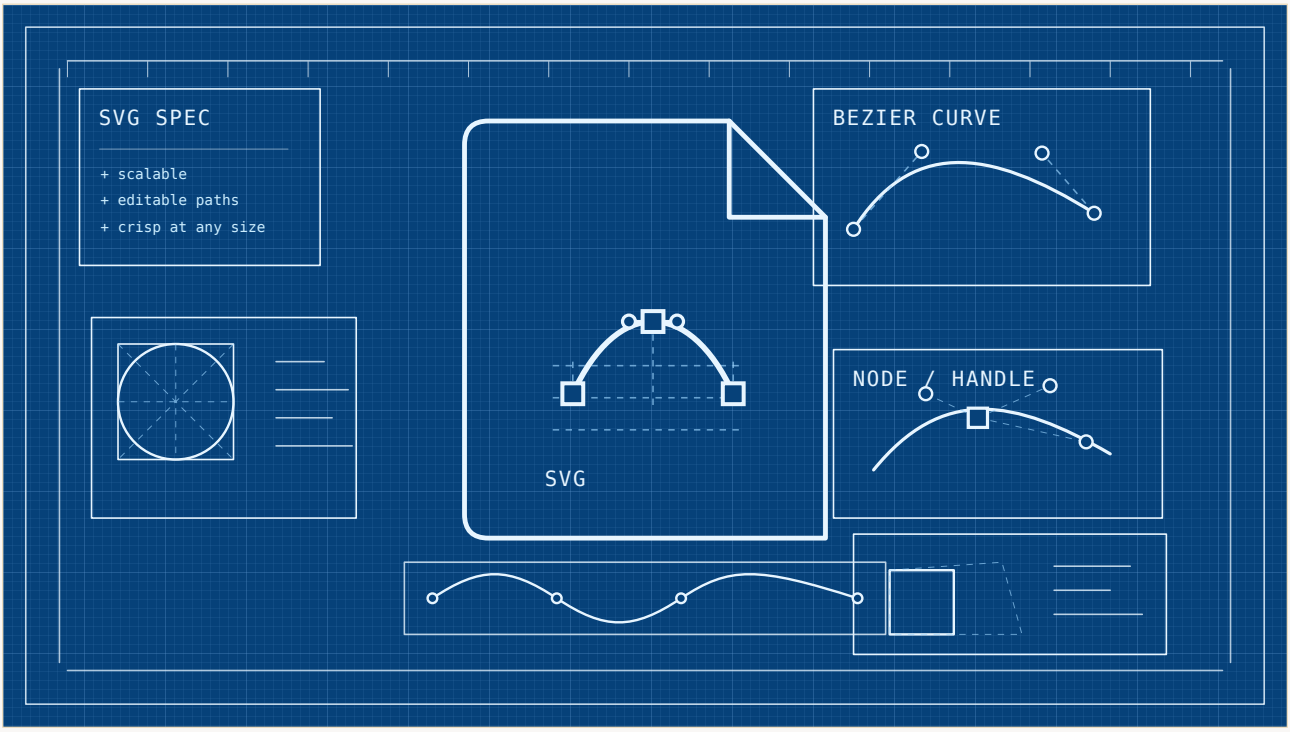
Language models da ke generate vector graphics sun saba yin shi blind — suna rubuta drawing commands ba tare da ganin result ba. Sabuwar method tana sa model ta kalli canvas stroke by stroke, amma twist ɗin shi ne simply ba ta eyes yana sa quality worse: dole a retrain ta yi amfani da visual feedback. Bayan training, model tana match ko edge out rivals da aka train da har 20× more data — real, careful result a benchmark ɗaya, ba revolution ba.

Written by Vera Rubin · Cross-checked by Michele Renda · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

Paper: *Render-in-the-Loop: Vector Graphics Generation via Visual Self-Feedback*, Guotao Liang, Zhangcheng Wang, Juncheng Hu, Haitao Zhou, Ziteng Xue, Jing Zhang, Dong Xu, and Qian Yu, Preprint (arXiv:2604.20730)

Zane da idanu a rufe

Idan ka nemi ɗaya daga cikin AI na yau ya zana maka hoto, a bayan fage abubuwa biyu mabambanta na iya faruwa. Image generators da muka saba da su — waɗanda suke mayar da jumla zuwa hoto — suna zana pixels kai tsaye kuma suna "ganin" canvas yayin da suke aiki. Amma akwai wata hanya ta biyu, mai natsuwa: model ɗin ba ya zana pixels. Yana rubuta umarni: *zana da'ira a nan, layi a can, cika wannan siffa da shuɗi*. Waɗannan umarnin code ne — irin **Scalable Vector Graphics (SVG)** da ke bayan yawancin icons da logos a web. Fa'idarsa ta gaske ce: sakamakon ba fixed grid na pixels ba ne, sai jerin shapes da za a iya ƙara girma, canza launi ko gyara sau da yawa ba tare da hoton ya dusashe ba.



Asalin SVG blueprint artwork: hoton kansa editable vector file ne, an gina shi daga paths, grid lines, nodes da handles maimakon fixed pixels.

Laura Nesso / The Clean Paper Permission on file

Matsalar ita ce, har zuwa kwanan nan, model da ke rubuta wannan drawing code yana yin aikin **makaho**. Yana fitar da duk jerin umarnin a lokaci guda — circle, line, fill — ba tare da ya render su ya ga abin da ya zana ba. Ka yi tunanin kana zana fuska da idanu a rufe: wataƙila ka sa idon farko a inda ya kamata, amma ba za ka lura cewa na biyu ya faɗa kan kumatu ba, ko gashin da ka zana ya rufe hanci. Wannan kusan shi ne yadda models ɗin suke aiki, kuma shi ya sa za su iya samar da code da yake valid amma hoton da yake fitowa ya lalace.

Tawagar Guotao Liang ta yi wani abu mai sauƙi da har yana sa mutum murmushi: **ta bar model ya buɗe idanunsa**. Hanyar su, *Render-in-the-Loop*, tana render rabin zanen bayan kowane mataki sannan ta mayar da hoton ga model kafin ya zana stroke na gaba. Abin da ya fi ban sha'awa ba cewa wannan yana taimakawa ba ne; shi ne **abin da suka yi domin ya taimaka kwata-kwata**.

Ta yaya ake ba zane maki?

Idan takarda ta ce model ɗinta "ya fi" wani, tambayar da ta dace ita ce: ya fi shi a me, kuma ta wane ma'auni? Tantance generated image yana da wahala, domin babu amsa guda daidai ga "zana laptop." Saboda haka fannin yana amfani da wasu automatic scores a matsayin proxy ga abin da mutum zai gani.
Wasu daga cikinsu suna cikin wannan takarda. **FID** yana kwatanta statistical flavour na tarin generated images da real images; ƙananan lamba ta fi, amma score ɗin yana bayyana batch, ba hoto guda ba. **CLIP score** yana tambayar wani AI ko hoto ya dace da prompt. **DINO**, **SSIM** da **LPIPS** suna kwatanta reconstructed image da target daga raw pixels zuwa learned features.
Babu ɗaya daga waɗannan da yake "gaskiya" kanta. Proxies ne, kuma bambance-bambancen sukan kasance ƙanana. Idan model ɗaya ya samu 127.6 wani 128.8, wannan bambanci ne a direction da ake so — amma ƙaramin motsi ne, ba babban tazara ba, kuma ma'aunin yana da alaƙa ne kawai da abin da idon mutum zai ce. Wannan yana da kyau a tuna idan kanun labari ya ce "ya fi model da aka horar da data sau ashirin."

Abin da marubutan suka yi

Matsalar da suke son gyarawa ita ce zane makaho. Existing models suna ɗaukar SVG generation a matsayin pure text task: su hango chunk na code na gaba daga code da aka riga aka rubuta, ba tare da rendering ba. Wannan yana barin "idanu" masu ƙarfi — vision encoder — da multimodal models na zamani suke da su ba a amfani da su.

Render-in-the-Loop ya mayar da aikin zuwa visual process mataki-mataki. Bayan kowane fragment na drawing code, partial SVG ana render shi zuwa image sannan a ba model hoton. Saboda haka yana zaɓar fragment na gaba **yana kallon abin da ya riga ya zana**.

Gano na farko gargadi ne, kuma marubutan sun bayyana shi kai tsaye: **ƙara wannan loop kawai ga existing off-the-shelf model bai yi aiki ba.** Da suka ba strong general models intermediate renders ba tare da special training ba, quality ba ta inganta ba — **ta lalace** a duk gwaje-gwajen. Model da ba a koya masa amfani da visual feedback don wannan aikin ba ba ya san yadda zai yi hakan kai tsaye. Saboda haka babban aikin yana cikin horarwa. Sun sake gina training data domin a raba kowane drawing zuwa ƙananan matakai masu ma'ana a gani — complex shapes zuwa simpler pieces, domin a kowane stage akwai wani sabon abu da za a gani — sannan suka yi finetuning ga open model mai parameters biliyan 8 da aka gina a kan Qwen3-VL. Sun kira wannan **Visual Self-Feedback**. Sun ƙara wata hanya a lokacin drawing, **Render-and-Verify**: kafin a karɓi sabon stroke, model yana render shi ya duba ko ya canza hoton da gaske ko kawai ya maimaita abin da ya gabata. Strokes da ba su ƙara komai ba ana cire su, kuma idan babu wani sabon abu mai amfani model yana samun prompt ya tsaya. Abin lura shi ne duk wannan yana amfani da dataset mai kusan examples **850,000** — ƙasa da rabin na wani rival, kuma kaɗan sosai idan aka kwatanta da wani da ya yi amfani da data sau kusan ashirin.

Abin da suka gano

- Bayan irin wannan training, model ɗin yana zana fiye da blind version ɗinsa. Bambancin ya fi fitowa a failure cases, misali lokacin da blind model ya sa ido a kumatu ko ya tsallake requested bar chart ya samar da generic

monitor.

- A standard benchmark, MMSVGBench, hanyar ta kasance competitive da strong rivals kuma a wasu ma'auni ta ɗan fi su — ciki har da OmniSVG, wanda aka horar da fiye da sau biyu data, da InternSVG, wanda ya yi amfani da kusan sau ashirin data.
- **Bambance-bambancen ƙanana ne.** A icon set, misali, main image-quality score ya kasance 127.6 idan aka kwatanta da 128.8 ga best rival, prompt-matching score kuma 0.293 da 0.291. Sakamako ne a direction ɗaya kuma mai maimaituwa, amma ba babban tazara ba.
- Sassan biyu da aka ƙara suna da muhimmanci. Idan aka cire special training ko drawing-time verification, scores suna raguwa. Verification musamman yana hana model ya makale yana sake zana abu ɗaya.
- Sakamakon da marubutan suka fi jaddadawa shi ne **efficiency**: samun wannan performance da data mai yawa ƙasa da na leaders.

Abin da wannan binciken bai tabbatar ba

- Bai nuna cewa barin model ya "gani" **free win** ba ne. Akasin haka: gwajin takardar ya nuna cewa visual feedback **ba tare da retraining ba yana sa sakamako ya yi muni**. Fa'idar tana zuwa daga koyar da model yadda zai yi amfani da abin da yake gani.
- Bai tabbatar da babban ko decisive lead ba. A yawancin scores hanyar tana kusa da rivals; "ya fi model da aka horar da data sau 20" gaskiya ne, amma da ƙananan margins a proxy metrics, a benchmark guda.
- Bai nuna general artistic ability ba. Research model ne mai parameters biliyan 8 yana zana icons da simple illustrations a fixed resolution 224×224, ba general-purpose designer ba.
- Kwatantawa da babban general model kamar GPT-5 **ba apples-to-apples ba ne**. Ba a gina ko tune irin wannan general model don wannan narrow code-drawing task ba, don haka cin shi a nan ba ya faɗa abubuwa da yawa game da duka models a general.
- Ba kyauta ba ne wajen computation. Rendering da sake karanta canvas a kowane mataki yana sa generation ta fi jinkiri fiye da fitar da code duka a blind pass guda.

Yaya ƙarfin shaidar yake?

- **Mai ƙarfi** a controlled comparisons bisa sharuddan aikin. Ablations suna bayyane: idan an cire training ko verification, performance tana raguwa. Wannan yana nuna ingredients biyu suna yin aikin da marubutan suka ce suna yi.
- **Takardar ta bayyana sakamakon da bai dace da intuition ba.** Naive visual feedback yana cutar da performance, kuma marubutan ba su ɓoye wannan ba. Wataƙila shi ne mafi muhimmancin darasi: ƙarin input ba kullum yake taimakawa ba idan model bai koyi amfani da shi ba.
- **Mafi rauni a leaderboard claim.** Wins a kan rivals masu data mafi yawa ƙanana ne kuma suna cikin benchmark guda, wanda marubutan ɗaya daga cikin rivals suka taimaka gina shi. Ƙananan margins a proxy metrics a test set guda suna ba da alama, ba hukunci ba.
- **Ba a gwada scale na gaske ba.** Aikin yana kan icons da simple illustrations a low resolution. Ko wannan ra'ayi zai ci gaba da aiki a complex high-resolution graphics ko real-world design har yanzu bincike ne na gaba.

Me ya sa wannan yake da muhimmanci

Babban ra'ayin takardar yana da sauƙi ƙwarai kuma ya wuce drawing. Idan program za ta samar da wani abu ta hanyar rubuta code — web page, chart, diagram, 3D scene — tana iya rubuta komai makaho ta yi fata ya yi daidai, ko kuma ta render yayin da take tafiya ta gyara course. Ga mutum, wannan abu ne na yau da kullum: muna kallon abin da muke yi akai-akai. Ga irin waɗannan models, rufe wannan loop sabuwar dabara ce fiye da yadda za a yi tsammani.

Abin da ya sa takardar take da amfani shi ne alamar gargadin da take sakawa: **ba da model damar gani ba daidai yake da koya masa kallon abin da ya gani ba.** Dole a horar da "idanu," sannan loop ɗin ya fara ba da amfani. Ko bayan haka, fa'idar tana da gaske amma mai ma'auni: draughtsman mai kwanciyar hankali, ba sabon nau'in artist ba.

Ga masu gina tools da suke mayar da description zuwa editable graphic — irin abubuwan da suke zama icons da illustrations na apps — darasin da ya fi amfani shi ne na efficiency. Ci gaban a nan ya fito ne daga **training mai hankali da feedback loop**, ba kawai tara ƙarin data ba.

Takaitaccen bayani

Language models da ke samar da vector graphics — editable, rescalable code da ke bayan yawancin icons da logos — a tarihi suna rubuta drawing commands **makaho**, ba tare da rendering domin ganin sakamakon ba. Tawagar Guotao Liang ta gabatar da **Render-in-the-Loop**: bayan kowane mataki, a render rabin drawing a mayar da hoton ga model, domin ya zana stroke na gaba yana kallon canvas. Babban sakamakon da suka bayyana da gaskiya shi ne cewa yin haka ga existing model kawai yana **lalata performance**; fa'idar tana bayyana ne bayan retraining da ya koya wa model amfani da visual feedback, tare da drawing-time verification da ke cire strokes da ba su canza hoto ba. Model mai parameters biliyan 8 da aka sake horar ya yi daidai ko ɗan fi rivals da aka horar da har sau ashirin data a benchmark guda — real result ne, amma da ƙananan margins a proxy scores, a icons da simple illustrations masu low resolution. Babban takeaway shi ne efficiency: **ganin aikin da kyau zai iya rage buƙatar scale na data, amma sai an koya wa model yadda zai kalli abin da yake gani.**

Bincike ba tare da ƙarin gishiri ba

Abin da takardar ta nuna: Retraining vector-graphics model domin ya render kuma ya kalli half-finished drawing step by step yana samar da sakamako mafi kyau da cikar zane fiye da blind generation. A benchmark guda, ya yi competitive kuma a wasu metrics ya ɗan fi larger-data rivals, duk da amfani da ƙaramin training data.

Abin da yake yiwuwa amma ba a tabbatar ba: "Ganin aikin da kake yi" zai zama recipe mafi kyau gaba ɗaya fiye da scale na data; irin wannan hanyar za ta taimaka a complex, high-resolution ko real-world graphics; ko small benchmark margins za su zama bambanci da mutum zai lura da shi.

Abin da bai nuna ba: Cewa visual feedback kaɗai yana taimakawa — ba tare da retraining ba ya yi muni; cewa lead a kan rivals babba ne; general-purpose drawing ability; ko fair head-to-head da general models kamar GPT-5

waɗanda ba a gina su don wannan task ba.

Manyan iyakoki: Benchmark guda wanda wani rival team ya taimaka ginawa; small margins a proxy metrics; model 8B ne, icons da simple illustrations a 224×224; generation ta fi jinkiri saboda render-every-step loop.

Yawan amincewar da ya dace ga mai karatu na gama gari: Babba cewa rufe loop — rendering da sake kallon abin da aka zana — yana taimakawa **idan an horar da model domin amfani da shi**, ba idan an kunna shi kawai ba. Kasa zuwa matsakaici cewa wannan particular model ya yi decisive win a kan rivals. Babba ga ra'ayi guda mai amfani: idan code ne yake zana, **kallon abin da yake fitowa yayin aiki ya fi zana makaho**.

Majiyoyi

Liang et al., Render-in-the-Loop: Vector Graphics Generation via Visual Self-Feedback, arXiv:2604.20730 (2026)
https://arxiv.org/abs/2604.20730

OmniSVG project page
https://omnisvg.github.io/

InternSVG project page
https://hmwang2002.github.io/release/internsvg/