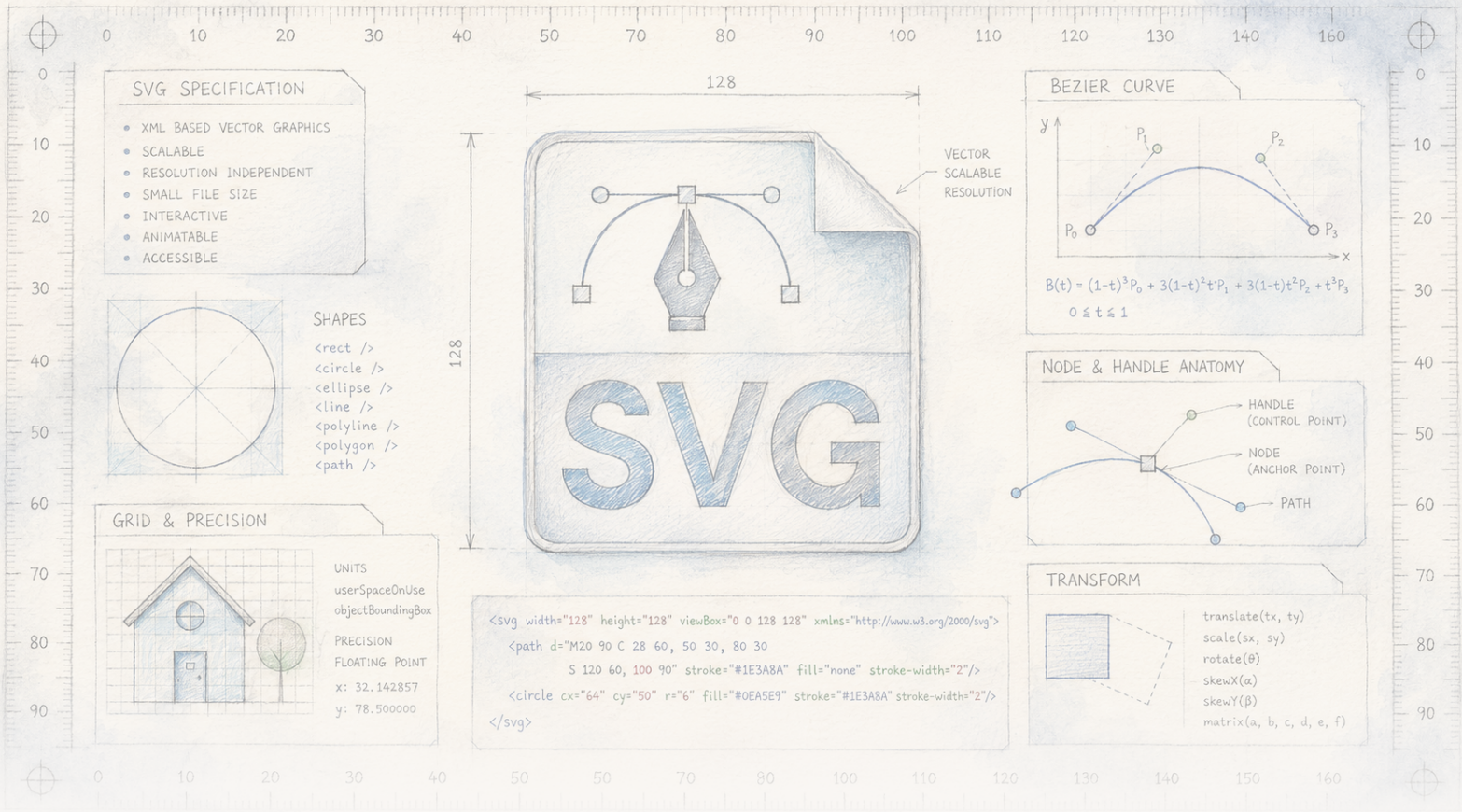


The Clean Paper

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

Kuifundisha AI inayochora kuangalia ukurasa

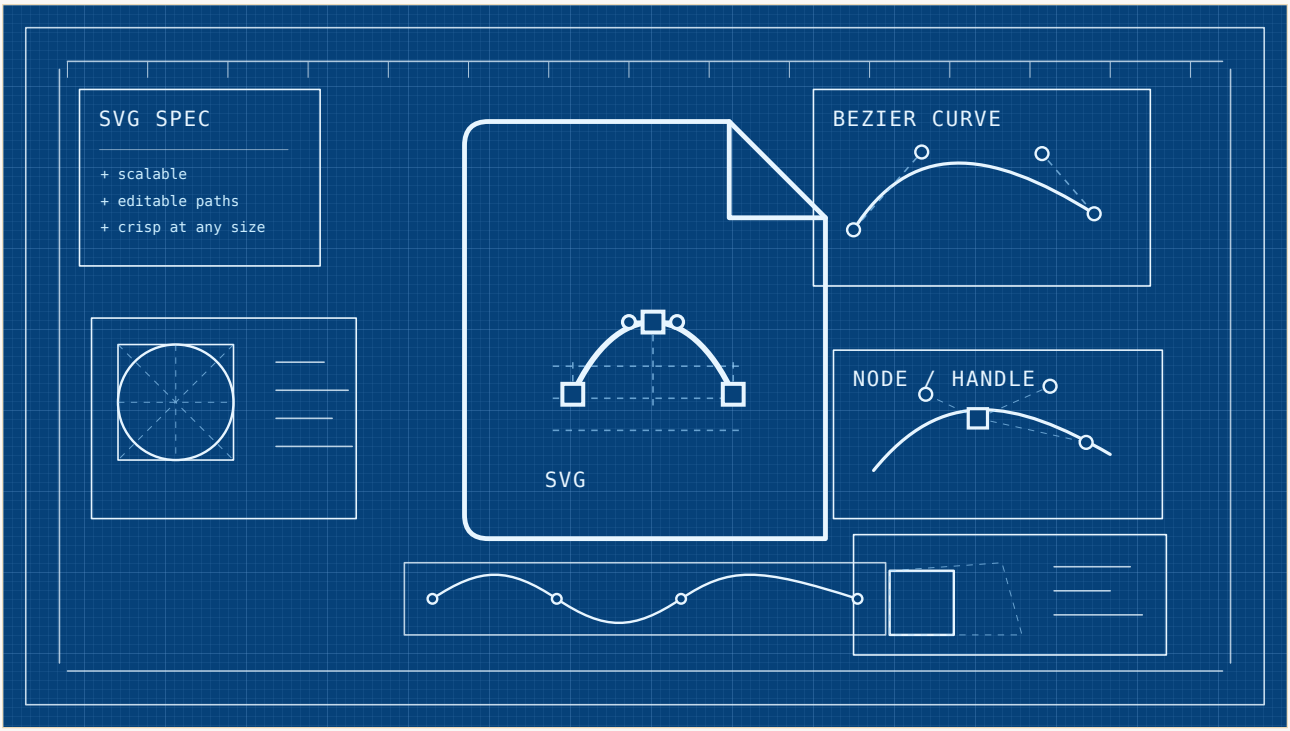
Language models zinazogenerate vector graphics hufanya kazi blind — zinaandika drawing commands bila kuona result. Method mpya inaruhusu modeli kutazama canvas yake ikijaa stroke kwa stroke, na twist ya uaminifu ni kwamba kumpa eyes tu hufanya mambo kuwa mabaya: lazima i-retrainiwe kutumia feedback hiyo. Payoff ni modeli inayolingana au kupita kidogo rivals trained kwenye hadi mara 20 data nyingi zaidi — real, careful result kwenye benchmark moja, si revolution.

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)

Kuchora ukiwa umefumba macho

Ukiomba mojawapo ya AI modeli za leo ikuchoree picha, chini ya hood jambo moja kati ya mawili tofauti sana hutokea. Image generators zinazojulikana — zinazotengeneza picha kutoka sentence — hupaka pixels moja kwa moja, na zinaweza kuona canvas zinapofanya kazi. Lakini kuna aina nyingine, tulivu zaidi ya kuchora, ambapo modeli haipaki pixels kabisa. Inaandika instructions: *chora circle hapa, line pale, jaza shape hii blue*. Instructions hizo ni code — Scalable Vector Graphics, au SVG, ileile iliyo nyuma ya icons na logos nyingi kwenye web — na faida yake ni halisi: matokeo si fixed grid ya pixels bali seti ya shapes unazoweza rescale, recolour na edit bila mwisho bila kuwa blurry.



Original SVG blueprint artwork: image yenyewe ni editable vector file, iliyojengwa kutoka paths, grid lines, nodes na handles badala ya fixed pixels.

Laura Nesso / The Clean Paper Permission on file

Tatizo ni kwamba, hadi hivi karibuni, modeli inayoandika drawing code hii ilikuwa inafanya hivyo kwa "upofu." Ilitoa sequence nzima ya instructions kwa pass moja — circle, line, fill — bila kutengeneza picha hata mara moja kuona ilichotengeneza. Fikiria kuchora uso umefumba macho: unaweza kuweka macho roughly mahali macho yanapaswa kuwa, lakini huwezi kuona kwamba la pili limeangukia shavuni, au shape ya nywele sasa imefunika pua. Ndivyo modeli hizi zilivyofanya kwa kiasi kikubwa, ndiyo maana mara nyingi zilitoa code valid kikamilifu lakini visually mess.

Timu inayoongozwa na Guotao Liang sasa imefanya jambo linaloonekana obvious kwa kuchekesha: imeiruhusu modeli kufungua macho. mbinu yao, *Render-in-the-mzunguko*, hu-render half-finished drawing baada ya kila step na kurudisha picha hiyo kwa modeli kabla ya stroke inayofuata. Sehemu interesting kweli si kwamba hii husaidia — ni kile walicholazimika kufanya ili *isaidie* kabisa.

Unapimaje drawing?

Makala inaposema modeli yake "imeishinda" nyingine, ni fair kuuliza: imeishinda kwenye nini, na imepimwa vipi? Kuhukumu generated image ni vigumu kweli — hakuna answer moja sahihi kwa "chora laptop" — kwa hiyo field hutegemea automatic alama kadhaa, kila moja ikiwa imperfect proxy ya mtu kuangalia matokeo.

Baadhi zinaonekana hapa. **FID** inalinganisha overall statistical flavour ya batch nzima ya generated images dhidi ya real ones; lower ni better, lakini inaeleza batch, si picha moja. **CLIP alama** inauliza AI nyingine kama image inalingana na words za prompt — muhimu, lakini sharpness yake inategemea judge huyo. **DINO**, **SSIM** na **LPIPS** hulinganisha reconstructed image na target, katika levels kutoka raw pixels hadi learned features.

Hakuna mojawapo ni truth. Ni proxies, na mara nyingi huhama kwa increments ndogo. Ukisoma kwamba modeli moja ina 127.6 na nyingine 128.8, hiyo ni real difference katika intended direction — lakini ni nudge, si landslide, na nudge katika number inayofuata kwa loose kile macho yako yangesema. Hilo ni muhimu kukumbuka kichwa cha habari inaposema "beats a modeli trained on twenty times the data."

Walichofanya waandishi

Tatizo la msingi walilotaka kurekebisha ni kuchora "bila kuona." Modeli nyingi zilizopo hutendea uandishi wa SVG kama kazi ya maandishi pekee: hutabiri kipande kinachofuata cha msimbo kutokana na msimbo uliopo bila kuonyesha picha. Hivyo uwezo wa kuona ambao modeli nyingi za kisasa za multimodal tayari zina hauhusishwi katika kazi hiyo. Render-in-the-Loop hubadilisha mchakato kuwa wa kuona hatua kwa hatua. Baada ya kila kipande cha msimbo wa mchoro, SVG ya muda hutengenezwa kama picha na kurudishwa kwa modeli, ili hatua inayofuata ichaguliwe huku modeli ikiangalia hali ya sasa ya turubai.

Matokeo yao ya kwanza ni onyo, na wanayaripoti wazi: kuongeza tu mzunguko huu kwenye modeli ya kawaida iliyopo hakutoshi. Waliporudisha picha za hatua za kati kwa modeli yenye uwezo mkubwa lakini isiyofundishwa mahsusi kwa kazi hii, ubora haukuongezeka—ulipungua. Modeli ambayo haijafundishwa kutumia taarifa ya kuona katika mchakato huu haijifunzi ghafla kufanya hivyo kwa sababu tu imepewa picha.

Kwa hiyo sehemu kubwa ya kazi iko katika mafunzo. Waandishi waliunda upya data ya mafunzo ili kila mchoro ugawanywe katika hatua nyingi ndogo zenye maana ya kuona, kwa kuvunja maumbo tata kuwa sehemu rahisi ambazo zinaongeza kitu kipya katika kila hatua. Kisha waliifanyia fine-tuning modeli wazi yenye vigezo bilioni nane, iliyojengwa juu ya Qwen3-VL, kwa mfululizo huo wa hatua kwa hatua. Wanaita mbinu hii **Visual Self-Feedback**. Pia wanaongeza utaratibu wa pili wakati wa kuchora, **Render-and-Verify**: kabla ya kukubali kila stroke mpya, modeli hui-render na kuangalia kama kweli imebadilisha picha au imerudia tu hatua ya mwisho. Stroke isiyoongeza kitu hutupwa; na inapofika mahali ambapo hakuna nyongeza inayosaidia, modeli huagizwa kusimama. Yote haya yanafanywa kwa hifadhidata ndogo kiasi—karibu mifano 850,000, chini ya nusu ya data ya mshindani mmoja na sehemu ndogo tu ya ya mwingine.

Walichogundua

- Ikiwa imefundishwa kwa njia hii, modeli huchora vizuri zaidi kuliko toleo lake linalochora bila kuona—hasa katika hali ambazo toleo la kawaida hushindwa, kwa mfano kuweka jicho kwenye shavu au kupuuza grafu ya pau iliyoombwa na kutoa monitor ya kawaida badala yake.
- Katika kipimo linganishi cha kawaida (MMSVGBench), mbinu inashindana vizuri na, kwa baadhi ya vipimo, iko mbele kidogo ya washindani wenye nguvu—ikiwemo OmniSVG, iliyofundishwa kwa zaidi ya mara mbili ya data, na InternSVG, iliyofundishwa kwa karibu mara 20 ya data.
- Tofauti ni ndogo. Katika seti ya ikoni, kwa mfano, alama kuu ya ubora wa picha ni 127.6 dhidi ya 128.8 ya mshindani bora, na alama ya kufuata prompt ni 0.293 dhidi ya 0.291. Tofauti hizo ni halisi na zinaelekea upande uleule, lakini si kubwa.
- Vipengele vyote viwili vilivyoongezwa vina mchango unaoweza kupimwa: ukiondoa mafunzo maalum au uthibitishaji wakati wa kuchora, alama hushuka. Hatua ya uthibitishaji hasa huzuia modeli kukwama ikichora kitu kilekile tena na tena.
- Jambo ambalo waandishi wenyewe wanasisitiza ni ufanisi: kufikia kiwango hiki kwa data ya mafunzo iliyo kidogo sana kuliko ile iliyotumiwa na mifumo inayoongoza.

Hili halithibitishi nini

- **Halionyeshi kwamba kuiruhusu modeli "ione" ni free win.** Kinyume chake: experiment ya makala yenyewe inaonyesha visual feedback *bila retraining* hufanya mambo kuwa mabaya. Gain inatoka retraining, si eyes pekee.
- **Halithibitishi lead kubwa au decisive.** Kwenye alama nyingi mbinu iko neck-and-neck na rivals; "beats a modeli trained on 20× the data" ni true, lakini kwa nudges kwenye vipimo vya mbadala, kwenye kipimo linganishi moja.
- **Halionyeshi uwezo wa jumla wa kisanii.** Hii ni modeli ya utafiti yenye vigezo bilioni 8 inayochora ikoni na michoro rahisi katika azimio dogo lililowekwa (pikseli 224×224), si mbunifu wa matumizi ya jumla.
- ulinganisho dhidi ya big general modeli (GPT-5) **si apples-to-apples**: modeli hiyo haijajengwa au tuned kwa narrow code-drawing task hii, kwa hiyo kuishinda hapa haisemi mengi kuhusu modeli hizo kwa ujumla.
- **Haiji bure.** Rendering na re-reading canvas kila step hufanya generation kuwa slower kuliko kutoa code yote katika blind pass moja — gharama ambayo waandishi wanakubali.

Ushahidi una nguvu kiasi gani?

- **Imara** ambapo ni ulinganisho uliodhibitiwa kwa terms zake. Ablations ni clean: ondoa training, au ondoa verification, numbers zinashuka — kwa hiyo ingredients hizo mbili kweli zinafanya kazi waandishi wanazodai.
- **Ina uaminifu kuhusu surprise yake.** matokeo kwamba naive visual feedback *inaumiza* quality

imeripotiwa, si kufichwa, na ndiyo sehemu interesting zaidi — correction muhimu kwa intuition kwamba more input daima ni better.
- **Nyembamba kwenye leaderboard dai.** Wins dhidi ya rivals wenye data kubwa ni small na ziko kwenye kipimo linganishi moja iliyojengwa na authors wa mmoja wa rivals hao. Small margins kwenye vipimo vya mbadala kwa one seti ya majaribio ni suggestive, si settled.
- **Haijajaribiwa at scale na in the wild.** Yote hapa ni icons na simple illustrations kwa low resolution. Kama idea hiyo hiyo inashikilia kwa complex, high-resolution au katika ulimwengu halisi design kazi imeachwa future kazi — waandishi wanasema hivyo.

Kwa nini ni muhimu

Idea ya katikati ya makala hii ni simple kiasi cha kuaibisha, na inaenda mbali zaidi ya drawing. Kama program itagenerate kitu kwa kuandika code — web page, chart, diagram, 3D scene — inaweza kuandika yote blind na kutumaini, au render inapoendelea na kusahihisha direction. Closing that mzunguko ni second nature kwa binadamu; tunaangalia page mara kwa mara. Ni move ya hivi karibuni kwa modeli hizi.

Kinachoifanya makala hii ivutie ni tahadhari yake. Kumpa modeli uwezo wa kuona si sawa na kuifundisha *kutazama*. Uwezo wa kuona lazima ufundishwe, ndipo mzunguko wa kuona na kurekebisha unaposaidia — na hata hivyo faida ni ya kiasi: mfumo unakuwa mchora picha mwenye uthabiti zaidi, si msanii wa aina mpya. Kwa zana zinazogeuza maelezo kuwa picha zinazoweza kuhaririwa, kama ikoni na michoro ya programu, somo la vitendo ni kwamba mafunzo yaliyobuniwa vizuri yanaweza kuwa muhimu kuliko kuongeza data bila mwisho.

Muhtasari safi

lugha modeli zinazogenerate vector graphics — editable, rescalable code iliyo nyuma ya web icons na logos nyingi — traditionally zimefanya hivyo “blind,” zikiandika drawing commands zote bila kuona matokeo. Timu inayoongozwa na Guotao Liang inapendekeza Render-in-the-mzunguko: render half-finished drawing baada ya kila step na irudishe kwa modeli, ili stroke inayofuata ichaguliwe modeli ikiwa inaangalia canvas. Central honest matokeo yao ni kwamba kufanya hivi tu kwa existing modeli kunaifanya *mbaya zaidi*; gain inaonekana baada ya retraining modeli kutumia visual feedback, ikisaidiwa na drawing-time check inayotupa strokes ambazo hazibadilishi kitu. Retrained 8B modeli inalingana au slightly beats rivals trained kwa hadi mara 20 data nyingi zaidi kwenye kipimo linganishi cha kawaida — genuine matokeo, lakini kwa small margins kwenye proxy alama, kwa icons na simple illustrations katika low resolution. Takeaway ambayo waandishi wanasisitiza ni efficiency: kuona kazi yako vizuri kunaweza kusimama badala ya sheer data scale.

Ukaguzi bila kupamba

Makala inaonyesha nini: Retraining vector-graphics modeli kutengeneza picha na kuangalia half-finished drawing yake, hatua kwa hatua, inatoa matokeo bora na complete zaidi kuliko drawing blind — competitive na slightly ahead ya larger-data rivals kwenye kipimo linganishi cha kawaida moja, kutoka training data kidogo.

Kinachowezekana lakini hakijathibitishwa: Kwamba "seeing your kazi" ni recipe bora kwa ujumla kuliko scaling data; kwamba idea hiyo itasaidia complex, high-resolution au katika ulimwengu halisi graphics; kwamba small kipimo linganishi margins zinawakilisha difference ambayo mtu angeona kweli.

Isichoonyeshwa: Kwamba visual feedback husaidia peke yake (bila retraining inaumiza); kwamba lead dhidi ya rivals ni large au decisive; general-purpose drawing ability; fair ulinganisho wa moja kwa moja na general modeli kama GPT-5 ambazo hazijajengwa kwa task hii.

Vikwazo vikuu: kipimo linganishi moja, partly built na authors wa rival; small margins kwenye vipimo vya mbadala; 8B modeli, icons na simple illustrations kwa 224×224; slower generation kwa sababu ya render-every-step mzunguko.

Msomaji wa kawaida anapaswa kuwa na uhakika kiasi gani? Uhakika mkubwa kwamba closing the mzunguko — rendering na re-reading unapochora — husaidia kweli, na lazima ifundishwe, si kuwashwa tu. Uhakika wa low-to-moderate kwamba particular modeli hii decisively beats rivals; "beats 20× the data" ichukuliwe kama real lakini modest single-kipimo linganishi matokeo. Uhakika mkubwa kwenye idea moja ya kukumbuka: kwa code inayochora, kuangalia unapoendelea ni bora kuliko drawing blind.

Vyanzo

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/