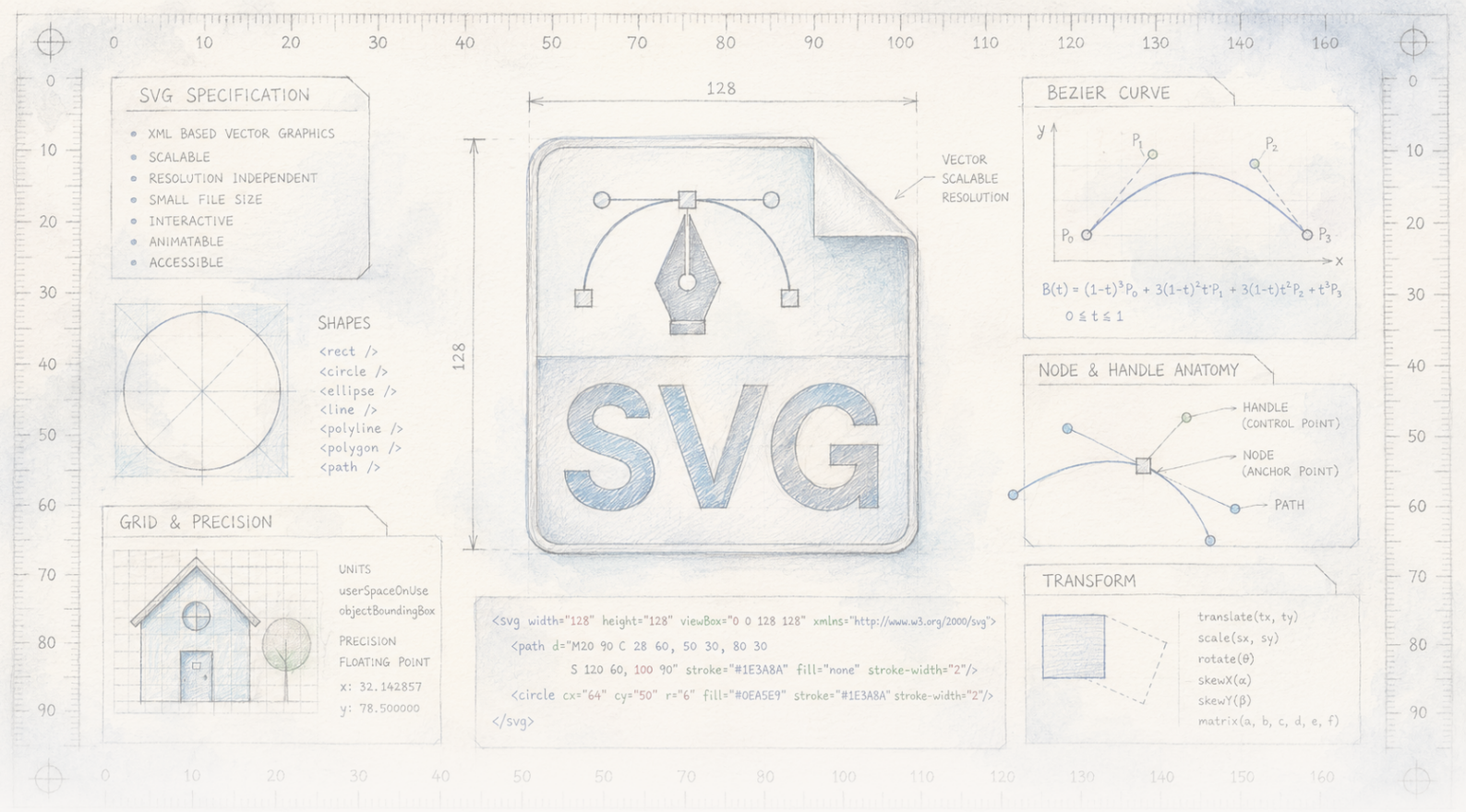


The Clean Paper

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

Naučiť kresliacu AI sledovať vlastné plátno

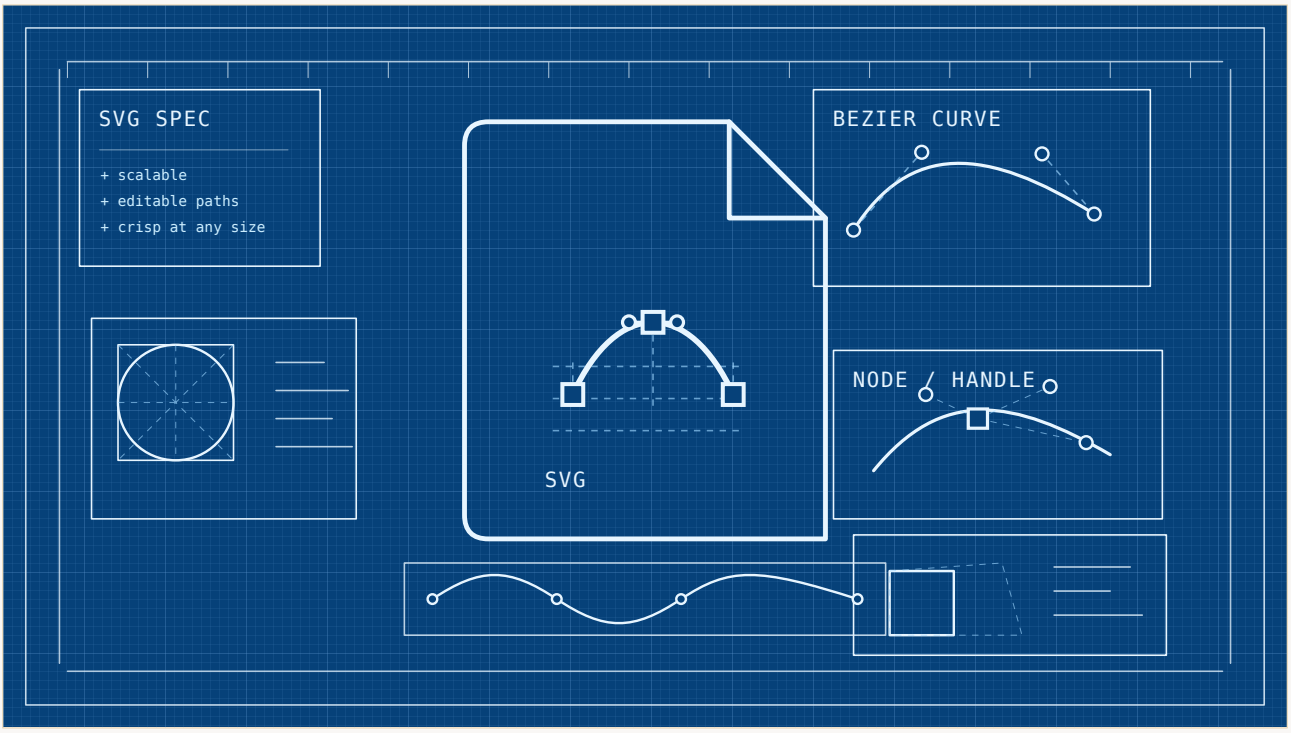
Jazykové modely, ktoré vytvárajú vektorovú grafiku, tradične kreslia naslepo — zapisujú príkazy bez toho, aby videli výsledok. Nová metóda umožňuje modelu sledovať, ako sa jeho vlastné plátno dopĺňa ťah po ťahu. Poctivým zvratom je, že samotné pridanie zraku výkon zhoršuje: model treba znovu natrénovať, aby vizuálnu spätnú väzbu vedel používať. Výsledný model sa vyrovnáva konkurentom trénovaným až na dvadsaťnásobnom množstve dát alebo ich mierne prekonáva — ide o skutočný, no skromný výsledok na jedinom benchmarku, nie o revolúciu.

Written by Vera Rubin · Cross-checked by Michele Renda · Edited by Michele Renda · Figures by Laura Nesso · 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)

Kreslenie so zatvorenými očami

Požiadajte niektorý z dnešných modelov AI, aby nakreslil obrázok, a v jeho vnútri sa odohrá jedna z dvoch úplne odlišných vecí. Známe generátory obrázkov — tie, ktoré z vety vytvoria fotografiu — maľujú priamo pixely a počas práce môžu vidieť plátno. Existuje však aj druhý, menej nápadný druh kreslenia, pri ktorom model vôbec nemaľuje. Píše pokyny: *nakresli tu kruh, tam čiaru, vyplň tento tvar modrou*. Pokyny tvoria kód — škálovateľnú vektorovú grafiku, teda SVG, ktorá stojí za väčšinou ikon a log na webe. Výhoda je zrejmá: výsledkom nie je pevná mriežka pixelov, ale súbor tvarov, ktoré možno ľubovoľne zväčšovať, prefarbovať a upravovať bez rozmazania.



Pôvodný technický nákres v SVG: obrázok je sám o sebe editovateľný vektorový súbor, vytvorený z ciest, mriežkových línií, uzlov a rukovätí, nie z pevných pixelov.

Laura Nesso / The Clean Paper Permission on file

Háčik spočíva v tom, že model píšuci takýto kresliaci kód donedávna pracoval naslepo. Naraz vygeneroval celú postupnosť pokynov — kruh, čiaru, výplň — bez priebežného vykreslenia výsledku. Predstavte si kreslenie tváre so zatvorenými očami: oči by možno skončili približne na správnom mieste, ale nemohli by ste si všimnúť, že druhé oko je na líci alebo že tvar vlasov prekryl nos. Modely fungovali približne takto, preto často vytvárali úplne platný, no vizuálne chaotický kód.

Tím vedený Guotao Liangom teraz urobil takmer komicky zrejmú vec: nechal model otvoriť oči. Ich metóda, *Render-in-the-Loop*, vykreslí nedokončený obrázok po každom kroku a odovzdá obrázok späť modelu pred ďalším ťahom. Naozaj zaujímavé nie je, že to pomáha — ale to, čo museli urobiť, aby to vôbec pomohlo.

Ako sa ohodnotí kresba?

Keď práca tvrdí, že ich model „prekonáva“ iný, je spravodlivé sa opýtať: v čom a podľa akého meradla? Posudzovať generovaný obrázok je naozaj ťažké — neexistuje jediná správna odpoveď na „nakresli notebook“ — takže odbor sa opiera o niekoľko automatických skóre, z ktorých každé je nedokonalým zástupcom osoby, ktorá sa pozerá na výsledok.
Niekoľko z nich sa objavuje aj v tejto práci. **FID** porovnáva celkový štatistický vzhľad dávky generovaných obrázkov s reálnymi; nižšia hodnota je lepšia, ale opisuje súbor, nie jediný obrázok. **Skóre CLIP** sa pýta samostatnej AI, či obrázok zodpovedá slovám zadania — je užitočné, ale iba také presné ako samotný hodnotiteľ. **DINO**, **SSIM** a **LPIPS** porovnávajú rekonštruovaný obraz s cieľom na úrovniach od surových pixelov po naučené príznaky.
Žiadna z týchto metrík nie je úplnou pravdou. Sú to zástupné ukazovatele a majú tendenciu pohybovať sa v malých krokoch. Keď si prečítate, že jeden model dosahuje 127,6 a iný 128,8, je to skutočný rozdiel v zamýšľanom smere — ale ide však len o mierny, nie drvivý posun v čísle, ktoré len voľne zodpovedá tomu, čo by povedalo vaše oko. Stojí za to si ho ponechať, keď titulok znie “prekonáva model trénovaný na dvadsaťnásobku dát.”

Čo autori urobili

Problém, ktorý sa autori snažia vyriešiť, je samotné kreslenie naslepo. Existujúce modely považujú písanie SVG za čisto textovú úlohu: predpovedať ďalší úsek kódu z doterajšieho kódu a nikdy ho nevykresliť. To necháva nečinné mocné “oči” — obrazový enkodér — ktorý už moderné multimodálne modely majú. Render-in-the-Loop reštrukturalizuje prácu ako vizuálny proces krok za krokom. Po každom fragmente výkresového kódu sa čiastočný SVG vykreslí do obrazu a vráti modelu ako obrázok, takže ďalší fragment sa vyberie pri skutočnom pohľade na doterajšie plátno.

Ich prvé zistenie je varovné a treba im uznať, že to jasne uvádzajú: jednoducho pripevniť tento cyklus na existujúci hotový model nefunguje. Keď podávali priebežné rendery silným všeobecným modelom bez špeciálneho tréningu, kvalita sa nezlepšila — *zhoršila* sa naprieč celým hodnotením. Model, ktorý nikdy nebol naučený používať zrak na tento účel, zrazu nevie ako.

Takže väčšina práce spočíva v trénovaní. Autori upravili trénovacie dáta tak, že každý výkres je rozdelený na mnoho malých, vizuálne zmysluplných krokov — rozdeľujú zložité tvary na jednoduchšie časti, aby bolo v každej fáze niečo nové na objavenie — a potom doladia otvorený model s osemmiliardami parametrov (postavený na Qwen3-VL) na týchto sekvenciách jednotlivých krokov. Toto nazývajú vizuálna spätná väzba na vlastný výstup. Pri kreslení pridávajú druhý mechanizmus, Render-and-Verify: pred prijatím každého nového ťahu ho model vykreslí a skontroluje, či skutočne zmenil obrázok, alebo len zopakoval ten posledný. Ťahy, ktoré nič nepridávajú, sa vyhadzujú, a keď už nič viac nepomáha, model je vyzvaný, aby sa zastavil. Pozoruhodné je, že všetko toto beží na pomerne malom dátovom súbore — približne 850 000 príkladov, menej ako polovica toho, čo používal jeden konkurent, a malý zlomok iného.

Čo našli

- Pri trénovaní týmto spôsobom model kreslí lepšie ako jeho slepý náprotivok — najviditeľnejšie v prípadoch zlyhania, keď slepý model priloží oko na líce alebo preskočí požadovaný stĺpcový graf a namiesto toho vyprodukuje generický monitor.
- V štandardnom benchmarku (MMSVGBench) je metóda konkurencieschopná a podľa niekoľkých ukazovateľov mierne predstihuje silných konkurentov — vrátane OmniSVG, trénovaného na viac než dvojnásobnom množstve dát, a InternSVG, trénovaného na približne dvadsaťnásobnom množstve dát.
- Rozdiely sú malé. V súbore ikon má napríklad hlavné skóre kvality obrazu 127,6 oproti 128,8 pri najlepšom súperovi a skóre zhody so zadaním 0,293 oproti 0,291 — skutočné a konzistentné, ale mierne rozdiely.
- Obe pridané súčasti si zaslúžia svoje miesto: vypnú špeciálny tréning alebo overovanie počas kreslenia a čísla merateľne klesajú. Najmä overovací krok zabraňuje tomu, aby sa model zasekol v opakovaní toho istého stále dokola.
- Výsledok, ktorý autori sami zdôrazňujú, je efektivita — dosiahnuť takýto výsledok z oveľa menšieho množstva trénovacích dát, než koľko použili lídri.

Čo to nedokazuje

- Neukazuje, že nechať model „vidieť“ prináša zlepšenie zadarmo. Práve naopak: experiment samotného článku ukazuje, že vizuálna spätná väzba *bez* preškolenia situáciu zhoršuje. Prínos pochádza z nového trénovania, nie len z očí.
- Nevytvára veľký alebo rozhodujúci náskok. Vo väčšine prípadov je metóda vyrovnaná so svojimi konkurentmi; “Prekonáva model trénovaný na 20× dát” je pravda, ale len na základe malého rozdielu v zástupných metrikách, na jednom benchmarku.
- Nepreukazuje všeobecnú umeleckú schopnosť. Ide o výskumný model s osem miliardami parametrov, ktorý kreslí ikony a jednoduché ilustrácie v malom pevnom rozlíšení (224×224 pixelov), nie o všeobecného dizajnéra.
- Porovnanie s veľkým všeobecným modelom (GPT-5) nie je porovnaním **rovnakého s rovnakým**: tento model nie je postavený ani naladený na túto úzku úlohu kreslenia kódu, takže jeho prekonanie tu veľa nehovorí o žiadnom z modelov všeobecne.
- **Nie je bez nákladov.** Vykresľovanie a opätovné čítanie plátna v každom kroku spomaľuje generovanie oproti jednorazovému vytvoreniu kódu naslepo, čo autori výslovne uznávajú.

Aké silné sú dôkazy

- **Pevné**, kde ide o kontrolované porovnanie na vlastných podmienkach. Ablácie sú čisté: ak odstránite tréning, alebo overenie, čísla klesnú — takže tieto dve zložky naozaj robia to, čo autori tvrdia.

- **Úprimné voči vlastnému prekvapeniu.** Zistenie, že naivná vizuálna spätná väzba *škodí*, je uvedené, nie ukryté, a je to najzaujímavejšia vec v článku — užitočná oprava intuície, že viac vstupov je vždy lepšie.
- **Slabšie tam, kde je to tvrdenie o rebríčku.** Víťazstvá nad rivalmi s väčšími dátami sú malé a pochádzajú z jediného benchmarku, ktorý vytvorili autori jedného z týchto rivalov. Malé rozdiely v proxy metrikách na jednom testovacom sete sú skôr náznakom než rozhodnutím.
- **Netestované vo veľkom rozsahu a v reálnom prostredí.** Všetko tu sú ikony a jednoduché ilustrácie v nízkom rozlíšení. Či už rovnaká myšlienka platí pre zložité, vysoko rozlíšené alebo reálne dizajnové práce, zostáva otázkou pre budúci výskum — autori to jasne hovoria.

Prečo je to dôležité

Myšlienka v centre tejto práce je takmer trápne jednoduchá a siaha ďaleko za hranice kreslenia. Ak má program niečo vygenerovať písaním kódu — webovú stránku, graf, diagram, 3D scénu — môže buď napísať celý proces naslepo a dúfať, alebo vykresľovať za pochodu a upraviť smer. Uzavrieť túto slučku je pre človeka samozrejmosťou; stále sa pozeráme na stránku. Je to prekvapivo nedávny krok pre tieto modely.

To, čo robí článok hodným čítania, je výhrada, ktorú pripája. Dať modelu spôsob, ako vidieť, nie je to isté ako naučiť ho pozerať sa. Oči treba trénovať, a až potom sa slučka vypláca — a aj vtedy je odmena skutočná, ale vyvážená: istejší kresliar, nie iný typ umelca. Pre každého, kto vytvára nástroje, ktoré premenia popis na upraviteľnú grafiku — niečo, čo skončí za ikonami a ilustráciami aplikácie — je užitočná tichá, praktická lekcia. Víťazstvo tu prišlo z úspornejšieho a inteligentnejšieho trénovania, nie z väčšieho množstva dát. To je užitočnejšie ponaučenie než ďalší bod v rebríčku.

Čisté zhrnutie

Jazykové modely, ktoré generujú vektorovú grafiku — upraviteľný, škálovateľný kód tvoriaci väčšinu webových ikon a log — tradične pracovali „naslepo“: zapísali všetky príkazy na kreslenie bez toho, aby si výsledok priebežne vykreslili. Tím vedený Guotao Liangom navrhuje metódu Render-in-the-Loop: po každom kroku vykreslí rozpracovaný obrázok a vráti ho modelu, ktorý potom kreslí ďalší ťah pri pohľade na plátno. Ústredným a poctivým zistením je, že samotné pridanie tohto správania existujúci model *zhorší*. Zlepšenie prichádza až po novom natrénovaní na vizuálnu spätnú väzbu, podporenom kontrolou počas kreslenia, ktorá zahadzuje ťahy bez viditeľnej zmeny. Pretrénovaný model s ôsmimi miliardami parametrov sa v štandardnom benchmarku vyrovná konkurentom trénovaným na až dvadsaťnásobne väčšom objeme dát alebo ich mierne prekoná — je to skutočný výsledok, ale s malým rozdielom v zástupných metrikách pri ikonách a jednoduchých ilustráciách s nízkym rozlíšením. Zapamätania hodný záver autorov sa týka efektívnosti: pozorné sledovanie vlastnej práce môže nahradiť časť obrovského množstva dát.

Kontrola bez nezmyslov

Čo ukazuje práca: Pretrénovanie vektorovo-grafického modelu na vykreslenie a sledovanie vlastného nedokončeného výkresu, krok za krokom, prináša lepšie a úplnejšie výsledky než kreslenie naslepo — konkurencieschopné a mierne lepšie než výsledky súperov s väčším množstvom dát v jednom štandardnom benchmarku, a to vďaka menšiemu množstvu trénovacích dát.

Čo je pravdepodobné, ale nie dokázané: Že "vidieť svoju prácu" je všeobecne lepší recept než škálovanie dát; že rovnaká myšlienka pomôže pri zložitej, vysoko rozlíšenej alebo reálnej grafike; že malé rozdiely v benchmarku odrážajú rozdiel, ktorý by človek skutočne zaznamenal.

Čo neukazuje: Že vizuálna spätná väzba pomáha sama o sebe (bez nového trénovania škodí); že náskok pred súpermi je veľký alebo rozhodujúci; schopnosť všeobecného kreslenia; férový súboj s bežnými modelmi ako GPT-5, ktoré na túto úlohu nie sú stavané.

Hlavné obmedzenia: Jediný benchmark, čiastočne vytvorený autormi konkurenčného systému; malé rozdiely v zástupných metrikách; model s 8 miliardami parametrov, ikony a jednoduché ilustrácie s rozlíšením 224×224; pomalšie generovanie pre vykresľovanie v každom kroku.

Koľko dôvery by mal mať bežný čitateľ? Vysokú v to, že uzatváranie slučky — renderovanie a opätovné čítanie počas kreslenia — naozaj pomáha, a že to musí byť trénované, nie len zapnuté. Nízku až strednú v to, že tento konkrétny model rozhodne poráža svojich konkurentov; považujte vetu „prekonáva model s 20-násobkom dát" ako skutočný, ale skromný výsledok v rámci jedného benchmarku. Vysokú v jednu myšlienku, ktorú stojí za to si zapamätať: pri kóde, ktorý kreslí, je lepšie pozerať sa ako kresliť naslepo.

Zdroje

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/