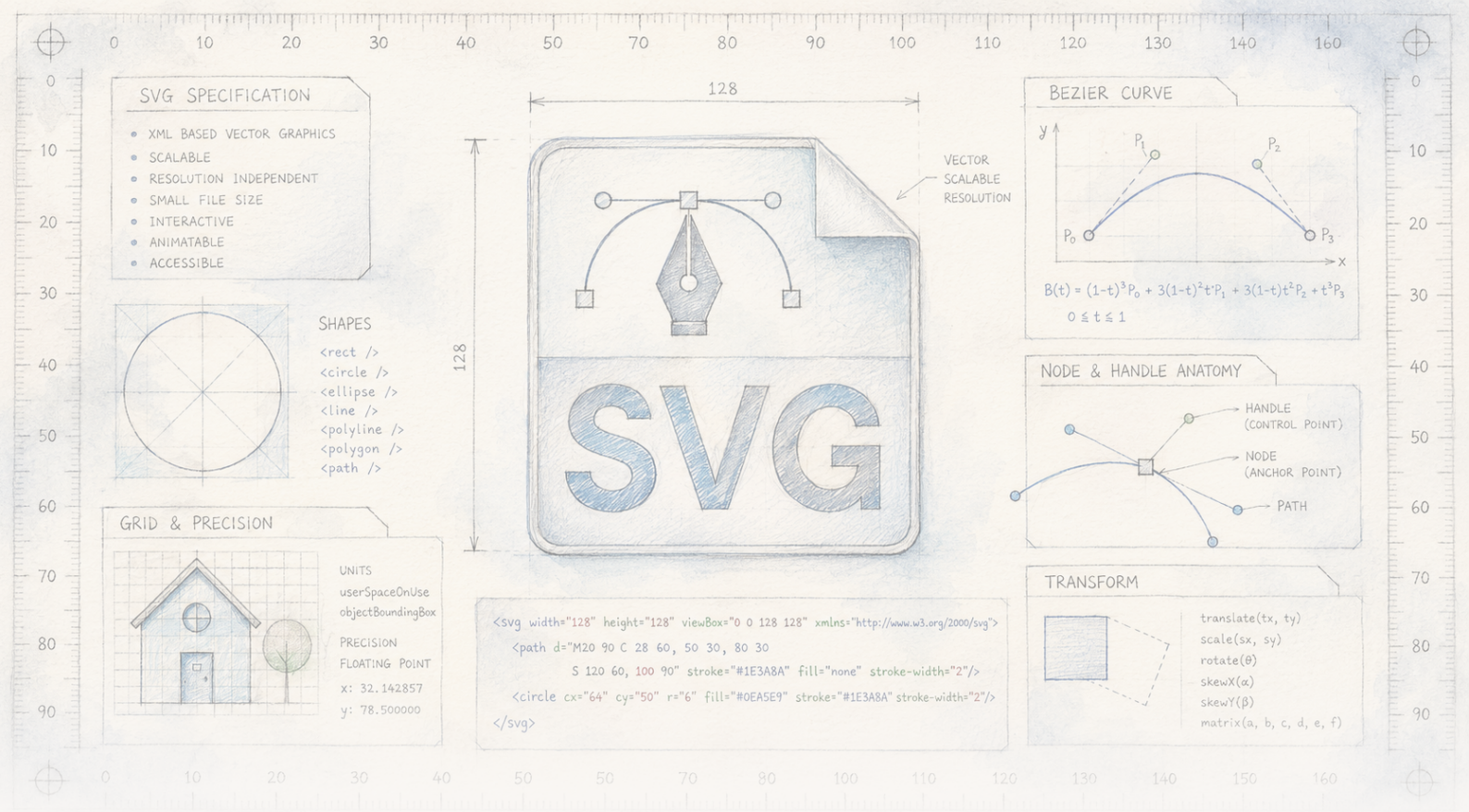


The Clean Paper

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

Mengajar AI yang melukis untuk melihat halaman

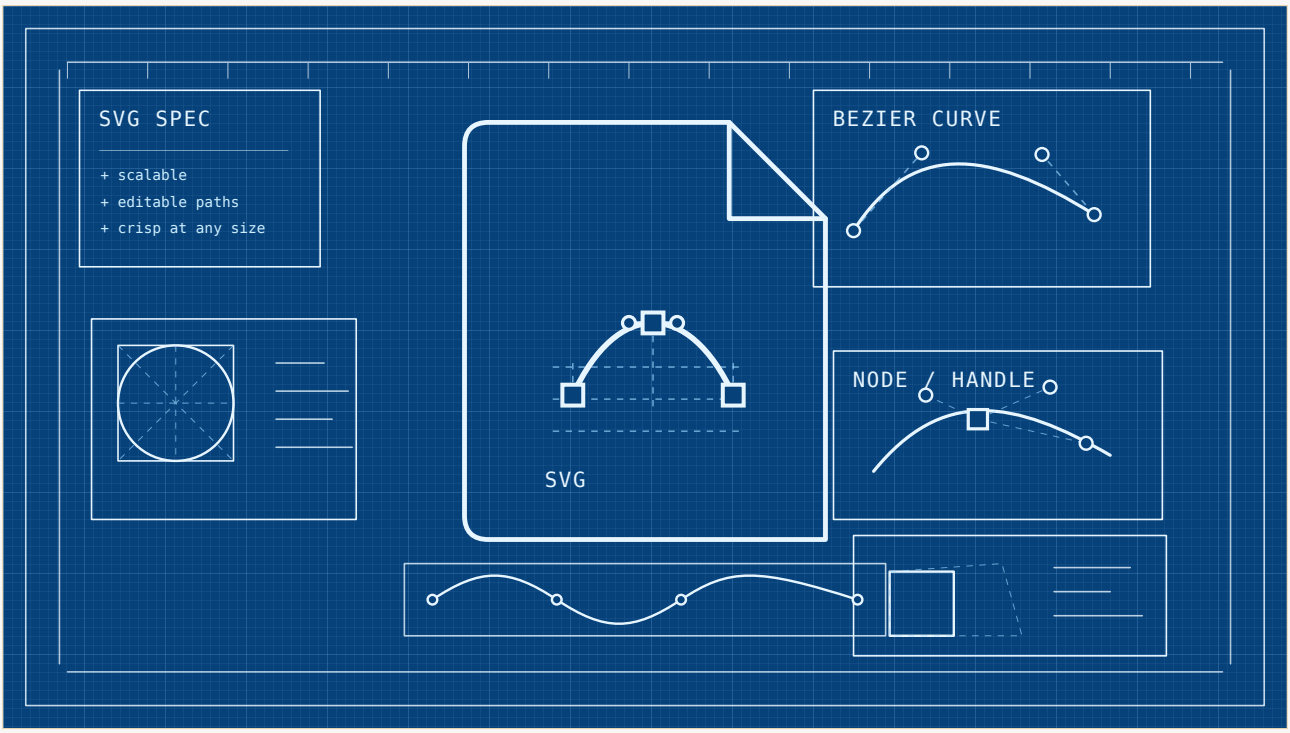
Model bahasa yang menjana grafik vektor melakukannya secara buta — menulis arahan lukisan tanpa pernah melihat hasilnya. Kaedah baharu membolehkan model melihat kanvasnya sendiri terbentuk, goresan demi goresan, dan kelainan yang jujur ialah sekadar memberinya mata sebenarnya memburukkan hasil: ia perlu dilatih semula untuk menggunakan penglihatan itu. Hasilnya ialah model yang menyamai atau sedikit mengatasi pesaing yang dilatih dengan data sehingga dua puluh kali lebih banyak — hasil nyata dan berhati-hati pada satu penanda aras, bukan satu revolusi.

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)

Melukis dengan mata tertutup

Minta salah satu model AI hari ini melukis sebuah gambar untuk anda, dan di sebalik tabir, salah satu daripada dua perkara yang sangat berbeza akan berlaku. Penjana imej yang biasa kita kenali — yang boleh menghasilkan foto daripada satu ayat — melukis piksel secara langsung, dan ia boleh melihat kanvas ketika bekerja. Tetapi ada satu lagi jenis lukisan yang lebih senyap, di mana model langsung tidak melukis. Ia menulis arahan: *lukis bulatan di sini, garisan di sana, warnakan bentuk ini biru*. Arahan itu ialah kod — Scalable Vector Graphics, atau SVG, yang sama seperti yang digunakan di sebalik kebanyakan ikon dan logo di web — dan kelebihannya memang nyata: hasilnya bukan grid piksel tetap, tetapi sekumpulan bentuk yang boleh dibesarkan, dikecilkan, ditukar warna dan diedit tanpa had tanpa menjadi kabur.



Karya pelan teknikal SVG asli: imej itu sendiri ialah fail vektor yang boleh diedit, dibina daripada laluan, garisan grid, nod dan pemegang, bukannya piksel tetap.

Laura Nesso / The Clean Paper Permission on file

Masalahnya ialah, sehingga baru-baru ini, model yang menulis kod lukisan ini melakukannya secara buta. Ia mengeluarkan seluruh urutan arahan dalam satu laluan — bulatan, garisan, isian — tanpa pernah merendernya untuk melihat apa yang telah dihasilkan. Bayangkan melakar wajah dengan mata tertutup: anda mungkin meletakkan mata lebih kurang di tempat yang betul, tetapi tiada cara untuk menyedari bahawa mata kedua jatuh di pipi, atau bentuk rambut yang anda lukis kini menutupi hidung. Lebih kurang begitulah cara model-model ini bekerja, sebab itu ia kerap menghasilkan kod yang sah sepenuhnya tetapi rupa visualnya bersepah.

Pasukan yang diketuai Guotao Liang kini melakukan sesuatu yang hampir terlalu jelas sehingga lucu: mereka membenarkan model membuka matanya. Kaedah mereka, *Render-in-the-Loop*, merender

lukisan yang separuh siap selepas setiap langkah dan memberikan imej itu semula kepada model sebelum ia membuat goresan seterusnya. Bahagian yang benar-benar menarik bukanlah bahawa ini membantu — tetapi apa yang perlu mereka lakukan supaya ia membantu sama sekali.

Bagaimana anda menilai sebuah lukisan?

Apabila sebuah makalah mengatakan modelnya "mengalahkan" model lain, wajar untuk bertanya: mengalahkannya dalam apa, dan diukur bagaimana? Menilai gambar yang dijana memang sukar — tiada satu jawapan betul untuk arahan "lukis sebuah komputer riba" — jadi bidang ini bergantung pada beberapa skor automatik, setiap satunya hanyalah pengganti yang tidak sempurna untuk manusia yang melihat hasilnya.

Beberapa daripadanya muncul dalam makalah ini. **FID** membandingkan corak statistik keseluruhan bagi satu kelompok imej yang dijana dengan imej sebenar; lebih rendah lebih baik, tetapi ia menerangkan kelompok itu, bukan mana-mana satu gambar. **Skor CLIP** meminta AI berasingan menilai sama ada imej sepadan dengan kata-kata dalam arahan — berguna, tetapi ketepatannya bergantung pada ketajaman penilai itu. **DINO**, **SSIM** dan **LPIPS** membandingkan imej yang dibina semula dengan sasaran, pada tahap daripada piksel mentah hingga ciri yang dipelajari.

Tiada satu pun daripada skor ini ialah kebenaran mutlak. Ia ialah proksi, dan biasanya berubah dalam langkah yang kecil. Apabila anda membaca bahawa satu model mendapat 127.6 manakala yang lain mendapat 128.8, itu memang perbezaan sebenar ke arah yang dikehendaki — tetapi ia cuma sedikit kelebihan, bukan kemenangan besar, dan pada nombor yang hanya berkait secara longgar dengan apa yang akan dinilai oleh mata manusia. Perkara ini patut diingat apabila tajuknya berbunyi "mengalahkan model yang dilatih dengan data dua puluh kali ganda".

Apa yang dilakukan oleh para penulis

Masalah yang cuba dibaiki oleh para penulis ialah lukisan secara buta itu sendiri. Model sedia ada menganggap penulisan SVG sebagai tugas teks semata-mata: ramalkan cebisan kod seterusnya daripada kod yang sudah ada, dan jangan sekali-kali merendernya. Ini membiarkan "mata" yang berkuasa — pengekod penglihatan — yang sudah dimiliki model multimodal moden tidak digunakan. Render-in-the-Loop menyusun semula tugas itu sebagai proses visual langkah demi langkah. Selepas setiap cebisan kod lukisan, SVG separa dirender menjadi imej dan dimasukkan semula kepada model sebagai gambar, supaya cebisan berikutnya dipilih sambil benar-benar melihat keadaan kanvas setakat itu.

Penemuan pertama mereka bersifat peringatan, dan mereka patut diberi kredit kerana melaporkannya dengan terus terang: sekadar memasang gelung ini pada model sedia ada tidak berkesan. Apabila mereka memberikan render perantaraan kepada model umum yang kuat tanpa latihan khusus, kualiti tidak bertambah baik — ia *merosot* secara menyeluruh. Model yang tidak pernah diajar menggunakan matanya untuk tugas ini tidak tiba-tiba tahu caranya.

Jadi, sebahagian besar kerja sebenar terletak pada latihan. Mereka membina semula data latihan supaya setiap lukisan dipecahkan kepada banyak langkah kecil yang bermakna secara visual — membahagikan bentuk kompleks kepada bahagian yang lebih ringkas supaya ada sesuatu yang baharu untuk dilihat pada setiap peringkat — kemudian menala halus model terbuka lapan bilion parameter (berasaskan Qwen3-VL) pada urutan langkah demi langkah ini. Mereka menamakannya Visual Self-Feedback. Mereka turut menambah mekanisme kedua ketika melukis, Render-and-Verify: sebelum menerima setiap goresan baharu, model merendernya dan memeriksa sama ada ia benar-benar mengubah gambar, atau sekadar mengulangi yang terakhir. Goresan yang tidak menambah apa-apa dibuang, dan apabila tiada lagi perubahan yang membantu, model diarahkan untuk berhenti. Yang penting, semua ini berjalan pada set data yang agak kecil — kira-kira 850,000 contoh, kurang daripada separuh jumlah yang digunakan oleh satu pesaing dan hanya sebahagian kecil daripada jumlah pesaing lain.

Apa yang mereka temui

- Setelah dilatih dengan cara ini, model melukis lebih baik daripada versi butanya — paling jelas pada kes kegagalan, apabila model buta meletakkan mata di pipi, atau mengabaikan carta bar yang diminta lalu menghasilkan monitor biasa sahaja.
- Pada penanda aras piawai (MMSVGBench), kaedah ini bersaing rapat dengan, dan pada beberapa ukuran sedikit mendahului, pesaing yang kuat — termasuk OmniSVG, yang dilatih dengan lebih daripada dua kali ganda data, dan InternSVG, yang dilatih dengan kira-kira dua puluh kali ganda data.
- Marginnya kecil. Pada set ikon, misalnya, skor utama kualiti imejnya ialah 127.6 berbanding 128.8 untuk pesaing terbaik, dan skor padanan arahan 0.293 berbanding 0.291 — nyata dan konsisten, tetapi tipis.
- Kedua-dua komponen tambahan memang menyumbang: apabila latihan khas atau pengesahan semasa melukis dimatikan, skornya turun dengan ketara. Langkah pengesahan khususnya menghalang model daripada terperangkap melukis semula perkara yang sama berulang kali.
- Hasil yang paling ditekankan oleh para penulis sendiri ialah kecekapan — mencapai tahap ini dengan data latihan yang jauh lebih sedikit daripada yang digunakan oleh model-model teratas.

Apa yang tidak dibuktikan oleh kajian ini

- Ia **tidak** menunjukkan bahawa membenarkan model “melihat” ialah kemenangan percuma. Malah sebaliknya: eksperimen makalah ini sendiri menunjukkan bahawa maklum balas visual *tanpa* latihan semula menjadikan hasil lebih buruk. Peningkatan datang daripada latihan semula, bukan daripada mata semata-mata.
- Ia **tidak** membuktikan kelebihan yang besar atau muktamad. Pada kebanyakan skor, kaedah ini hampir seri dengan pesaing; “mengalahkan model yang dilatih dengan data 20× lebih banyak” memang benar, tetapi hanya melalui perbezaan kecil pada metrik proksi, dalam satu penanda aras.

- Ia **tidak** menunjukkan kebolehan artistik umum. Ini ialah model penyelidikan lapan bilion parameter yang melukis ikon dan ilustrasi ringkas pada resolusi tetap yang kecil (224×224 piksel), bukan pereka serba guna.
- Perbandingan dengan model umum yang besar (GPT-5) **bukan** perbandingan yang setara: model itu tidak dibina atau ditala untuk tugas khusus melukis melalui kod ini, jadi mengalahkannya di sini tidak banyak memberitahu kita tentang kedua-dua model secara umum.
- Ia **bukan** percuma dari segi kos pengiraan. Merender dan membaca semula kanvas pada setiap langkah menjadikan penjanaan lebih perlahan daripada mengeluarkan kod dalam satu laluan buta — satu kos yang diakui oleh para penulis.

Sejauh mana kukuh buktinya?

- **Kukuh** apabila ia melibatkan perbandingan terkawal dalam skopnya sendiri. Ujian ablasi jelas: buang latihan, atau buang pengesahan, dan skornya jatuh — jadi kedua-dua ramuan itu memang melakukan kerja yang didakwa oleh para penulis.
- **Jujur tentang kejutan sendiri.** Penemuan bahawa maklum balas visual naif sebenarnya *memburukkan* hasil dilaporkan, bukan disembunyikan, dan itulah bahagian paling menarik dalam makalah — pembetulan berguna terhadap intuisi bahawa lebih banyak input sentiasa lebih baik.
- **Lebih lemah apabila menjadi dakwaan papan kedudukan.** Kemenangan berbanding pesaing yang menggunakan lebih banyak data adalah kecil dan hanya muncul pada satu penanda aras yang dibina oleh penulis salah satu pesaing tersebut. Margin kecil pada metrik proksi dalam satu set ujian memberi petunjuk, bukan keputusan muktamad.
- **Belum diuji pada skala besar atau dalam penggunaan sebenar.** Semua yang diuji di sini ialah ikon dan ilustrasi ringkas pada resolusi rendah. Sama ada idea yang sama berkesan untuk reka bentuk kompleks, resolusi tinggi atau kerja dunia sebenar masih tinggal untuk kajian akan datang — dan para penulis sendiri menyatakannya.

Mengapa ia penting

Idea di tengah makalah ini hampir memalukan kerana terlalu mudah, tetapi ia melangkaui lukisan. Jika sebuah program akan menjana sesuatu dengan menulis kod — halaman web, carta, rajah, adegan 3D — ia boleh sama ada menulis semuanya secara buta dan berharap hasilnya betul, atau merender sambil bergerak dan membetulkan arah. Bagi manusia, menutup gelung ini adalah perkara semula jadi; kita sentiasa menjeling halaman ketika bekerja. Untuk model-model ini, langkah tersebut ternyata masih agak baharu.

Apa yang menjadikan makalah ini berbaloi dibaca ialah tanda bintang yang ditambahkannya. Memberi model cara untuk melihat tidak sama dengan mengajarnya cara memandang. Matanya perlu dilatih, dan hanya selepas itu gelung ini memberi hasil — dan walaupun begitu, hasilnya nyata tetapi sederhana: pelukis teknikal yang lebih konsisten, bukan jenis artis yang berbeza. Bagi sesiapa yang membina alat untuk menukar penerangan menjadi grafik boleh diedit — jenis grafik yang akhirnya

digunakan untuk ikon dan ilustrasi aplikasi — pengajaran praktikal yang tenang inilah yang berguna. Kemenangan di sini datang daripada latihan yang lebih murah dan lebih pintar, bukan daripada lebih banyak data. Itu lebih berharga untuk dipelajari daripada satu lagi mata di papan kedudukan.

Ringkasan utama

Model bahasa yang menjana grafik vektor — kod boleh diedit dan diskalakan semula yang berada di sebalik kebanyakan ikon dan logo web — secara tradisinya melakukannya "secara buta", menulis semua arahan lukisan tanpa pernah merendernya untuk melihat hasil. Pasukan yang diketuai Guotao Liang mencadangkan Render-in-the-Loop: render lukisan separuh siap selepas setiap langkah dan berikan semula kepada model, supaya ia membuat goresan seterusnya sambil melihat kanvas. Penemuan utama mereka yang jujur ialah sekadar melakukan ini pada model sedia ada menjadikannya *lebih buruk*; peningkatan hanya muncul selepas model dilatih semula untuk menggunakan maklum balas visual, dibantu oleh pemeriksaan semasa melukis yang membuang goresan yang tidak mengubah apa-apa. Model lapan bilion parameter yang dilatih semula itu menyamai atau sedikit mengatasi pesaing yang dilatih dengan data sehingga dua puluh kali lebih banyak pada satu penanda aras piawai — hasil yang sebenar, tetapi dengan margin kecil pada skor proksi, untuk ikon dan ilustrasi ringkas pada resolusi rendah. Kesimpulan yang ditekankan oleh para penulis, dan yang patut diingati, ialah kecekapan: melihat hasil kerja dengan baik boleh menggantikan sebahagian daripada kelebihan skala data semata-mata.

Semakan tanpa gembar-gembur

Apa yang ditunjukkan oleh makalah: Melatih semula model grafik vektor untuk merender dan melihat lukisan separuh siapnya sendiri, langkah demi langkah, menghasilkan gambar yang lebih baik dan lebih lengkap daripada melukis secara buta — bersaing dengan, dan sedikit mendahului, pesaing yang dilatih dengan lebih banyak data pada satu penanda aras piawai, walaupun menggunakan data latihan yang lebih sedikit.

Apa yang munasabah tetapi belum dibuktikan: Bahawa "melihat kerja sendiri" secara umum ialah strategi yang lebih baik daripada sekadar menambah skala data; bahawa idea yang sama akan membantu grafik kompleks, resolusi tinggi atau penggunaan sebenar; bahawa margin kecil pada penanda aras itu mencerminkan perbezaan yang benar-benar akan diperhatikan manusia.

Apa yang tidak ditunjukkan: Bahawa maklum balas visual membantu dengan sendirinya (tanpa latihan semula ia memburukkan hasil); bahawa kelebihan berbanding pesaing besar atau muktamad; kebolehan melukis serba guna; atau perbandingan yang adil dengan model umum seperti GPT-5, yang tidak dibina untuk tugas ini.

Had utama: Satu penanda aras, yang sebahagiannya dibina oleh penulis model pesaing; margin kecil pada metrik proksi; model 8B, ikon dan ilustrasi ringkas pada 224×224; penjanaan lebih perlahan kerana gelung yang merender pada setiap langkah.

Berapa banyak keyakinan patut diberikan pembaca umum? Tinggi bahawa menutup gelung — merender dan membaca semula sambil melukis — benar-benar membantu, dan bahawa ia perlu dilatih, bukan sekadar dihidupkan. Rendah hingga sederhana bahawa model khusus ini mengalahkan pesaingnya secara tegas; anggap frasa “mengalahkan 20× data” sebagai hasil sebenar tetapi sederhana daripada satu penanda aras. Tinggi terhadap satu idea yang patut diingati: untuk kod yang melukis, melihat sambil bergerak lebih baik daripada melukis secara buta.

Sumber

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/