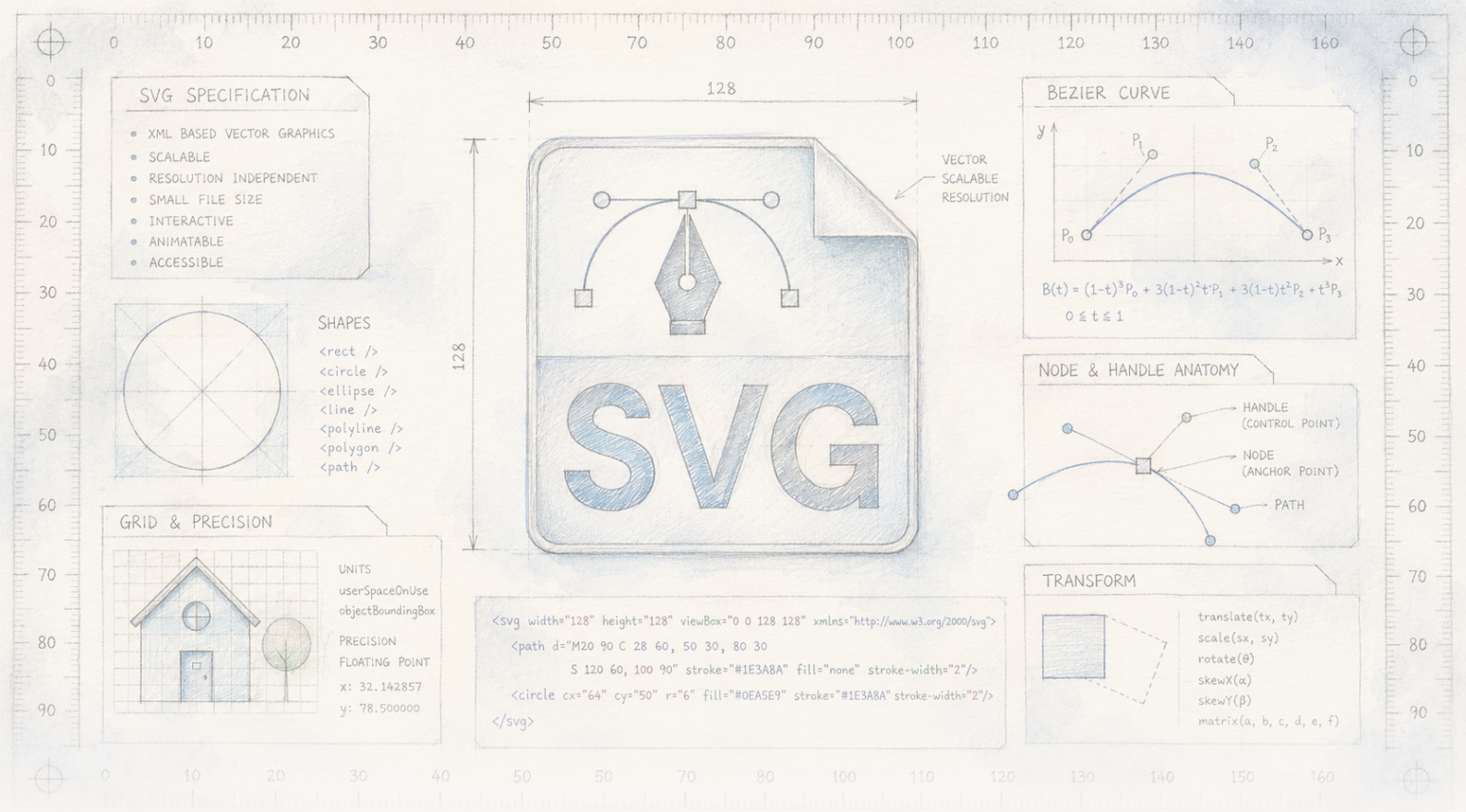


The Clean Paper

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

تعليم ذكاء اصطناعي للرسم أن ينظر إلى الصفحة

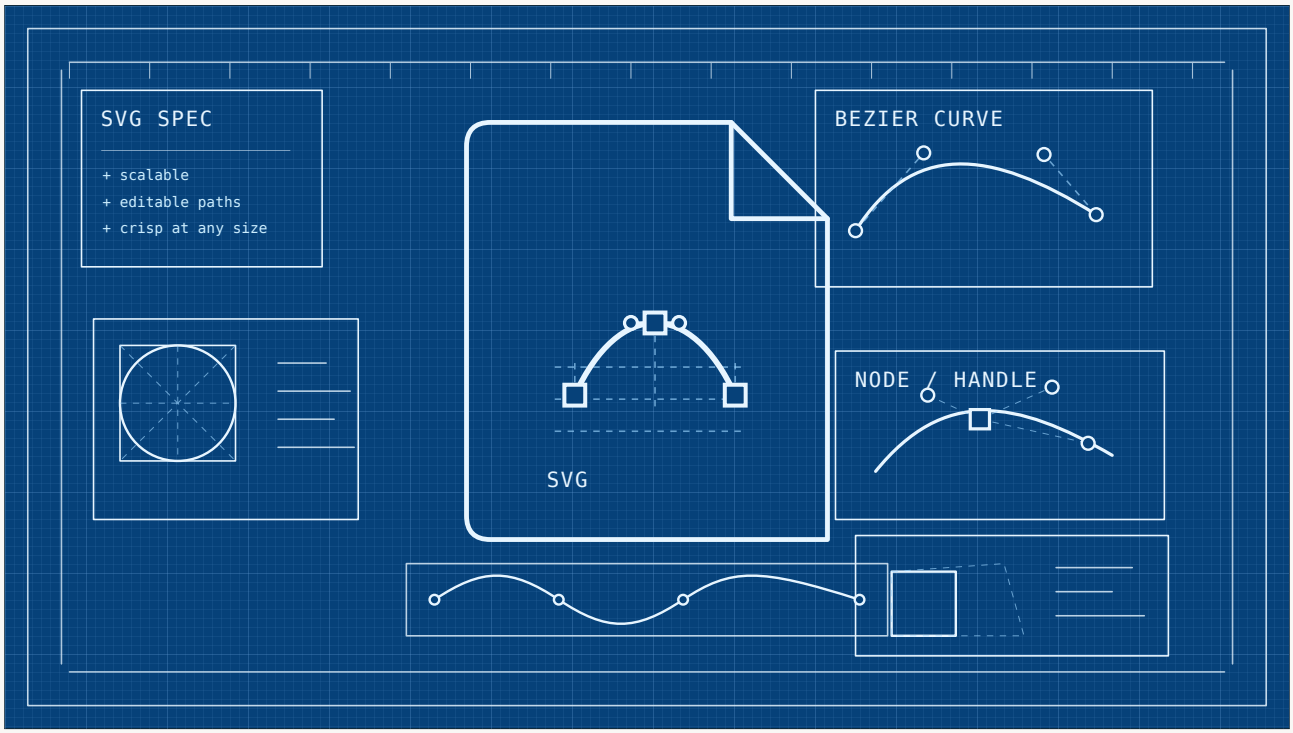
نماذج اللغة التي تولّد رسومات متجهية تعمل عادةً وهي عمياء — تكتب أوامر الرسم من دون أن ترى النتيجة. تسمح طريقة جديدة للنموذج بمشاهدة لوحته وهي تمتلئ ضربة بعد ضربة، والمفارقة الصريحة أن مجرد منحه «عينين» يجعل الأداء أسوأ: يجب إعادة تدريبه كي يتعلم استخدامهما. والنتيجة نموذج يطابق أو يتقدم قليلًا على منافسين تدربوا على بيانات أكثر بما يصل إلى عشرين ضعفًا — نتيجة حقيقية ومدروسة على اختبار واحد، لا ثورة.

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 Sefl-Feedback*, Guotao Liang, Zhangcheng Wang, Juncheng Hu, Haitao Zhou, Ziteng Xue, Jing Zhang, Dong Xu, and Qian Yu, Preprint (arXiv:2604.20730)

الرسم وعيناك مغمضتان

اطلب من أحد نماذج الذكاء الاصطناعي الحالية أن يرسم لك صورة، وسيحدث في الداخل واحد من أمرين مختلفين جدًا. مولدات الصور المألوفة — التي تستحضر صورة فوتوغرافية من جملة — ترسم البكسلات مباشرة، ويمكنها «رؤية» اللوحة أثناء العمل. لكن هناك نوعًا ثانيًا أكثر هدوءًا من الرسم لا يرسم فيه النموذج شيئًا بالفعل، بل يكتب تعليمات: ارسم دائرة هنا، وخطًا هناك، واملأ هذا الشكل بالأزرق. هذه التعليمات هي شيفرة — صيغة Graphics Vector Scalable أو SVG نفسها التي تقف خلف معظم الأيقونات والشعارات على الويب — ومغرياتها حقيقية: الناتج ليس شبكة ثابتة من البكسلات، بل مجموعة أشكال يمكن تكبيرها وتصغيرها وإعادة تلوينها وتحريرها بلا نهاية من دون أن تصبح ضبابية.



عمل أصلي على هيئة مخطط تقني بصيغة SVG: الصورة نفسها ملف متجهي قابل للتحرير، مبني من مسارات وخطوط شبكة وعُقَد ومقابض تحكم بدلًا من بكسلات ثابتة.

Laura Nesso / The Clean Paper Permission on file

المشكلة أن النموذج الذي يكتب شيفرة الرسم كان، حتى وقت قريب، يفعل ذلك وهو أعمى. كان يولّد سلسلة التعليمات كاملة في تمريرة واحدة — دائرة، خط، تعبئة — من دون أن يرسمها بين الخطوات ليرى ما صنعه. تخيّل أنك ترسم وجهًا وعيناك مغمضتان: قد تضع العينين تقريبًا حيث ينبغي، لكنك لن تلاحظ أن الثانية انتهت على الخد أو أن شكل الشعر صار يغطي الأنف. هكذا كانت هذه النماذج تعمل إلى حد بعيد، ولهذا كثيرًا ما أنتجت شيفرة صحيحة تمامًا من الناحية التقنية لكنها فوضى بصريًا.

قام فريق يقوده Liang Guotao الآن بما يبدو حلًا واضحًا إلى حد الطرافة: سمحوا للنموذج بأن يفتح عينيه. طريقتهم، *Render-in-the-Loop*، ترسم الصورة غير المكتملة بعد كل خطوة ثم تعيد هذه الصورة إلى النموذج قبل أن يضيف الضربة التالية. والجزء المثير للاهتمام حقًا ليس أن هذا يساعد، بل ما اضطروا إلى فعله كي يساعد أصلًا.

كيف نمنح الرسم درجة؟

صورة تقييم ذلك؟ قيس وكيف ماذا، في يتفوق نسأل: أن العادل فمن آخر، نموذج على «يتفوق» نموذجها إن ورقة تقول عندما من صغير عدد على المجال يعتمد لذلك — محمولًا» حاسوبًا «ارسم مثل لطلب وحيدة صحيحة إجابة توجد فلا — فعلًا صعب مولدة النتيجة. إلى ينظر شخص عن كامل غير بديل منها واحد وكل الآلية، المقاييس تظهر عدة مقاييس في هذه الورقة. يقارن **FID** الطابع الإحصائي العام لمجموعة كاملة من الصور المولدة بصور حقيقية؛ الأقل أفضل، لكنه يصف المجموعة لا صورة بعينها. ويسأل **score CLIP** نموذج ذكاء اصطناعي آخر هل تتطابق الصورة مع كلمات الطلب — وهو مفيد بقدر دقة ذلك الحكم. أما **DINO** و**SSIM** و**LPIPS** فتقارن صورة معاد بناؤها بصورة هدف على مستويات تمتد من البكسلات الخام إلى السمات المتعلمة.

ولا واحد من هذه المقاييس هو «الحقيقة». إنها مؤشرات بديلة، وغالبًا ما تتحرك بفروق صغيرة. عندما تقرأ أن نموذجًا يسجل 6.127 وآخر 8.128، فهذه بالفعل حركة في الاتجاه المقصود — لكنها دفعة صغيرة لا انهيار ساحق، كما أنها دفعة في رقم لا يتطابق إلا بصورة تقريبية مع ما ستقوله عينك. ومن المفيد تذكر ذلك عندما يكون العنوان «يتفوق على نموذج تدرب على عشرين ضعف البيانات».

ماذا فعل الباحثون؟

المشكلة التي أراد الباحثون إصلاحها هي الرسم الأعمى نفسه. تتعامل النماذج الموجودة مع كتابة SVG بوصفها مهمة نصية خالصة: توقّع الجزء التالي من الشيفرة اعتمادًا على الشيفرة السابقة، ولا ترسم النتيجة مطلقًا. وهذا يترك «العينين» القويتين — مشفر الرؤية — الموجودتين أصلًا في النماذج الحديثة متعددة الوسائط من دون استخدام. تعيد Render-in-the-Loop بناء المهمة كعملية بصرية خطوة بخطوة. بعد كل جزء من شيفرة الرسم، يُرسم SVG الجزئي إلى صورة وتُعاد الصورة إلى النموذج، فيختار الجزء التالي وهو ينظر فعلًا إلى اللوحة حتى تلك اللحظة.

أولى نتائجهم تحذيرية، ويحسب لهم أنهم يعرضونها بوضوح: مجرد إضافة هذه الحلقة إلى نموذج جاهز لا ينجح. عندما أعادوا الصور الوسيطة إلى نماذج عامة قوية من دون تدريب خاص، لم تتحسن الجودة — بل **ساءت في كل الحالات**. فالنموذج الذي لم يتعلم استخدام عينيه لهذه المهمة لا يعرف فجأة ماذا يفعل بهما.

لذلك يقع معظم العمل في التدريب. أعاد الفريق بناء بيانات التدريب بحيث يُقسّم كل رسم إلى خطوات صغيرة ذات معنى بصري — مع تفكيك الأشكال المعقدة إلى أجزاء أبسط كي يكون هناك شيء جديد يمكن رؤيته في كل مرحلة — ثم أجروا ضبطًا دقيقًا لنموذج مفتوح من ثمانية مليارات مُعامل (مبني على Qwen3-VL) على هذه السلاسل خطوة بخطوة. ويسمون ذلك Visual Sefl-Feedback. كما أضافوا آلية ثانية عند الرسم، Render-and-Verfiy: قبل قبول كل ضربة جديدة، يرسمها النموذج ويتحقق هل غيّرت الصورة فعلًا أم أنها كررت الضربة السابقة فقط. تُحذف الضربات التي لا تضيف شيئًا، وإذا لم يعد شيء مفيدًا يُطلب من النموذج أن يتوقف. واللافت أن هذا كله يستخدم مجموعة بيانات صغيرة نسبيًا — نحو 850 ألف مثال، أقل من نصف ما استخدمه منافس وأقل بكثير من جزء البيانات التي استخدمها منافس آخر.

ماذا وجدوا؟

- بعد هذا التدريب يرسم النموذج أفضل من نسخته العمياء — ويظهر ذلك بوضوح خصوصًا في حالات الفشل، مثل وضع نموذج أعمى عينًا على الخد، أو تجاهله مخطط أعمدة مطلوبًا وإخراجه شاشة عامة بدلًا منه.

- على الاختبار القياسي MMSVGBench، تصبح الطريقة منافسة للنماذج القوية وتتقدم عليها قليلًا في عدة مقاييس — ومنها OmniSVG الذي تدرب على أكثر من ضعف البيانات، وInternSVG الذي تدرب على نحو عشرين ضعفًا.
- الفروق صغيرة. في مجموعة الأيقونات مثلًا، تبلغ قيمة المقياس الرئيسي لجودة الصورة 6.127 مقابل 8.128 لأفضل منافس، وتبلغ درجة مطابقة الطلب 293.0 مقابل 291.0 — فروق حقيقية ومتسقة، لكنها نحيلة.
- كلا الجزأين المضافين يبرران وجودهما: إذا أوقفت التدريب الخاص أو التحقق أثناء الرسم، تهبط الأرقام بصورة قابلة للقياس. وتمنع خطوة التحقق خصوصًا النموذج من أن يعلق وهو يعيد رسم الشيء نفسه مرارًا.
- النتيجة التي يؤكدها الباحثون أنفسهم هي الكفاءة — بلوغ هذا المستوى ببيانات تدريب أقل بكثير مما استخدمه المتصدرون.

ما الذي لا تثبته الدراسة؟

- لا تُظهر أن السماح للنموذج بأن «يرى» مكسب مجاني. بل العكس: تجربة الورقة نفسها تُظهر أن التغذية الراجعة البصرية **من دون إعادة تدريب تجعل الأداء أسوأ**. المكسب يأتي من إعادة التدريب، لا من «العينين» وحدهما.
- لا تثبت تفوقًا كبيرًا أو حاسمًا. في معظم المقاييس تتقارب الطريقة مع منافسيها؛ وعبارة «يتفوق على نموذج تدرب على 20× من البيانات» صحيحة، لكن بفروق طفيفة في مقاييس بديلة وعلى اختبار واحد.
- لا تثبت قدرة فنية عامة. هذا نموذج بحثي من ثمانية مليارات مُعامل يرسم أيقونات ورسومًا بسيطة بدقة صغيرة ثابتة (224×224 بكسل)، وليس مصممًا عامًا.
- المقارنة مع نموذج عام كبير (GPT-5) **ليست مقارنة متكافئة**: ذلك النموذج لم يُبن أو يُضبط لهذه المهمة الضيقة المتمثلة في الرسم بالشيفرة، ولذلك لا يخبرنا التفوق عليه هنا بالكثير عن أي من النموذجين عمومًا.
- ولا تأتي الطريقة **بلا تكلفة**. رسم اللوحة وإعادة قراءتها عند كل خطوة يجعل التوليد أبطأ من إخراج الشيفرة كلها في تمريرة عمياء واحدة — وهي تكلفة يعترف بها الباحثون.

ما مدى قوة الأدلة؟

- الأدلة **قوية** عندما تكون المقارنة مضبوطة ضمن شروط الدراسة نفسها. تجارب إزالة المكونات واضحة: احذف التدريب أو احذف التحقق، فتهبط النتائج — ما يعني أن هذين العنصرين يفعلان بالفعل ما ينسبه إليهما الباحثون.
- الورقة **صريحة بشأن مفاجأتها الخاصة**. اكتشاف أن التغذية الراجعة البصرية الساذجة تضر بالأداء يُعرض بدلًا من إخفائه، وهو أكثر ما يثير الاهتمام في الورقة: تصحيح مفيد للحدس القائل إن المزيد من المدخلات أفضل دائمًا.
- الأدلة **رقيقة عندما تتحول إلى ادعاء ترتيب على لوحة الصدارة**. الانتصارات على منافسين تدربوا على بيانات أكثر صغيرة، وتعيش كلها على اختبار واحد بناه مؤلفو أحد هؤلاء المنافسين. فروق صغيرة في مقاييس بديلة على مجموعة اختبار واحدة تعطي إشارة، لكنها لا تحسم المسألة.
- الطريقة **غير مختبرة على نطاق واسع أو في الاستخدام الحقيقي**. كل شيء هنا أيقونات ورسوم بسيطة بدقة منخفضة. وما إذا كانت الفكرة نفسها تعمل على أعمال تصميم معقدة أو عالية الدقة أو واقعية تُرك للمستقبل — والباحثون يقولون ذلك أيضًا.

لماذا يهم الأمر؟

الفكرة في قلب هذه الورقة تكاد تكون بسيطة على نحو محرج، وهي تمتد أبعد من الرسم بكثير. إذا كان برنامج سيولّد شيئًا بكتابة شيفرة — صفحة ويب، أو مخططًا، أو رسمًا بيانيًا، أو مشهدًا ثلاثي الأبعاد — فيمكنه أن يكتب كل شيء وهو أعمى ويأمل في الأفضل، أو أن يرسم ما يصنعه أثناء الطريق ويصحح مساره. إغلاق هذه الحلقة أمر بديهي للإنسان؛ فنحن ننظر إلى الصفحة باستمرار. لكنه خطوة حديثة بصورة مفاجئة لهذه النماذج.

وما يجعل الورقة جديرة بالقراءة هو النجمة التي تضعها بجانب الفكرة: منح النموذج وسيلة للرؤية ليس هو نفسه تعليمه كيف ينظر. يجب تدريب «العينين»، وعندها فقط تصبح الحلقة مفيدة — وحتى حينها فالمكسب حقيقي لكنه محسوب: رسام أكثر ثباتًا، لا نوعًا جديدًا من الفنانين. ولمن يبني أدوات تحول وصفًا إلى رسم متجهي قابل للتحرير — من النوع الذي ينتهي خلف أيقونات التطبيقات ورسومها — فالدرس العملي الهادئ هو الأهم. جاء المكسب هنا من تدريب أذكى وأرخص، لا من مزيد من البيانات. وهذا أَولى بالتعلم من نقطة إضافية على لوحة ترتيب.

الخلاصة

كانت نماذج اللغة التي تولّد الرسوم المتجهية — أي الشيفرة القابلة للتحرير والتكبير التي تقف خلف معظم أيقونات وشعارات الويب — تقوم بذلك تقليديًا وهي «عمياء»: تكتب أوامر الرسم كلها من دون أن ترسمها لترى النتيجة. يقترح فريق يقوده Liang Guotao طريقة Render-in-the-Loop: ارسم الصورة غير المكتملة بعد كل خطوة وأعدها إلى النموذج، بحيث يرسم الضربة التالية وهو ينظر إلى اللوحة. والنتيجة الصريحة المركزية هي أن مجرد فعل ذلك مع نموذج موجود يجعله **أسوأ**؛ ولا يظهر المكسب إلا بعد إعادة تدريب النموذج على استخدام التغذية الراجعة البصرية، بمساعدة فحص أثناء الرسم يحذف الضربات التي لا تغيّر شيئًا. يطابق النموذج المعاد تدريبه، ذي الثمانية مليارات مُعامل، منافسين تدربوا على بيانات أكثر بما يصل إلى عشرين ضعفًا أو يتفوق عليهم قليلًا في اختبار قياسي — نتيجة حقيقية، لكن بفروق صغيرة في مقاييس بديلة، ولأيقونات ورسوم بسيطة منخفضة الدقة. والخلاصة التي يؤكدها الباحثون، والتي تستحق الاحتفاظ بها، تتعلق بالكفاءة: رؤية عملك بطريقة جيدة قد تعوض جزءًا من التوسع الخام في البيانات.

مراجعة بلا تهويل

ما الذي تُظهره الورقة: إعادة تدريب نموذج للرسوم المتجهية كي يرسم وينظر إلى صورته غير المكتملة خطوة بخطوة تنتج رسومًا أفضل وأكثر اكتمالًا من الرسم الأعمى — منافسة لنماذج تدربت على بيانات أكثر ومتقدمة عليها قليلًا في اختبار قياسي واحد، باستخدام بيانات تدريب أقل.

ما هو معقول لكنه غير مثبت: أن «رؤية ما تصنعه» وصفة أفضل عمومًا من توسيع البيانات؛ وأن الفكرة نفسها ستساعد مع الرسوم المعقدة أو عالية الدقة أو الواقعية؛ وأن الفروق الصغيرة في الاختبار تعكس فرقًا سيلاحظه الإنسان فعلًا.

ما الذي لا تُظهره: أن التغذية الراجعة البصرية تساعد من تلقاء نفسها (من دون إعادة التدريب تضر)؛ أو أن التفوق على المنافسين كبير أو حاسم؛ أو قدرة رسم عامة؛ أو مقارنة عادلة وجهًا لوجه مع نماذج عامة مثل GPT-5 لم تُبن لهذه المهمة.

القيود الرئيسية: اختبار واحد بناه جزئيًا مؤلفو نموذج منافس؛ فروق صغيرة في مقاييس بديلة؛ نموذج 8B، وأيقونات ورسوم بسيطة بدقة 224×224؛ وتوليد أبطأ بسبب حلقة الرسم بعد كل خطوة.

ما مقدار الثقة المناسبة للقارئ العام؟ ثقة عالية بأن إغلاق الحلقة — الرسم ثم إعادة النظر أثناء العمل — يساعد فعلًا، وأنه يجب أن يُدرَّب عليه لا أن يُشغَّل كمفتاح فقط. وثقة منخفضة إلى متوسطة في أن هذا النموذج بعينه يتفوق بصورة حاسمة على منافسيه؛ تعامل مع عبارة «يتفوق رغم 20× من البيانات» بوصفها نتيجة حقيقية لكنها متواضعة وعلى اختبار واحد. وثقة عالية في الفكرة الوحيدة التي تستحق التذكر: عندما ترسم الشيفرة، النظر أثناء العمل أفضل من الرسم وأنت أعمى.

المصادر

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

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

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