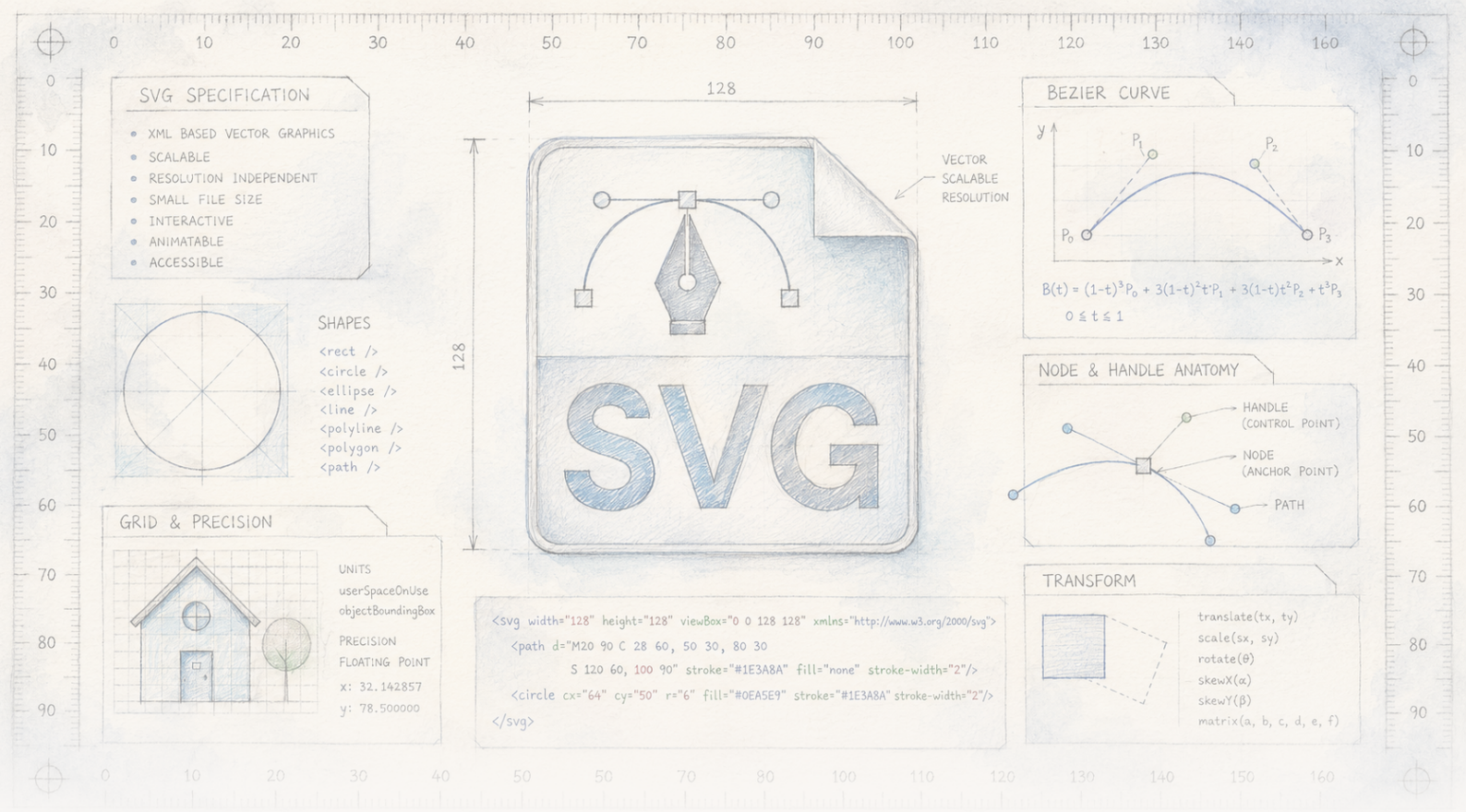


The Clean Paper

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

آنکھیں بند کر کے ڈرائنگ کرنا

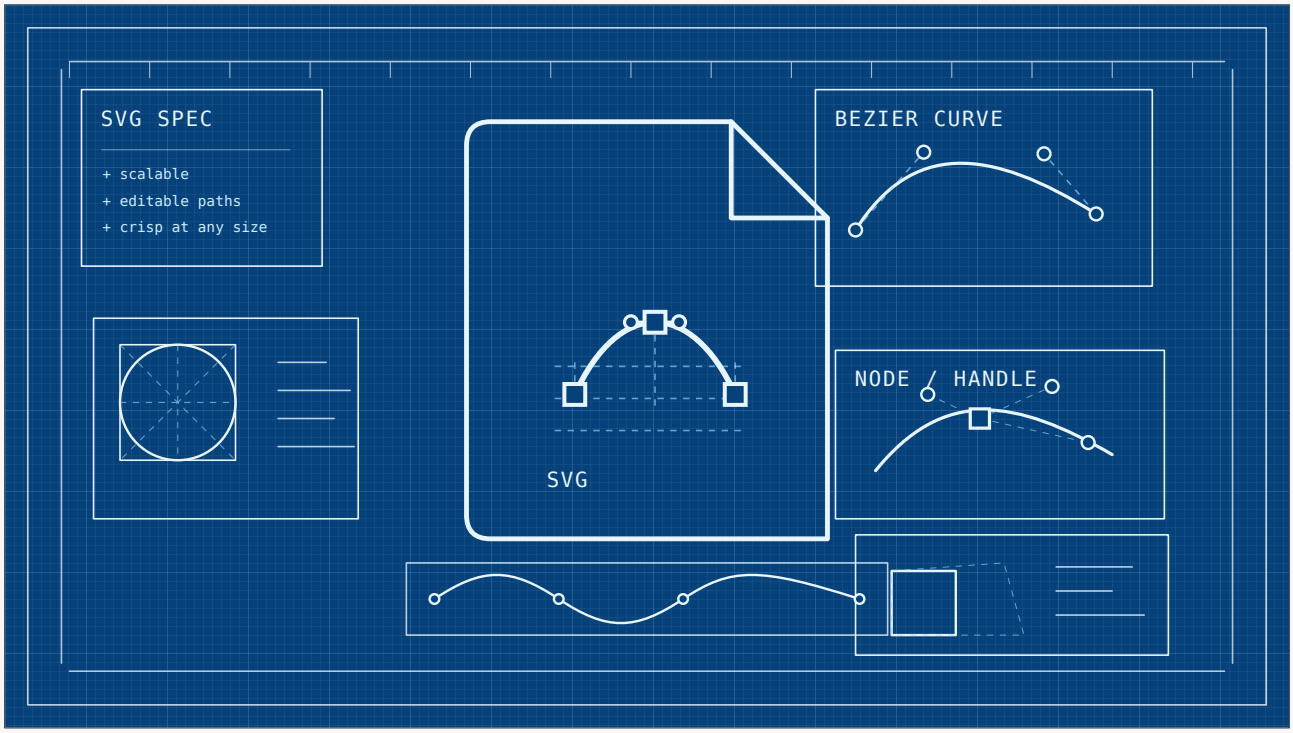
generate graphics vector کرنے والے زبان ماڈلز روایتی طور پر "بلائنڈ" تھے: تمام *SVG* ڈرائنگ *commands* لکھتے تھے مگر نتیجہ *render* کر کے نہیں دیکھتے تھے۔ *Liang Guotao* کی *team* لوپ-*Render-in-the-propose* کرتی ہے: ہر مرحلہ کے بعد *hafl-finished* ڈرائنگ *render* کریں اور ماڈل کو واپس دیں تا کہ اگلا *canvas stroke* دیکھتے ہوئے بنے۔ *central* دیانت دار *finding* یہ ہے کہ *existing* ماڈل پر یہ بس *feedback* کر دیں تو ماڈل ***worse*** ہوتا ہے؛ صرف *gain* معیاری بینچ مارک پر *rivals* کو مطابقت یا *beat slightly* کرتا ہے، جن میں 20× زیادہ ڈیٹا پر *trained* ماڈل بھی شامل ہے—*genuine* نتیجہ، مگر چھوٹا میٹرک-*proxy*، *margins* کم-*res* سادہ/*icons illustrations* پر۔ مصنفین کا اہم *takeaway* کارکردگی ہے: اپنے کام کو صحیح طرح "دیکھنا" *sheer* ڈیٹا پیمانہ کا کچھ متبادل بن سکتا ہے۔ *retraining* کے بعد آتا ہے، ساتھ ڈرائنگ-وقت جانچ جو نہیں-تبدیلی *discard strokes* کرتی ہے۔ *8B retrained* ماڈل

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)

آنکھیں بند کر کے ڈرائنگ کرنا

آج کے اے آئی ماڈل سے تصویر بنانے کو کہیں تو اندر سے دو بہت مختلف چیزوں میں سے ایک ہوتی ہے۔ familiar تصویر جو—sentence generators سے photograph بنا دیتے ہیں—pixels کو براہِ راست paint کرتے ہیں اور کام کرتے ہوئے canvas "دیکھ" سکتے ہیں۔ لیکن ڈرائنگ کی ایک دوسری، quieter قسم بھی ہے، جس میں ماڈل paint نہیں کرتا۔ وہ ہدایات لکھتا ہے: یہاں *circle* بناؤ، وہاں لائن، اس شکل کو *fill blue* کرو۔ یہ ہدایات کوڈ ہیں—وہی Scalable سمتی مقدار، Graphics، یا SVG جو web کے زیادہ تر icons اور logos کے پیچھے ہوتے ہیں۔ فائدہ حقیقی ہے: نتیجہ pixels کی مقرر گرڈ نہیں بلکہ ساختیں کا سیٹ ہے جسے بار بار، recolour rescale اور ترمیم کیا جا سکتا ہے، blur ہوئے بغیر۔



اصل artwork: blueprint SVG تصویر خود editable سمتی مقدار file ہے، مقرر pixels کے بجائے راستے، گرڈ لائنیں، nodes اور handles سے بنی ہوئی۔

Laura Nesso / The Clean Paper Permission on file

مسئلہ یہ تھا کہ حال تک ماڈل یہ ڈرائنگ کوڈ تقریباً بلائنڈ لکھتا تھا۔ وہ پوری ہدایت سلسلہ ایک pass میں emit کرتا—circle، لائن، بغیر—fill کبھی رینڈر کر کے دیکھے کہ بنایا کیا ہے۔ آنکھیں بند کر کے چہرہ sketch کرنے کا تصور کریں: شاید eyes تقریباً درست جگہ آئیں، مگر آپ notice نہیں کر سکتے کہ دوسری cheek eye پر چلی گئی یا hair کا شکل nose ڈھانپ رہا ہے۔ تقریباً یہی ماڈلز کے ساتھ ہوتا تھا، اسی لیے کوڈ perfectly درست ہوتے ہوئے بھی بصری نتیجہ mess بن سکتا تھا۔

Liang Guotao کی lead میں ٹیم نے تقریباً مزاحیہ حد تک واضح قدم اٹھایا: ماڈل کو "آنکھیں" کھولنے دیں۔ ان کا طریقہ، رینڈر-لوپ-*in-the*، ہر مرحلہ کے بعد ادھوری ڈرائنگ رینڈر کرتا ہے اور اگلا خط بنانے سے پہلے وہ تصویر ماڈل کو واپس دیتا ہے۔ واقعی دلچسپ حصہ یہ نہیں کہ ردِعمل مدد کر سکتی ہے—بلکہ یہ کہ مدد کروانے کے لیے ماڈل کو کیا سکھانا پڑا۔

ڈرائنگ کو اسکور کیسے کرتے ہیں؟

کیا؟ پیمائش کیسے میں، چیز کس ہے: سوال fair تو ہے کرتا رہا" بہتر "سے کو دوسرے ماڈل ہے کہتا مقالہ جب خودکار چند شعبہ لیے نہیں—اس جواب درست واحد کا کرو" draw ہے—"laptop مشکل کرنا judge تصویر generated ہے۔ imperfect proxy اسکور ہر ہے، کرتی rely پر اسکورز

یہ مقالہ کئی میٹرکس استعمال کرتا ہے۔ **FID** generated تصاویر کے گروہ کی مجموعی شماریاتی flavour کو حقیقی تصاویر سے موازنہ کرتا ہے، کم بہتر ہے، مگر گروہ describe کرتا ہے، فرد تصویر نہیں۔ **CLIP اسکور** ایک الگ اے آئی سے پوچھتا ہے تصویر ہدایت کے الفاظ سے کتنی مطابقت کرتی ہے—مفید مگر judge ماڈل کی quality پر انحصار کرنا۔ **DINO، SSIM** اور **LPIPS** reconstructed تصویر کو ہدف سے مختلف سطحیں پر موازنہ کرتے ہیں، pixels raw سے سیکھا خصوصیات تک۔

ان میں سے کوئی "truth" نہیں۔ یہ proxies ہیں، اور تبدیلیاں اکثر چھوٹی ہوتی ہیں۔ اگر ماڈل 6.127 اسکور کرے اور rival 8.128، intended سمت میں حقیقی فرق ہے—مگر landslide نہیں، اور میٹرک آنکھ کی judgement کو loose طور پر نظر رکھتی ہے۔ "20× ڈیٹا پر تربیت یافتہ ماڈل کو سے بہتر رہا کیا" سرخی پڑھتے وقت یہ caveat یاد رکھنا چاہیے۔

مصنفین نے کیا کیا

مصنفین کا ہدف بلائنڈ ڈرائنگ مسئلہ تھا۔ موجود ماڈلز SVG لکھنے کو خالص متن کام سمجھتے ہیں: اب تک کا کوڈ دیکھ کر اگلا predict chunk کرو، رینڈر نہ کرو۔ اس سے جدید multimodal ماڈلز کا vision اینکوڈر idle رہتا ہے۔ رینڈر-in-the-لوپ کام کو مرحلہ-by-مرحلہ بصری عمل بناتا ہے۔ ڈرائنگ کوڈ کے ہر ٹکڑا کے بعد SVG partial تصویر میں رینڈر ہوتی ہے اور ماڈل کو تصویر کے طور پر خوراک لینا ہوتی ہے، اس لیے اگلا ٹکڑا canvas دیکھ کر choose ہوتا ہے۔

پہلا نتیجہ cautionary ہے، اور credit یہ کہ مصنفین اسے صاف رپورٹ کرتے ہیں: موجود off-the-shefl ماڈل پر لوپ صرف bolt کر دینے سے کام نہیں بنتا۔ مضبوط عمومی ماڈلز کو renders intermediate خاص تربیت کے بغیر دیں تو quality بہتر کرنا نہیں ہوئی—**ہر طرف degrade ہوئی**۔ ماڈل جسے اپنی بصری ردِعمل استعمال کرنا سکھایا نہیں گیا، وہ اچانک اسے سمجھ نہیں لیتا۔

اس لیے اصل کام تربیت میں ہے۔ تربیتی ڈیٹا دوبارہ بنایا گیا تاکہ ہر ڈرائنگ بہت سے چھوٹے، بصری طور پر معنی خیز مراحل میں ٹوٹے—پیچیدہ ساختوں کو سادہ حصوں میں بانٹا گیا تاکہ ہر مرحلے پر کچھ نیا نظر آئے—پھر آٹھ ارب پیرامیٹر والا کھلا ماڈل Qwen3-VL) پر مبنی) مرحلہ وار تسلسل پر fine-tune کیا گیا۔ اسے صرف آخری تصویر نہیں، بننے کا عمل بھی دکھایا گیا۔ ڈرائنگ وقت پر دوسرا طریقۂ کار Render-and-Verfiy شامل ہوا: ہر نیا stroke قبول کرنے سے پہلے render کر کے جانچ کیا جاتا ہے کہ تصویر واقعی بدلی یا آخری stroke دھرایا گیا۔ تبدیلی نہ لانے والے strokes خارج کر دیے جاتے ہیں، اور جب مزید مفید تبدیلی نہ ہو تو ماڈل کو رکنے کا prompt دیا جاتا ہے۔ یہ سب تقریباً 850,000 مثالوں کے نسبتاً چھوٹے ڈیٹا سیٹ پر چلتا ہے—ایک حریف کے نصف سے کم اور دوسرے کے ڈیٹا سیٹ کا چھوٹا حصہ۔

انہوں نے کیا پایا

- اس تربیت کے بعد ماڈل بلائنڈ counterpart سے بہتر draw کرتا ہے، خاص طور پر ناکامی کیسز میں: بلائنڈ ماڈل eye cheek پر رکھ سکتا ہے یا chart bar requested چھوڑ کر عمومی monitor دے سکتا ہے۔
- معیاری بینچ مارک MMSVGBench پر طریقہ مضبوط rivals کے competitive ہے اور کئی پیمائشیں میں تھوڑا آگے بھی—OmniSVG جسے >2× ڈیٹا ملا، اور InternSVG جسے تقریباً 20× ڈیٹا ملا۔

- margins چھوٹا ہیں۔ icon سیٹ پر اہم تصویر-quality اسکور 6.127 ہے بمقابلہ بہترین rival 8.128، ہدایت-matching 293.0 بمقابلہ 291.0—ہم آہنگ اور حقیقی، مگر slender۔
- دونوں اجزا contribute کرتے ہیں: خاص تربیت یا ڈرائنگ-وقت verfiication بند کریں تو اعداد قابلِ پیمائش طور پر drop۔ verfiication ماڈل کو ایک ہی thing بار بار redraw کرنے میں stuck ہونے سے روکتی ہے۔
- مصنفین خود کارکردگی emphasize کرتے ہیں—leaders کے مقابلے میں بہت کم تربیت ڈیٹا سے اتنی کارکردگی۔

یہ کیا ثابت نہیں کرتا

- یہ نہیں دکھاتا کہ ماڈل کو "دیکھنے" دینا آزاد win ہے۔ مقالہ خود دکھاتا ہے بصری ردِعمل **بغیر retraining** چیزیں worse کرتی ہے۔ فائدہ تربیت سے ہے، eyes اکیلا سے نہیں۔
- یہ بڑی یا فیصلہ کن برتری ثابت نہیں کرتا۔ زیادہ تر اسکور بہت قریب ہیں؛ یہ کہنا درست ہے کہ ماڈل نے تقریباً 20 گنا زیادہ ڈیٹا والے حریف سے بہتر کارکردگی دکھائی، مگر فرق ایک بینچ مارک کے metrics proxy پر نسبتاً چھوٹا ہے۔
- یہ عمومی demonstrate ability artistic نہیں کرتا۔ یہ 8B تحقیق ماڈل icons اور سادہ illustrations 224×224 pixels پر draw کرتا ہے، عمومی-مقصد designer نہیں۔
- GPT-5 جیسے big عمومی ماڈل سے موازنہ **apples-to-apples نہیں**: وہ محدود کوڈ-ڈرائنگ کام کے لیے تیار/tuned نہیں، اس لیے یہاں سے بہتر رہا کرنا عمومی صلاحیت کے بارے میں کم بتاتا ہے۔
- یہ آزاد computationally بھی نہیں۔ ہر مرحلہ پر رینڈر اور re-read کرنے سے نسل بلائنڈ ایک-pass کوڈ اخراج سے زیادہ سست ہے۔

شواہد کتنے مضبوط ہیں

- **ٹھوس** اپنے کنٹرول شدہ موازنے میں۔ ablations صاف ہیں: تربیت یا remove verfiication کریں تو اسکورز fall، یعنی دونوں ingredients واقعی contribute کرتے ہیں۔
- **اپنی surprise کے بارے میں دیانت دار۔** naive بصری ردِعمل *hurt* کرتی ہے—یہ نتیجہ رپورٹ کی گئی، hide نہیں۔ مقالے کا شاید سب سے دلچسپ lesson یہی ہے: زیادہ input ہمیشہ بہتر نہیں۔
- **leaderboard دعویٰ میں پتلا۔** ڈیٹا-larger rivals پر wins چھوٹا ہیں، ایک بینچ مارک پر، اور بینچ مارک خود ایک rival کے مصنفین نے بنایا تھا۔ چھوٹا میٹرک-proxy margins اشارہ دینے والا ہیں، settled نہیں۔
- **پیمانہ/جنگلی میں untested۔** سادہ/icons کم-res۔ illustrations کمپلیکس، بلند-res یا حقیقی دنیا ڈیزائن میں خیال hold کرے گی یا نہیں آئندہ کام ہے۔

یہ کیوں اہم ہے

مقالے کی مرکزی خیال تقریباً embarrassingly سادہ ہے اور ڈرائنگ سے باہر جاتی ہے۔ اگر پروگرام کوڈ لکھ کر کچھ generate کر رہا ہے—web، page، chart، diagram 3D منظر—تو یا پوری چیز بلائنڈ لکھ کر hope کرے، یا as-it-goes رینڈر کر کے course درست کرے۔ انسان کے لیے لوپ قریب کرنا قدرتی ہے؛ ہم page بار بار دیکھتے ہیں۔ ماڈلز کے لیے یہ surprisingly recent حرکت ہے۔

مقالے کو واقعی قابلِ مطالعہ بنانے والی بات یہی شرط ہے: ماڈل کو دیکھنے کا راستہ دینا اور اسے **دیکھنا سکھانا** ایک چیز نہیں۔ بصری ردِعمل استعمال کرنے کی تربیت دینی پڑتی ہے، تب یہ چکر فائدہ دیتا ہے۔ فائدہ بھی ناپا گیا ہے: زیادہ مستحکم خاکہ نگار، کوئی بالکل نئی قسم کا فنکار نہیں۔ متن سے قابلِ ترمیم گرافکس بنانے والے آلات — مثلاً ایپ آئیکنز اور سادہ تصویریں — کے لیے خاموش مگر عملی سبق اہم ہے۔ یہاں بہتری زیادہ ڈیٹا سے نہیں بلکہ زیادہ ہوشیار اور کم خرچ تربیت سے آئی۔

صاف خلاصہ

ویکٹر گرافکس بنانے والے زبان ماڈل روایتی طور پر "اندھے" تھے: وہ تمام SVG ڈرائنگ احکامات لکھتے تھے مگر نتیجہ رینڈر کر کے دیکھتے نہیں تھے۔ Liang Guotao کی ٹیم **رینڈر-in-the-لوپ** تجویز کرتی ہے: ہر مرحلے کے بعد ادھوری ڈرائنگ رینڈر کر کے ماڈل کو واپس دکھائی جائے، تاکہ اگلی لکیر بناتے وقت وہ موجودہ کینوس دیکھ سکے۔ مرکزی اور اہم نتیجہ یہ ہے کہ کسی موجودہ ماڈل میں بس یہ بصری ردِعمل فعال کر دینا کارکردگی **خراب** کرتا ہے؛ فائدہ صرف دوبارہ تربیت کے بعد ملا، ساتھ ایسی جانچ کے جو ڈرائنگ میں کوئی تبدیلی نہ لانے والے خطوط کو خارج کرتی ہے۔ دوبارہ تربیت یافتہ 8B ماڈل معیاری بینچ مارکس پر حریفوں کے برابر یا قدرے بہتر رہا، حتیٰ کہ بعض ایسے ماڈلوں کے مقابلے میں جو 20 گنا زیادہ ڈیٹا پر تربیت یافتہ تھے۔ یہ حقیقی نتیجہ ہے، مگر فرق چھوٹا ہے اور آزمائش کم ریزولوشن آئیکنز اور سادہ تصویروں جیسے نمائندہ پیمانوں پر ہوئی۔ مصنفین کا مفید نکتہ کارکردگی ہے: اپنے کام کو درست طور پر "دیکھنا" بعض اوقات محض مزید ڈیٹا بڑھانے کا جزوی متبادل بن سکتا ہے۔

بلا مبالغہ جانچ

مقالے کیا دکھاتا ہے: سمتی مقدار-graphics ماڈل کو ادھوری ڈرائنگ رینڈر کر کے مرحلہ-by-مرحلہ دیکھنا سکھانے سے بلائنڈ نسل کے مقابلے میں بہتر/مکمل نتائج آتے ہیں؛ ایک بینچ مارک پر کم ڈیٹا کے باوجود ڈیٹا-larger rivals کے competitive یا slightly ahead۔

کیا قرینِ قیاس مگر ثابت شدہ نہیں: your "seeing" کام" وسیع طور پر ڈیٹا پیمانہ بندی سے بہتر نسخہ ہو سکتی ہے؛ کمپلیکس/بلند-res graphics میں ایک ہی اثر؛ چھوٹا میٹرک margins انسان کو noticeable visibly ہوں گے۔

کیا نہیں دکھاتا: بصری ردِعمل اکیلا مدد کرتی ہے—بغیر hurt retraining کرتی ہے؛ rivals پر فیصلہ کن ؛lead عمومی-مقصد ڈرائنگ ؛GPT-5 ability کے ساتھ fair عمومی موازنہ۔

اہم حدود: ایک بینچ مارک؛ چھوٹا اسکور-proxy ؛8B margins ماڈل؛ 224×224 سادہ/icons ؛illustrations رینڈر-every-مرحلہ کی وجہ سے زیادہ سست نسل۔

اعتماد: بلند کہ لوپ قریب کرنا مدد کرتا ہے اگر اسے تربیت دی جائے۔ rivals پر فیصلہ کن برتری کے لیے کم-to-درمیانہ۔ یاد رکھنے والا خیال: کوڈ سے ڈرائنگ میں، بلائنڈ لکھنے کے بجائے رینڈر کر کے دیکھتے رہنا بہتر ہے—مگر "eyes" استعمال کرنا سکھانا پڑتا ہے۔

ماخذ

(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/