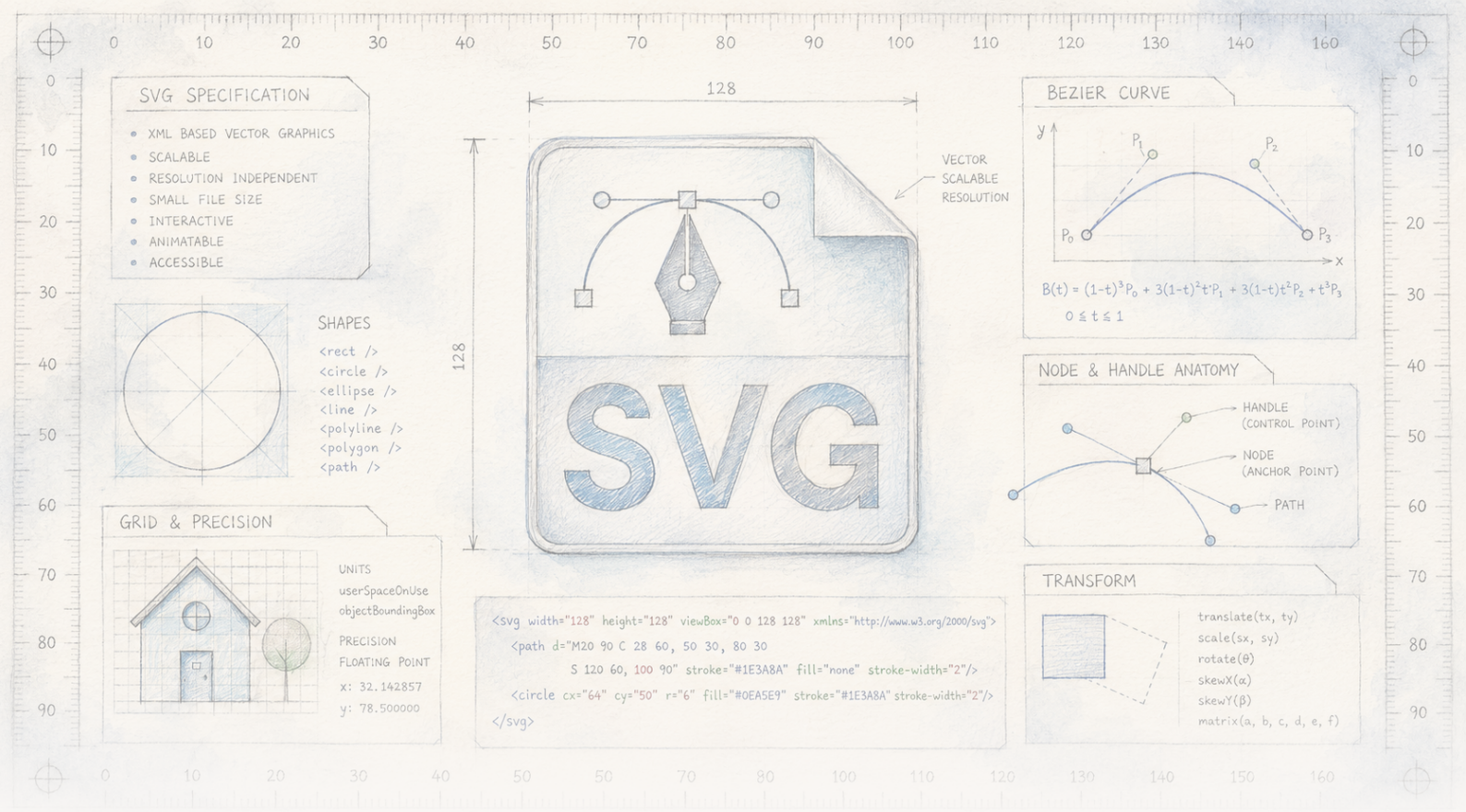


The Clean Paper

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

Μαθαίνοντας σε μια AI που ζωγραφίζει να κοιτάζει τη σελίδα

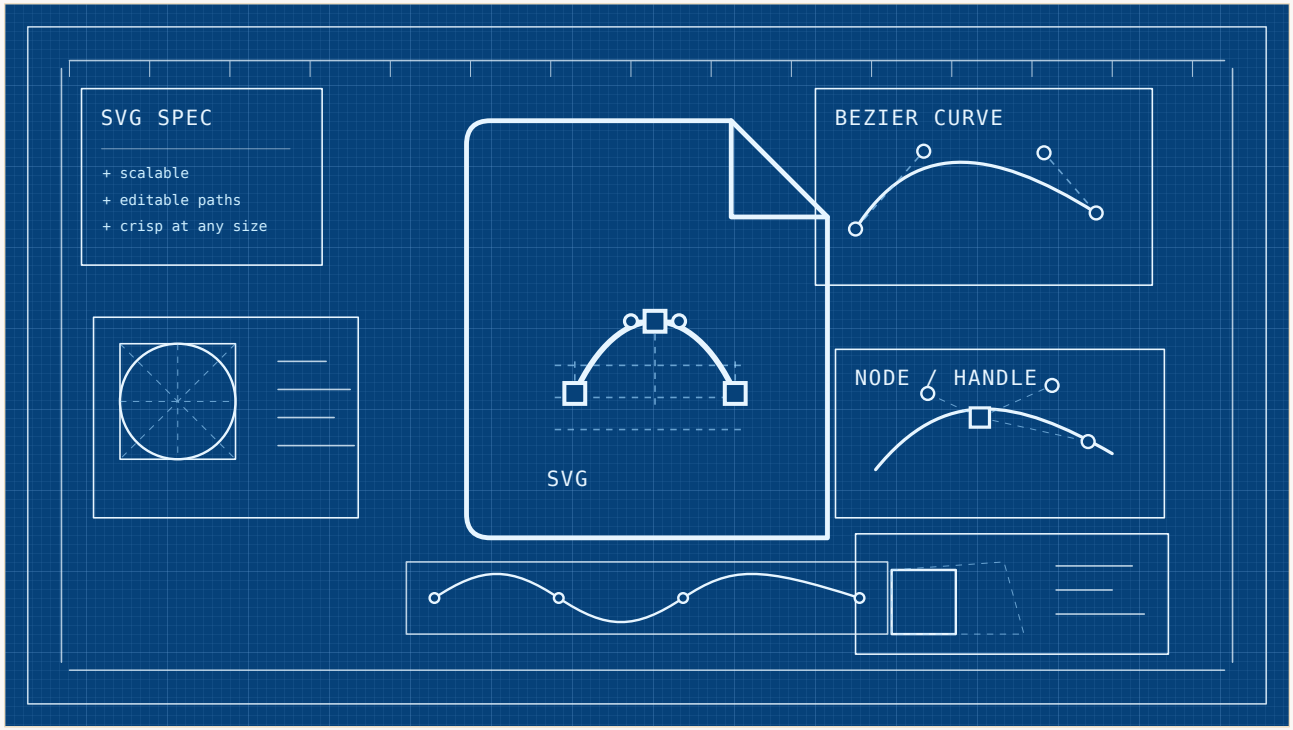
Τα γλωσσικά μοντέλα που παράγουν διανυσματικά γραφικά το κάνουν στα τυφλά — γράφουν τις εντολές σχεδίασης χωρίς να βλέπουν ποτέ το αποτέλεσμα. Μια νέα μέθοδος αφήνει το μοντέλο να παρακολουθεί τον καμβά του να γεμίζει, γραμμή προς γραμμή, και η έντιμη ανατροπή είναι ότι το να του δώσεις απλώς μάτια κάνει τα πράγματα χειρότερα: πρέπει να επανεκπαιδευτεί για να τα χρησιμοποιεί. Το αποτέλεσμα είναι ένα μοντέλο που ισοφαρίζει ή ξεπερνά οριακά ανταγωνιστές εκπαιδευμένους με έως είκοσι φορές περισσότερα δεδομένα — πραγματικό, προσεκτικό αποτέλεσμα σε ένα benchmark, όχι επανάσταση.

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)

Ζωγραφίζοντας με κλειστά μάτια

Ζητήστε από ένα σημερινό μοντέλο AI να σας ζωγραφίσει μια εικόνα και, κάτω από την επιφάνεια, συμβαίνει ένα από δύο πολύ διαφορετικά πράγματα. Οι οικείες γεννήτριες εικόνων — εκείνες που δημιουργούν μια φωτογραφία από μια πρόταση — ζωγραφίζουν απευθείας pixels και μπορούν να βλέπουν τον καμβά καθώς δουλεύουν. Υπάρχει όμως κι ένα δεύτερο, πιο ήσυχο είδος σχεδίασης, όπου το μοντέλο δεν ζωγραφίζει καθόλου. Γράφει οδηγίες: *σχεδίασε έναν κύκλο εδώ, μια γραμμή εκεί, γέμισε αυτό το σχήμα μπλε.* Αυτές οι οδηγίες είναι κώδικας — το ίδιο Scalable Vector Graphics, ή SVG, που βρίσκεται πίσω από τα περισσότερα εικονίδια και λογότυπα στο web — και το πλεονέκτημα είναι πραγματικό: το αποτέλεσμα δεν είναι ένα σταθερό πλέγμα pixels αλλά ένα σύνολο σχημάτων που μπορείς να μεγεθύνεις, να αλλάξεις χρώμα και να επεξεργαστείς επ' αόριστον χωρίς να θολώνει.



Πρωτότυπο σχέδιο blueprint σε SVG: η ίδια η εικόνα είναι επεξεργάσιμο διανυσματικό αρχείο, χτισμένο από διαδρομές, γραμμές πλέγματος, κόμβους και λαβές και όχι από σταθερά pixels.

Laura Nesso / The Clean Paper Permission on file

Το πρόβλημα είναι ότι, μέχρι πρόσφατα, ένα μοντέλο που έγραφε αυτόν τον κώδικα σχεδίασης το έκανε στα τυφλά. Παρήγαγε ολόκληρη την ακολουθία εντολών με μία φορά — κύκλος, γραμμή, γέμισμα — χωρίς ποτέ να την αποδώσει οπτικά για να δει τι είχε φτιάξει. Φανταστείτε να σχεδιάζετε ένα πρόσωπο με κλειστά μάτια: μπορεί να βάλετε περίπου τα μάτια εκεί όπου πρέπει, αλλά δεν θα έχετε τρόπο να παρατηρήσετε ότι το δεύτερο έπεσε στο μάγουλο ή ότι το σχήμα των μαλλιών βρίσκεται πλέον πάνω από τη μύτη. Έτσι περίπου δούλευαν αυτά τα μοντέλα, γι' αυτό και συχνά παρήγαγαν κώδικα απολύτως έγκυρο που οπτικά ήταν χάος.

Μια ομάδα με επικεφαλής τον Guotao Liang έκανε τώρα το σχεδόν κωμικά προφανές: άφησε

το μοντέλο να ανοίξει τα μάτια του. Η μέθοδός τους, ***Render-in-the-Loop***, αποδίδει την ημιτελή εικόνα μετά από κάθε βήμα και δίνει αυτή την εικόνα πίσω στο μοντέλο πριν σχεδιάσει την επόμενη γραμμή. Το πραγματικά ενδιαφέρον μέρος δεν είναι ότι αυτό βοηθά — είναι τι χρειάστηκε να κάνουν ώστε να βοηθήσει καθόλου.

> **Πώς βαθμολογείται ένα σχέδιο;**
>
> Όταν μια εργασία λέει ότι το μοντέλο της «νικά» ένα άλλο, είναι δίκαιο να ρωτήσουμε: σε τι ακριβώς και πώς μετρήθηκε; Η αξιολόγηση μιας παραγόμενης εικόνας είναι πραγματικά δύσκολη — δεν υπάρχει μία σωστή απάντηση στο «ζωγράφισε ένα laptop» — έτσι ο κλάδος στηρίζεται σε λίγες αυτόματες βαθμολογίες, καθεμία ένα ατελές υποκατάστατο του ανθρώπινου βλέμματος.
> Αρκετές εμφανίζονται εδώ. Το **FID** συγκρίνει τη συνολική στατιστική «γεύση» μιας ολόκληρης παρτίδας παραγόμενων εικόνων με πραγματικές· χαμηλότερο είναι καλύτερο, αλλά περιγράφει την παρτίδα, όχι μία εικόνα. Το **CLIP score** ρωτά μια ξεχωριστή AI αν η εικόνα ταιριάζει στις λέξεις του prompt — χρήσιμο, αλλά τόσο καλό όσο ο κριτής. Τα **DINO**, **SSIM** και **LPIPS** συγκρίνουν μια ανακατασκευασμένη εικόνα με έναν στόχο, σε επίπεδα που κυμαίνονται από ωμά pixels έως μαθημένα χαρακτηριστικά. Κανένα από αυτά δεν είναι η αλήθεια. Είναι proxies και συνήθως κινούνται σε μικρές διαφορές. Όταν διαβάζετε ότι ένα μοντέλο παίρνει 127.6 ενώ ένα άλλο 128.8, αυτό είναι πραγματική διαφορά προς την επιθυμητή κατεύθυνση — αλλά είναι μικρή μετατόπιση, όχι κατολίσθηση, και μάλιστα σε έναν αριθμό που μόνο χαλαρά ακολουθεί όσα θα έλεγε το μάτι σας. Αξίζει να το θυμόμαστε όταν ο τίτλος είναι «νικά μοντέλο εκπαιδευμένο με είκοσι φορές περισσότερα δεδομένα».

Τι έκαναν οι συγγραφείς

Το πρόβλημα που θέλησαν να διορθώσουν ήταν ακριβώς η τυφλή σχεδίαση. Τα υπάρχοντα μοντέλα αντιμετωπίζουν τη συγγραφή SVG ως καθαρά γλωσσική εργασία: προβλέπουν το επόμενο κομμάτι κώδικα από τον κώδικα που έχει γραφτεί μέχρι εκείνη τη στιγμή και δεν τον αποδίδουν ποτέ. Έτσι μένουν αχρησιμοποίητα τα ισχυρά «μάτια» — ο vision encoder — που ήδη διαθέτουν τα σύγχρονα πολυτροπικά μοντέλα. Το Render-in-the-Loop αναδομεί την εργασία ως βηματική οπτική διαδικασία. Μετά από κάθε τμήμα κώδικα σχεδίασης, το μερικό SVG αποδίδεται ως εικόνα και τροφοδοτείται ξανά στο μοντέλο, ώστε το επόμενο τμήμα να επιλέγεται ενώ το μοντέλο κοιτά πραγματικά τον καμβά μέχρι εκείνη τη στιγμή.

Το πρώτο τους εύρημα είναι προειδοποιητικό και, προς τιμήν τους, το αναφέρουν καθαρά: το να κολλήσεις απλώς αυτόν τον βρόχο σε ένα υπάρχον έτοιμο μοντέλο δεν λειτουργεί. Όταν έδωσαν ενδιάμεσες αποδόσεις σε ισχυρά γενικά μοντέλα χωρίς ειδική εκπαίδευση, η ποιότητα δεν βελτιώθηκε — *χειροτέρεψε* παντού. Ένα μοντέλο που δεν διδάχθηκε ποτέ να χρησιμοποιεί τα μάτια του γι' αυτή την εργασία δεν ξέρει ξαφνικά πώς να τα χρησιμοποιήσει.

Άρα το μεγαλύτερο μέρος της δουλειάς βρίσκεται στη διδασκαλία. Ξαναφτιάχνουν τα

δεδομένα εκπαίδευσης έτσι ώστε κάθε σχέδιο να χωρίζεται σε πολλά μικρά, οπτικά ουσιαστικά βήματα — διασπώντας σύνθετα σχήματα σε απλούστερα κομμάτια ώστε σε κάθε στάδιο να υπάρχει κάτι νέο να δει — και έπειτα κάνουν fine-tuning σε ένα ανοικτό μοντέλο οκτώ δισεκατομμυρίων παραμέτρων (βασισμένο στο Qwen3-VL) πάνω σε αυτές τις βηματικές ακολουθίες. Το ονομάζουν Visual Self-Feedback. Προσθέτουν δεύτερο μηχανισμό κατά τη σχεδίαση, Render-and-Verify: πριν δεχθεί κάθε νέα γραμμή, το μοντέλο την αποδίδει και ελέγχει αν όντως άλλαξε την εικόνα ή απλώς επανέλαβε την προηγούμενη. Οι γραμμές που δεν προσθέτουν τίποτα απορρίπτονται και, όταν τίποτα άλλο δεν βοηθά, ζητείται από το μοντέλο να σταματήσει. Αξιοσημείωτα, όλα αυτά εκπαιδεύονται σε σχετικά μικρό dataset — περίπου 850.000 παραδείγματα, λιγότερο από το μισό ενός ανταγωνιστή και μικρό κλάσμα ενός άλλου.

Τι βρήκαν

- Εκπαιδευμένο έτσι, το μοντέλο σχεδιάζει καλύτερα από την τυφλή εκδοχή του — πιο εμφανώς στις αποτυχίες, όπου ένα τυφλό μοντέλο βάζει ένα μάτι στο μάγουλο ή παραλείπει το ζητούμενο ραβδόγραμμα και βγάζει έναν γενικό monitor.
- Στο καθιερωμένο benchmark (MMSVGBench), η μέθοδος είναι ανταγωνιστική και σε αρκετές μετρήσεις λίγο καλύτερη από ισχυρούς αντιπάλους — μεταξύ τους το OmniSVG, εκπαιδευμένο με πάνω από διπλάσια δεδομένα, και το InternSVG, εκπαιδευμένο με περίπου είκοσι φορές περισσότερα.
- Τα περιθώρια είναι μικρά. Στο σύνολο εικονιδίων, για παράδειγμα, η κύρια βαθμολογία ποιότητας εικόνας είναι 127.6 έναντι 128.8 του καλύτερου ανταγωνιστή και η βαθμολογία αντιστοίχισης με το prompt 0.293 έναντι 0.291 — πραγματικά και συνεπή, αλλά λεπτά κέρδη.
- Και τα δύο πρόσθετα κομμάτια δικαιολογούν την ύπαρξή τους: αν απενεργοποιηθεί η ειδική εκπαίδευση ή η επαλήθευση κατά τη σχεδίαση, οι αριθμοί πέφτουν μετρήσιμα. Η επαλήθευση ειδικά εμποδίζει το μοντέλο να κολλά ξανασχεδιάζοντας το ίδιο πράγμα ξανά και ξανά.
- Το αποτέλεσμα που τονίζουν οι ίδιοι οι συγγραφείς είναι η αποδοτικότητα — να φτάνεις τόσο μακριά με πολύ λιγότερα δεδομένα εκπαίδευσης από όσα χρησιμοποίησαν οι πρωτοπόροι.

Τι δεν αποδεικνύει

- **Δεν** δείχνει ότι το να αφήσεις ένα μοντέλο να «βλέπει» είναι δωρεάν κέρδος. Στην πραγματικότητα το αντίθετο: το ίδιο το πείραμα δείχνει ότι οπτική ανάδραση *χωρίς* επανεκπαίδευση κάνει τα πράγματα χειρότερα. Το κέρδος προέρχεται από την επανεκπαίδευση, όχι μόνο από τα μάτια.
- **Δεν** αποδεικνύει μεγάλο ή αποφασιστικό προβάδισμα. Στις περισσότερες βαθμολογίες η μέθοδος βρίσκεται σχεδόν ισόπαλη με τους αντιπάλους· το «νικά μοντέλο με 20×

περισσότερα δεδομένα» είναι αληθινό, αλλά με μικρές μετατοπίσεις σε proxy metrics, σε ένα benchmark.
- **Δεν** αποδεικνύει γενική καλλιτεχνική ικανότητα. Πρόκειται για ερευνητικό μοντέλο οκτώ δισεκατομμυρίων παραμέτρων που σχεδιάζει εικονίδια και απλές εικονογραφήσεις σε μικρή σταθερή ανάλυση (224×224 pixels), όχι σχεδιαστή γενικής χρήσης.
- Η σύγκριση με ένα μεγάλο γενικό μοντέλο (GPT-5) **δεν** είναι apples-to-apples: εκείνο το μοντέλο δεν είναι κατασκευασμένο ή ρυθμισμένο για αυτή τη στενή εργασία σχεδίασης με κώδικα, άρα η νίκη εδώ λέει λίγα για οποιοδήποτε από τα δύο γενικά.
- **Δεν** είναι χωρίς κόστος. Η απόδοση και εκ νέου ανάγνωση του καμβά σε κάθε βήμα κάνει τη δημιουργία πιο αργή από το να εκπέμπεται όλος ο κώδικας σε ένα τυφλό πέρασμα — κόστος που οι συγγραφείς αναγνωρίζουν.

Πόσο ισχυρά είναι τα στοιχεία

- **Στέρεα** εκεί όπου πρόκειται για ελεγχόμενη σύγκριση με τους δικούς της όρους. Τα ablations είναι καθαρά: αφαιρείς την εκπαίδευση ή την επαλήθευση και οι αριθμοί πέφτουν — άρα τα δύο συστατικά πράγματι κάνουν τη δουλειά που ισχυρίζονται οι συγγραφείς.
- **Έντιμα ως προς τη δική τους έκπληξη.** Το εύρημα ότι η αφελής οπτική ανάδραση βλάπτει αναφέρεται και δεν θάβεται, και είναι ίσως το πιο ενδιαφέρον σημείο της εργασίας — χρήσιμη διόρθωση της διαίσθησης ότι περισσότερη είσοδος είναι πάντα καλύτερη.
- **Λεπτά ως προς τον ισχυρισμό leaderboard.** Οι νίκες απέναντι σε ανταγωνιστές με περισσότερα δεδομένα είναι μικρές και βρίσκονται σε ένα benchmark που κατασκευάστηκε από τους συγγραφείς ενός από αυτούς τους ανταγωνιστές. Μικρά περιθώρια σε proxy metrics σε ένα test set είναι ενδεικτικά, όχι οριστικά.
- **Αδοκίμαστα σε μεγάλη κλίμακα και στον πραγματικό κόσμο.** Όλα εδώ είναι εικονίδια και απλές εικονογραφήσεις χαμηλής ανάλυσης. Το αν η ίδια ιδέα ισχύει για σύνθετη, υψηλής ανάλυσης ή πραγματική σχεδιαστική εργασία μένει για το μέλλον — οι συγγραφείς το λένε ρητά.

Γιατί έχει σημασία

Η ιδέα στο κέντρο της εργασίας είναι σχεδόν ντροπιαστικά απλή και φτάνει πολύ πέρα από το σχέδιο. Αν ένα πρόγραμμα πρόκειται να δημιουργήσει κάτι γράφοντας κώδικα — μια ιστοσελίδα, ένα γράφημα, ένα διάγραμμα, μια τρισδιάστατη σκηνή — μπορεί είτε να γράψει όλο το πράγμα στα τυφλά και να ελπίζει είτε να το αποδίδει όσο προχωρά και να διορθώνει πορεία. Το κλείσιμο αυτού του βρόχου είναι δεύτερη φύση για τον άνθρωπο· κοιτάζουμε συνέχεια τη σελίδα. Για αυτά τα μοντέλα είναι εκπληκτικά πρόσφατη κίνηση.

Αυτό που κάνει την εργασία άξια ανάγνωσης είναι ο αστερίσκος που προσθέτει. Το να δώσεις σε ένα μοντέλο τρόπο να βλέπει δεν είναι το ίδιο με το να του μάθεις να κοιτάζει. Τα μάτια πρέπει να εκπαιδευτούν, και μόνο τότε αποδίδει ο βρόχος — και ακόμη τότε το

κέρδος είναι πραγματικό αλλά μετρημένο: πιο σταθερός σχεδιαστής, όχι διαφορετικό είδος καλλιτέχνη. Για όποιον φτιάχνει εργαλεία που μετατρέπουν μια περιγραφή σε επεξεργάσιμο γραφικό — το είδος που καταλήγει πίσω από εικονίδια και εικονογραφήσεις εφαρμογών — το ήσυχο πρακτικό μάθημα είναι το χρήσιμο. Το κέρδος εδώ ήρθε από φθηνότερη, εξυπνότερη εκπαίδευση, όχι από περισσότερα δεδομένα. Αυτό είναι καλύτερο μάθημα από ακόμη έναν πόντο σε leaderboard.

Καθαρή σύνοψη

Τα γλωσσικά μοντέλα που παράγουν διανυσματικά γραφικά — τον επεξεργάσιμο, κλιμακώσιμο κώδικα πίσω από τα περισσότερα web icons και λογότυπα — παραδοσιακά το κάνουν «στα τυφλά», γράφοντας όλες τις εντολές σχεδίασης χωρίς ποτέ να αποδίδουν την εικόνα για να δουν το αποτέλεσμα. Μια ομάδα με επικεφαλής τον Guotao Liang προτείνει το Render-in-the-Loop: μετά από κάθε βήμα αποδίδει το ημιτελές σχέδιο και το τροφοδοτεί ξανά, ώστε το μοντέλο να ζωγραφίζει την επόμενη γραμμή κοιτάζοντας τον καμβά. Το κεντρικό και έντιμο εύρημα είναι ότι αν το κάνεις απλώς σε ένα υπάρχον μοντέλο, γίνεται *χειρότερο*· το κέρδος εμφανίζεται μόνο αφού το μοντέλο επανεκπαιδευτεί ώστε να χρησιμοποιεί την οπτική ανάδραση, βοηθούμενο από έλεγχο κατά τη σχεδίαση που απορρίπτει γραμμές που δεν αλλάζουν τίποτα. Το επανεκπαιδευμένο μοντέλο οκτώ δισεκατομμυρίων παραμέτρων ισοφαρίζει ή ξεπερνά ελαφρά ανταγωνιστές με έως είκοσι φορές περισσότερα δεδομένα σε ένα καθιερωμένο benchmark — πραγματικό αποτέλεσμα, αλλά με μικρά περιθώρια σε proxy scores, για εικονίδια και απλές εικονογραφήσεις χαμηλής ανάλυσης. Το συμπέρασμα που τονίζουν οι συγγραφείς και αξίζει να μείνει αφορά την αποδοτικότητα: το να βλέπεις καλά τη δουλειά σου μπορεί να υποκαταστήσει μέρος της ακατέργαστης κλίμακας δεδομένων.

Έλεγχος χωρίς υπερβολές

Τι δείχνει η εργασία: Η επανεκπαίδευση ενός μοντέλου vector graphics ώστε να αποδίδει και να κοιτά το μισοτελειωμένο σχέδιό του βήμα-βήμα παράγει καλύτερα και πληρέστερα αποτελέσματα από την τυφλή σχεδίαση — ανταγωνιστικά και ελαφρά καλύτερα από αντιπάλους με περισσότερα δεδομένα σε ένα καθιερωμένο benchmark, με λιγότερα δεδομένα εκπαίδευσης.

Τι είναι εύλογο αλλά δεν έχει αποδειχθεί: Ότι το «να βλέπεις τη δουλειά σου» είναι γενικά καλύτερη συνταγή από την κλιμάκωση δεδομένων· ότι η ίδια ιδέα θα βοηθήσει σε σύνθετα, υψηλής ανάλυσης ή πραγματικά γραφικά· ότι τα μικρά περιθώρια στο benchmark αντανακλούν διαφορά που θα παρατηρούσε ένας άνθρωπος.

Τι δεν δείχνει: Ότι η οπτική ανάδραση βοηθά από μόνη της (χωρίς επανεκπαίδευση βλάπτει)· ότι το προβάδισμα έναντι των αντιπάλων είναι μεγάλο ή αποφασιστικό· γενική ικανότητα σχεδίασης· δίκαιη σύγκριση με γενικά μοντέλα όπως GPT-5 που δεν έχουν κατασκευαστεί γι’

αυτή την εργασία.

Κύριοι περιορισμοί: Ένα benchmark, χτισμένο εν μέρει από συγγραφείς ανταγωνιστικού μοντέλου· μικρά περιθώρια σε proxy metrics· μοντέλο 8B, εικονίδια και απλές εικονογραφήσεις στα 224×224· πιο αργή δημιουργία λόγω του βρόχου απόδοσης σε κάθε βήμα.

Πόση εμπιστοσύνη πρέπει να έχει ένας γενικός αναγνώστης; Υψηλή ότι το κλείσιμο του βρόχου — να αποδίδει και να ξαναδιαβάζει καθώς σχεδιάζει — πράγματι βοηθά και ότι πρέπει να εκπαιδευτεί, όχι απλώς να ενεργοποιηθεί. Χαμηλή έως μέτρια ότι το συγκεκριμένο μοντέλο νικά αποφασιστικά τους αντιπάλους· η φράση «νικά 20× τα δεδομένα» είναι πραγματική αλλά μέτρια, από ένα benchmark. Υψηλή για τη μία ιδέα που αξίζει να θυμόμαστε: για κώδικα που σχεδιάζει, το να κοιτάζεις καθώς προχωράς είναι καλύτερο από το να σχεδιάζεις στα τυφλά.

Πηγές

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/`