VISUAL CHAIN OF THOUGHT AS A RESOURCE-BOUNDED MACHINE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal models now produce intermediate images and read them back through a connector. Text chain of thought theory counts only tokens, so the visual loop therefore lacks a formal account. In this paper, we model visual chain of thought as a resource bounded machine with two operators. A paint writes a pixel grid through a constant depth local decoder, and a read returns a bounded number of tokens through a connector encoder. Every bound is governed by three quantities, the pixels written at each step, the read bandwidth of the connector, and the state entropy the next step must observe. Storing one incompressible working image in text requires a quadratic number of tokens, while a single paint produces the same image in one step. Because a local decoder of constant radius implements one cellular automaton update, source to target connectivity on an open grid admits a linear number of paints, whereas a packed parent pointer transcript remains faster than linear in the diameter. When the read bandwidth falls below the state entropy, native regeneration cannot reconstruct the tape, and answers on the language side no longer depend on the image, so the intermediate image is decorative. The same bound determines what to draw, and the working image is a sufficient statistic of the state at low entropy. Frozen open weight vision language models reproduce this separation on low entropy queries. Accuracy stays at the text baseline until the connector retains enough tokens, and a corrupted render or token stream changes the answer. Larger models cross the same threshold at a smaller token budget.

1 INTRODUCTION

Chain of thought is the intermediate text that a decoder only transformer writes before its final answer (Nye et al., 2022; Wei et al., 2022). For log precision models of constant depth, that text changes the circuit class. Without it, such models lie in uniform TC^0, the class of constant depth threshold circuits (Merrill & Sabharwal, 2023b;a; Strobl et al., 2024). With T text tokens they recover serial depth. Boolean circuits of size T sit inside text chain of thought (Li et al., 2024), and polynomial steps under generalised prenorm characterise polynomial time (Merrill & Sabharwal, 2024). Hard attention lower bounds still require the number of text steps to grow with the input, even for some problems in TC^0 (Bavandpour et al., 2025). All of these theorems nevertheless count tokens, and each forward pass outputs one vocabulary symbol. A visual step falls outside that accounting, because one decoder call can update an entire grid while the language head may see only a few connector tokens.

Once that step writes an image, systems store it in different places. Visual Sketchpad and Whiteboard of Thought keep an external canvas that can be read again (Hu et al., 2024; Menon et al., 2024). Multimodal visualisation of thought and ThinkMorph instead write frames that hold spatial state absent from the text (Li et al., 2025a; Gu et al., 2026). OpenAI o3 and ReFocus edit inside that loop, and InterSketch also calls tools on the sketch (OpenAI, 2025; Fu et al., 2025; Ning et al., 2026). Closer to text, LLaVA-CoT keeps successive text stages (Xu et al., 2025), while other work inserts image regions or grounded object tokens into that rationale (Gao et al., 2025; Li et al., 2025b). Latent visual tokens avoid a pixel grid (Yang et al., 2026). See2Think compares text only answers, visual tokens without a faithful render, the full loop, and a corrupted render, and finds that a relevant visual action often leaves the answer unchanged (Yan et al., 2026). Single pass generation bounds do not iterate that image as memory (Ke et al., 2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

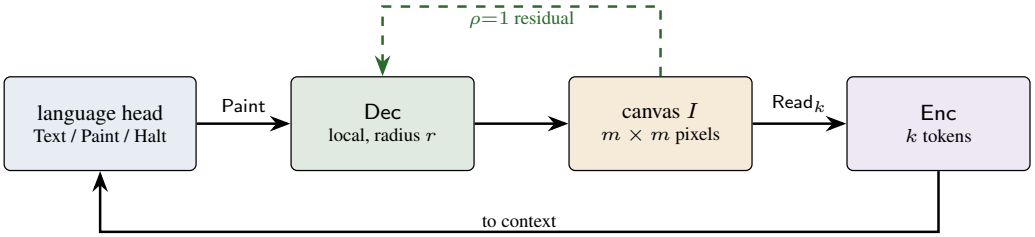


Figure 1. One round of the visual working tape. The local decoder writes the canvas, and the connector reads k tokens back into the context.

This paper therefore treats interleaved vision language chain of thought as a resource bounded machine, aligned with Merrill & Sabharwal (2024) except for the two operators in the loop, as shown in Figure 1. Paint produces an $m \times m$ image through a constant depth local decoder (Dec). Read_k returns k tokens through a connector encoder (Enc). Let B_{read} denote the resulting read bandwidth, which satisfies

$$B_{\mathrm{read}} \;\le\; k\, C_q \tag{1}$$

where C_q is the per token capacity. The residual bit ρ distinguishes an external canvas that keeps pixels (ρ=1) from native regeneration that must reencode the whole image (ρ=0). It is the canvas half of that resource distinction. The read bandwidth in equation 1 is the connector half. Together they decide whether a paint is a working tape or an image that the language head never sees.

Three quantities then control every bound below. They are the number of pixels produced per step, the read bandwidth in equation 1, and the state entropy $I_{\mathrm{state}} = H(S)$ of the task state S that the next step must see. The information theoretic floor, the light cone depth, and the text token lower bound are

$$k^{\star}_{\mathrm{lb}} \;=\; \lceil I_{\mathrm{state}}/C_q \rceil, \qquad t^{\star} \;=\; \lceil D/r \rceil, \qquad T \;\ge\; \Omega(t^3/B), \tag{2}$$

together with a rate distortion rule. Here D is the dependency diameter, r the decoder neighbourhood radius, t the number of local iterations, and B is the number of cell layer pairs a text step may name. The working image should be a sufficient statistic of the state at rate B_{read} rather than a photograph (Theorem 9). The first identity is an information theoretic floor $k^{\star}_{\mathrm{lb}}$. A particular encoder and head may become reliable only at the class reachability point $k_{\mathrm{reach}} = c_{\mathcal{E}}\, k^{\star}_{\mathrm{lb}}$ (Proposition 1), with encoder class factor $c_{\mathcal{E}} \ge 1$.

Polynomial text chain of thought is already P (Merrill & Sabharwal, 2024), so a visual tape does not escape P. What changes is the arity of a forward pass. A text token carries $O(\log |V|)$ bits, while one paint produces $m^2 \log |\Gamma|$ bits that are already laid out for a local rule. Storing one incompressible working image on a text tape therefore costs $\Omega(m^2)$ tokens, while one paint produces it in a single step (Theorem 1). A radius r decoder implements one two dimensional cellular automaton step, so open grid source to target connectivity is $O(m)$ paints by breadth first search (Theorems 2 and 3). Occupancy materialising text search pays $\Theta(m^2)$ tokens. Packing parent pointers into a polynomial vocabulary still pays $\Theta(m^2/\log m)$ (Theorem 3). A snake of length $\Theta(m^2)$ still needs $\Theta(m^2)$ rounds in both models. The visual tape is a different output arity inside P, and it is not a substitute for a calculator (Theorem 10).

When $B_{\mathrm{read}} < I_{\mathrm{state}}$, a generator without residual pixels cannot reconstruct the tape (Theorem 6), and language side answers A satisfy $I(A; R \mid \mathrm{Txt}) \le B_{\mathrm{read}}$ (Theorem 7), where R is the read back and Txt the text transcript. If the text already determines the answer, intervening on the image cannot move it, so the sketchpad is then decoration. This is a complexity account of the See2Think protocol, and it is the content of equation 2 on the read side. Uptake, the event that the text mentions the image, is not mutual information with the answer. The same task also splits on the residual bit. Residual breadth first search finishes on the canvas at $k = O(1)$, while the same local rule without residual pixels needs $k = \Omega(m^2/C_q)$ (Theorem 8).

This paper makes three contributions.

- A visual working tape with Paint, Read_k, and residual flag ρ, together with an unconditional simulation cost and a write arity tradeoff $T = \Omega(t^3/B)$, so one paint produces the same grid in one forward pass.

- A decoration threshold with information theoretic floor $k^{\star}_{\mathrm{lb}}$ and encoder class reachability k_{reach}, together with a residual separation on connectivity under which a residual canvas stays correct at $k = O(1)$ while native regeneration collapses.
- Resource slopes on executable constructions, and a frozen open weight vision language scan under the four See2Think settings, with a token path intervention that flips the answer at fixed pixels once the head can represent the query.

2 Related Work

Text chain of thought. Log precision transformers without intermediate tokens lie in uniform TC^0 (Hahn, 2020; Merrill & Sabharwal, 2023b;a; Strobl et al., 2024). Constant bit precision tightens that inclusion to AC^0, the class of constant depth circuits built from AND and OR gates (Li et al., 2024). Intermediate text recovers serial depth. Circuits of size T sit inside T step text chain of thought (Li et al., 2024), and polynomial steps under generalised prenorm characterise P (Merrill & Sabharwal, 2024). The same device lets a bounded transformer construct arithmetic and dynamic programs that would otherwise require superpolynomial size (Feng et al., 2023). Later bounds remain on this token tape. Hard attention still requires a problem dependent number of text steps (Bavandpour et al., 2025), and length generalisation of learned chain of thought is an optimisation guarantee for acquiring those steps (Huang et al., 2026). Infinite precision with unbounded steps is Turing complete, and surveys of formal languages place that regime outside log precision decoding (Pérez et al., 2019; Strobl et al., 2024). Looped latent thought replaces the tape by one hidden vector per round (Xu & Sato, 2026). Each of these models therefore writes either a one dimensional string or a single global state.

Visual intermediate steps. The same serial resource becomes an image once the trace is multimodal, but published systems still differ in where that image is stored. Visual Sketchpad and Whiteboard of Thought keep an external canvas, and both find that an abstract drawing can stand in for a photograph (Hu et al., 2024; Menon et al., 2024). Multimodal visualisation of thought and visual planning write the next frame from the model itself (Li et al., 2025a; Xu et al., 2026), and ThinkMorph argues that the frame should carry spatial state that the text does not copy (Gu et al., 2026). Editing sits between these two designs. ReFocus changes the current image and then continues from the edited view (Fu et al., 2025), while InterSketch also calls deterministic tools on the sketch (Ning et al., 2026). Other methods leave most of the chain in text. Staged rationales, region crops, and grounded object tokens insert only a fragment of the image (Xu et al., 2025; Shao et al., 2024; Gao et al., 2025; Li et al., 2025b). Interleaved visual tokens, large mixed traces, and multitask fusion of visual evidence extend that fragment (Cheng et al., 2026; Chen et al., 2026; Li et al., 2026; Lou et al., 2026), and latent visual tokens drop the pixel grid (Yang et al., 2026). These designs report task accuracy. See2Think finds that a relevant visual action often leaves the answer unchanged, and that protocol is the empirical content of Theorem 7 (Yan et al., 2026). A circuit bound on one visual autoregressive pass does not supply the missing account, because the generated image is not then read again as memory (Ke et al., 2025).

Grid reachability. The spatial tasks used to test those systems already have a classical complexity account, and that account identifies what a visual step saves. Undirected grid source to target reachability is in log space (Reingold, 2008) and is hard for logarithmic depth circuits under AC^0 reductions (Allender et al., 2009). A zero step log precision transformer therefore cannot decide it when $\mathsf{TC}^0 \neq \mathsf{NC}^1$, while a polynomial text chain of thought can. Communication lower bounds for connectivity do not carry over, because attention already grants random access to the input. The resource that remains is the cost of storing the occupancy map, whether as m^2 text tokens, as a packed spanning tree, as $O(D)$ local paints, or as a $\log m$ bit pose that walks for $\Theta(m^2)$ steps. Merrill & Sabharwal (2024) and Li et al. (2024) count forward passes that emit one token, while the visual machine counts forward passes that emit m^2 pixels. Once that arity is named, the simulation cost in Theorem 1 is an entropy calculation and does not require a new circuit separation. A looped hidden state updates one global vector and can climb TC^k (Xu & Sato, 2026), while a residual canvas updates m^2 coupled cells and realises a two dimensional automaton. Packing a canvas into that vector pays the quadratic token cost of Theorem 1 in reverse. This paper measures tokens and pixels inside the loop of Figure 1, and accuracy only shows that a head has crossed a bandwidth threshold or that a local rule has been implemented.

3 THE VISUAL WORKING TAPE

The bounds below are stated for one machine, so the operators and the budget are fixed before any theorem. The base model is a decoder only transformer $\mathcal{T}$ of constant depth L, embedding width d, and arithmetic precision $p = O(\log n)$, as in Merrill & Sabharwal (2024). The vocabulary V is polynomial, and one text token therefore carries $O(\log n)$ bits. One forward pass on a context $X_{1:t}$ produces a last position state $h_t \in \mathbb{R}^d$ and, when the action is text, a token $\mathrm{out}_V(h_t) \in V$. The constructions use average hard attention. Softmax attention with a polynomial inverse temperature realises the same primitive at log precision, as recorded in Appendix B.1. Three published facts are used as axioms and restated in Appendix B. Constant depth log precision transformers lie in TC^0. Text chain of thought of T steps contains $\mathsf{SIZE}[T]$. Undirected grid reachability lies in L and is NC^1 hard under AC^0 reductions. A visual step adds two operators around a finite image, and a bit that says whether the pixels survive the step.

Definition 1 (Working image). *A working image is an array $I \in \Gamma^{m\times m}$ over a finite palette with $|\Gamma| \geq 2$.*

Definition 2 (Local decoder and Paint). *A spatial decoder* Dec *assigns to each output pixel q a map $I^{\mathrm{new}}(q) = f_q\big(I^{\mathrm{prev}}|_{B_r(q)},\, b_t\big)$. Here $B_r(q)$ is the radius $r = O(1)$ neighbourhood of q, and b_t is an $O(\log n)$ bit broadcast from h_t. Each f_q is a constant depth threshold circuit on that neighbourhood. The residual bit $\rho \in \{0,1\}$ records whether the neighbourhood is read from the previous canvas (ρ=1) or must be rebuilt from tokens already returned (ρ=0). One call to* Paint *runs* Dec *once and writes every pixel in parallel. The decoder is not given search, shortest paths, or connectivity as a primitive.*

One paint is identified with one cellular automaton update only in Theorem 2. That claim is a construction. The definition itself only fixes the receptive field, the broadcast, and the residual bit.

Definition 3 (Read_k and bandwidth). *A task agnostic encoder* Enc *maps $\Gamma^{m\times m}$ to $\mathbb{R}^{k\times d}$ and appends k tokens. After p bit quantisation and projection onto a u dimensional language reachable subspace, the per token capacity C_q is an encoder parameter and satisfies*

$$C_q \;\leq\; \min\big\{pu,\, pd,\, \tfrac{u}{2}\log_2(1 + P/\sigma_0^2)\big\}, \qquad B_{\mathrm{read}} \;=\; kC_q. \tag{3}$$

Any later computation that sees only those k tokens satisfies $I(R; I) \leq B_{\mathrm{read}}$ by the data processing inequality.

The capacity C_q is a property of the encoder, and it is not a universal constant. Uniform eight bit scalar tokens give the textbook value $C_q^{\mathrm{tb}} = 8$, so a one bit query has floor $k_{\mathrm{lb}}^\star = 1$. A connector of restricted rank, at unit signal to noise, can be larger. A frozen vision transformer, probed linearly on a visited map, never reaches Hamming distance 0.05, and that encoder class therefore carries much less than eight bits per token. A frozen language head on a half plane query defines an effective capacity $C_q^{\mathrm{LM}} = I_{\mathrm{state}}/k_{\mathrm{emp}}$ that grows with scale, where k_{emp} is the accuracy and flip crossing measured in Section 5. The floor $k_{\mathrm{lb}}^\star$ is the least budget at which some encoder can carry the state S. A named encoder and head may become reliable only at $k_{\mathrm{reach}} = c_{\mathcal{E}}\, k_{\mathrm{lb}}^\star$. Every statement below is parametric in C_q, and every numerical claim names the encoder class that produced it.

Definition 4 (Complexity classes). $\mathrm{TextCoT}[T]$ *is the Merrill and Sabharwal text chain of thought machine, with t_{img}=0 and at most T text tokens.* $\mathrm{VisualCoT}[t, k, \rho]$ *is visual chain of thought with t paints, k read back tokens per paint, residual bit ρ, and $O(t)$ control tokens that do not store working state. The visual working tape* $\mathrm{VWT}[T_{\mathrm{txt}}, t_{\mathrm{img}}, k, \rho]$ *allows both tapes. Weights may be nonuniform in the input length. Every constructive upper bound below is logspace uniform.* $\mathrm{TextCoT}_B[T]$ *is the subclass in which each step appends at most B tokens, and each of those tokens names $O(1)$ cell layer pairs.*

The two values of ρ match two designs already used in practice. An external canvas keeps pixels on a renderer (ρ=1). A native generator rebuilds the frame from hidden state (ρ=0). Theorems 6 and 7 separate on that bit. With residual pixels the search can finish on the canvas while the language head reads only a short R. Without residual pixels, discarding the canvas leaves the next local step with the returned tokens alone.

Assumption 1 (Threat model). ***(A0)*** *The nonlocal input to* Dec *is $O(\log n)$ bits, and it cannot carry an $\omega(\log n)$ bit summary of the tape.* ***(A1)*** Enc *does not run the task.* ***(A2)*** *The ρ=1 canvas executes*

pixel updates and nothing else. Specialist tools form a separate tool tape (Theorem 10). ***(A3)*** *The input grid is already in context, and the working tape stores intermediate state.*

Assumptions A0 to A3 keep three channels outside the budget. Those channels are a global decoder broadcast, an encoder that solves the task, and a renderer that hides search. The assumptions name the machine whose pixels and tokens are counted.

One interleaved round then uses three tapes. The input tape holds the instance. The text tape receives one symbol per text action. The visual tape is the image I, overwritten in place by Paint. The language head reads the current context and emits a text token, a paint, or a halt. A paint runs Dec at every site in parallel, writes I^{new}, and immediately calls Read_k, whose k tokens are appended to the context. If $\rho{=}1$, the canvas keeps I^{new} for the next neighbourhood. If $\rho{=}0$, that neighbourhood must be rebuilt from the tokens. Control tokens that choose the action are $O(1)$ per round, and they are not allowed to store an occupancy map, since $O(t)$ such tokens carry only $O(t \log n)$ bits (Appendix C). A full run is therefore charged $(T_{\mathrm{txt}}, t_{\mathrm{img}})$ together with the per paint parameters (k, ρ, r, m). The sum of text steps and paints counts forward passes. The accounting that matters is which alphabet was written, and whether the read was wide enough to see the next state.

4 Resource Bounds

The machine of Section 3 has a write budget and a read budget, and the two are not interchangeable. The subsections below take them in that order, from the cost of writing a grid, through the light cone and the lossy read, to the image that is worth painting.

4.1 Simulation cost and local parallelism

A visual tape changes how many bits one forward pass may emit. The next theorem is the token cost of storing an incompressible grid in text, together with the matching fact that one Paint writes that grid in a single step.

Theorem 1 (Simulation cost)**.** *Let I be a working image with min entropy $H_\infty(I) \geq \eta m^2 \log|\Gamma|$. Any lossless text encoding then uses $T \geq \eta m^2 \log|\Gamma| / \log|V|$ tokens, while one* Paint *produces I in a single step. If Hamming distortion εm^2 is allowed on a binary palette, then $T \log|V| \geq \big(1 - h_2(\varepsilon)\big) m^2$, where h_2 is binary entropy. A sequence of t frames whose conditional entropy is at least βm^2 at each step needs $\Omega(t m^2)$ text tokens and only t paints.*

The lossless clause is the chain rule. The lossy clause is the rate distortion function of a fair Bernoulli source under Hamming distortion, and the frame clause is the same identity along a trajectory. A binary vocabulary makes the bound tight up to the factor η. Polynomial text chain of thought can still store the image, and it therefore remains inside P, but it pays a quadratic number of tokens for every frame.

Theorem 2 (One paint is one automaton step)**.** *There is a constant depth local decoder, using average hard attention and a finite lookup, such that one* Paint *implements an arbitrary radius r two dimensional cellular automaton update on $\Gamma^{m \times m}$. The construction is translation equivariant and logspace uniform in m.*

Each output site attends to its $(2r{+}1)^2$ neighbours by matching positions, concatenates their colours, and evaluates the local rule δ by a constant depth multilayer perceptron of width $O\big(|\Gamma|^{(2r+1)^2}\big)$. Residual reading takes the neighbourhood from the previous canvas. Reading without residual pixels uses the identity encoder at $k = m^2$. A global broadcast of $O(\log n)$ bits from h_t may select which local rule to apply. By assumption A0, that broadcast cannot carry an $\omega(\log n)$ bit summary of the image.

Remark 1 (Finite convolution)**.** *A finite local rule δ is a finite convolution on one hot colours, so the number of parameters does not grow with m. Supervised training on the new frontier of breadth first search fits this class to an F1 score above 0.98 across grid sizes, where F1 is the harmonic mean of precision and recall. A spatially shuffled target and a flattened perceptron both degrade as m grows (Section 5 and Appendix I).*

The same local update decides connectivity when the frontier is wide. On an open grid the diameter is linear, so a layer by layer paint is short, while a text archive of the same layers is quadratic.

Theorem 3 (Open grid reachability). *On an obstacle free $m \times m$ grid, residual visual chain of thought decides source to target connectivity with $t_{\mathrm{img}} = O(D)$ paints and $k = O(1)$, where $D \le 2m$ is the Manhattan diameter. Text search that writes the occupancy map uses $\Omega(m^2)$ tokens. Packing a parent pointer spanning tree into a polynomial vocabulary still uses $\Omega(m^2/\log m)$ tokens. On a width one snake of length $\Theta(m^2)$, both models use $\Theta(m^2)$ rounds.*

The visual upper bound is the usual breadth first layer update. Wall cells stay walls, visited cells stay visited, and an empty cell next to the frontier becomes the next frontier. Theorem 2 implements one layer per paint, and a single read of the target colour decides the answer. An optional parent channel stores a direction when a cell first joins the frontier, and that channel can be followed in $O(D)$ further paints if a path must be drawn. These text lower bounds apply to algorithms that name cells. They do not apply to every machine in $\mathrm{TextCoT}[\mathrm{poly}(m)]$. Polynomial text chain of thought can simulate the visual machine, because it characterises P, but the simulation writes a quadratic transcript for each frame. Wall following with a $\log m$ bit pose lies in log space (Reingold, 2008) and still takes $\Theta(m^2)$ steps on a spiral. The visual saving is therefore the parallel update of a wide front, and it is not a shorter path through the diameter. On a width one snake the front has size one. Both models then pay the length, and the arity of Paint does not reduce the number of rounds.

4.2 Write arity

The same geometry yields a space time tradeoff that does not depend on one dynamic program. The light cone fixes how many paints a local rule needs, and Theorem 4 converts that cone into the token lower bound $T = \Omega(t^3/B)$ for a text machine that names $O(B)$ cells per step.

Lemma 1 (Light cone). *After j paints of radius r, a query cell depends on the initial configuration only inside Manhattan radius rj. Any problem whose dependency set contains a cell at distance rt therefore requires $t_{\mathrm{img}} \ge t$.*

Theorem 4 (Write arity). *Let P_t be a radius r two dimensional cellular automaton, iterated t times and then queried at one cell, whose local rule is nonlinear or noncommutative and sensitive on the light cone. Any text machine that names at most B cell layer pairs per step satisfies $T = \Omega(t^3/B)$. Residual visual chain of thought uses t paints and updates $\Theta(t^2)$ cells in parallel at each step, independently of B.*

The dependency graph has $\sum_{i=0}^{t}(2ri+1)^2 = \Theta(t^3)$ nodes. A sensitivity argument shows that each node in the light cone is a necessary predecessor of the query (Appendix F). Each text step names $O(B)$ new nodes, and therefore $T = \Omega(t^3/B)$. The case $B = O(1)$ recovers a cubic token lower bound. The case $B = \Theta(t^2)$ matches the visual step count, while the total number of named cells remains $\Theta(t^3)$. The visual saving is spatial parallelism inside one forward pass, and the total work is not smaller. Linear rules over the two element field $\mathrm{GF}(2)$ may admit a closed form and are excluded. Folding t iterations into $o(t)$ constant depth threshold circuits would require $\mathsf{TC}^0 = \mathsf{NC}^1$, which is a separate conditional clause (Appendix F.1).

4.3 Decoration, reachability, and residual pairing

Each Read_k returns at most B_{read} bits, so the language head sees a lossy view of the working image. The next three statements turn that budget into a reconstruction barrier when pixels are discarded, a decoration criterion on the language side, and a residual pairing on one connectivity task.

Theorem 5 (Read bandwidth). *If $R = \mathrm{Enc}(I)$, then $I(R;S) \le \min\{B_{\mathrm{read}},\, I(I;S)\}$.*

Theorem 6 (Collapse without residual pixels). *Suppose ρ=0 and the next decoder must recover a state S with $H(S) \ge I_{\mathrm{state}}$. If $kC_q + O(dp) < I_{\mathrm{state}} - 2$, then any estimator of S from (R, h_t) has error at least $1/3$.*

Theorem 7 (Decoration). *Let A be the answer and Txt the text transcript. Then $I(A; R \mid \mathrm{Txt}) \le B_{\mathrm{read}}$. If $H(A \mid \mathrm{Txt}) > B_{\mathrm{read}} + \Delta$, an intervention on the image cannot move A except with error $\Omega(\Delta)$. In particular, if Txt already determines A, a corrupted render leaves the answer unchanged.*

Theorem 5 is data processing on the chain from S through I to R. Theorem 6 is Fano's inequality on the next paint when pixels are not kept. Theorem 7 is the same bound on the language side, and it is the formal content of a decorative sketch. See2Think records whether the text mentions

the image. Mutual information between the read back and the answer is a different quantity. The floor $k^\star_{\mathrm{lb}} = \lceil I_{\mathrm{state}}/C_q \rceil$ is necessary for every encoder, because $I(R;S) \le kC_q$. It is sufficient only when Enc is an optimal code of S on k tokens and the head can decode that code.

Proposition 1 (Encoder class reachability)**.** *For an encoder family $\mathcal{E}$ and a decoder class $\mathcal{G}$ there exists $c_{\mathcal{E}} \ge 1$ such that $k < c_{\mathcal{E}}\, k^\star_{\mathrm{lb}}$ implies error at least $1/3$ on that class. The multiplier $c_{\mathcal{E}}$ is the product of a support hitting factor and a head signal to noise factor (Proposition 2). It is not a new information theoretic constant. If $\mathcal{G}$ cannot represent the task, then $c_{\mathcal{E}} = \infty$.*

Proposition 2 (Sources of $c_{\mathcal{E}}$)**.** *Let N_v patch tokens be retained on an equispaced index set. (a) If a sufficient set U occupies a fraction α of the tokens, a hit requires $k = \Omega(1/\alpha)$. (b) A linear mean of noisy sufficient tokens has signal to noise ratio $\Theta(\alpha\sqrt{k})$ and needs $k = \Omega(1/\alpha^2)$ to cross a fixed threshold. (c) If patches are nearly independent and ρ=0, recovering a $\Theta(N_v)$ bit map still needs $k = \Omega(N_v)$ even when the answer is one bit.*

On a half plane, $\alpha = 1/2$, so both (a) and (b) predict $k = O(1)$. Measured heads on that query still need tens of tokens at the smallest scale, and the excess therefore sits in the head, above the constant budget predicted by the patch fraction. On a random single cell, $\alpha = 1/N_v$, and equispaced retention first hits near $k = 128$ out of $N_v = 196$.

Theorem 8 (Residual pairing)**.** *On the same open grid connectivity task, with the same local rule and the same $t = O(D)$, residual visual chain of thought lies in* $\mathrm{VisualCoT}[O(D), O(1), 1]$*. Any ρ=0 machine with $k = o(m^2/C_q)$ fails to reconstruct the next frontier.*

The residual side is Theorem 3. The side without residual pixels is Theorem 6 applied to a visited map of entropy $\Theta(m^2)$. The statement is a paired gap on one problem. It does not say that residual connections dominate on every geometry. The search may finish on the canvas while the language head reads only one bit, and Theorem 7 still applies to that head. A change to the pixels then breaks the canvas, and a change to the tokens breaks the head. Both interventions are needed to locate the causal path.

4.4 Sketches and incomparability

Once the read has a finite rate, the useful image is the one that spends that rate on the state. A tool and a local paint are cheap on different geometries, so neither replaces the other.

Theorem 9 (Sketch versus photograph)**.** *Fix Read_k. If a photograph is a sketch of S plus independent texture of entropy Ψ, then $I(R;S) \le \min\{B_{\mathrm{read}},\, H(I) - \Psi\}$. The working image that maximises $I(R;S)$ subject to $I(R;I) \le B_{\mathrm{read}}$ is an information bottleneck code of S at rate B_{read}.*

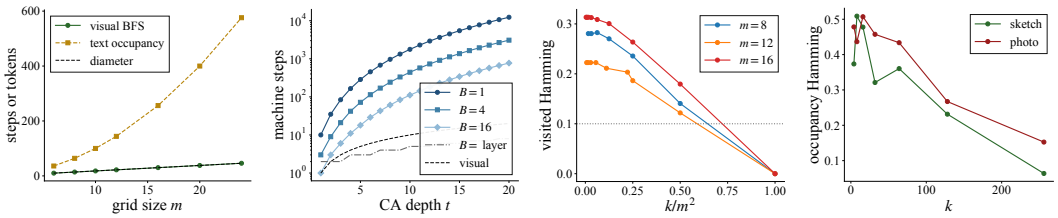
Lines, marks, and occupancy grids are discrete forms of that code. The rendering that the bound selects is a low entropy sufficient statistic of the state.

Theorem 10 (Incomparability)**.** *A tool that returns $O(\mathrm{poly}(\log n))$ bits per call needs $\Omega(m^2/\mathrm{poly}(\log n))$ calls to materialise an incompressible $m \times m$ configuration, which one paint produces in one step. Iterated modular squaring of an n bit integer for t rounds is t tool calls and $\Omega(tn/r)$ local paints. Superlinear visual savings exist on open grids and reverse on width one snakes. Increasing k does not reduce the light cone depth $\lceil D/r \rceil$.*

A visual tape is a local canvas. It is not a calculator, and it is not a routing oracle. If the renderer itself runs the search, the expressivity sits in the tool, and the tape is only the interface allowed by assumption A2.

A sandwich of bandwidth limited cellular automata can be defined by stacking t local maps with a B_{read} bit pipe between them (Appendix H.2). Placing that sandwich in NC^1 uses $\mathsf{TC}^0 \neq \mathsf{NC}^1$. The construction is a conditional bookkeeping layer, and it is not one of the unconditional bounds above.

The identities then have a direct design reading. The floor $k^\star_{\mathrm{lb}}$ is the least k at which some encoder can carry the state. The reachability point k_{reach} is the least k at which a named encoder and head actually do so. The depth $t^\star = \lceil D/r \rceil$ is the least number of local paints that can cover a dependency of diameter D. The write arity bound $T = \Omega(t^3/B)$ is the cost of naming the same light cone in text. The information bottleneck rule is to paint a sufficient statistic of S at rate B_{read}. None of these identities places visual chain of thought outside P. They specify how a forward pass should



(a) Open grid. (b) Write arity. (c) No residual. (d) Sketch and photo.

Figure 2. Executable checks of the resource slopes. Paints track the diameter of an open grid, and text steps track the cost of naming the same light cone. Without residual pixels the read needs many tokens, and a sketch stays easier to recover than a photograph.

spend pixels and tokens inside P. Section 5 measures those quantities on exact rules and on frozen heads, first as slopes of an executable machine and then as crossings of a frozen connector. Exact rules expose the write arity and the residual gap, and frozen heads expose how far a connector sits above the floor.

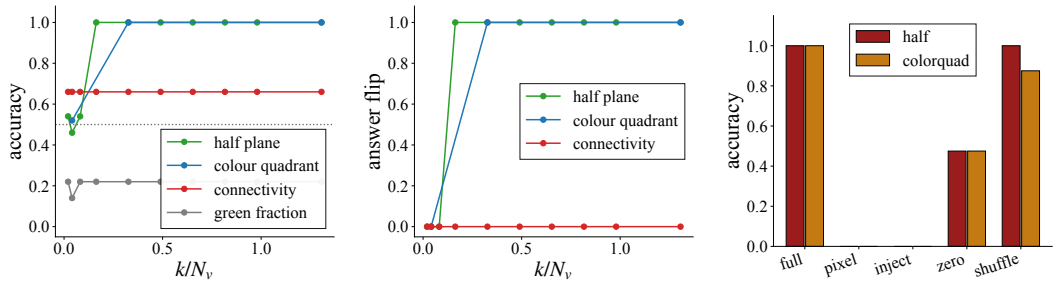
5 Experiments

The theorems predict slopes for an exact rule and crossings for a lossy read. This section checks both. Synthetic machines use exact local rules and simple encoders, so a slope can be read without a trained head. Frozen scans use Qwen3-VL Instruct at the 2B, 4B, and 8B scales, with a 448 pixel side, 196 visual tokens, and fifty instances in each cell (Bai et al., 2025). Every answer is a yes or no decision at the last position. The protocol stays fixed. Only the entropy of the state and the residual flag of the tape change.

5.1 Resource slopes

As shown in Figure 2, one paint is cheaper than a text archive of the same map because the two machines write different amounts in one step. On obstacle free grids the paint count tracks the diameter. At side 24 the visual machine uses 46 paints, text search that stores the occupancy map uses 576 tokens, and a packed parent pointer transcript uses 126 tokens. The packed transcript is the strongest text baseline here, and it still grows as the square of the side divided by the log of the side. At side 24 it remains about ten times the visual step count. Both machines accept the same yes instances. The gap is the cost of writing the map. A copy north cone of depth 20 repeats the comparison at fixed depth. The cone contains 12,341 cell layer nodes. One name per step uses 12,341 text steps, a budget of 16 names uses 772 steps, and naming one whole layer uses 8 steps, while residual paints use 20 steps. Text work divided by the cube of the depth is 1.54, and smaller depths follow the same cube divided by the naming budget, which is the bound in Theorem 4. Visual steps stay at one paint per layer. A text budget large enough to name a layer removes the advantage in step count and leaves the parallel update, since one paint still touches every cell in the layer.

The read side of the same rule splits on residual pixels, again as shown in Figure 2. With residual pixels the colour of the target is already correct at one token, because the canvas keeps the frontier and the head only has to see a one bit readout. Without residual pixels the next local step must rebuild the frontier from the tokens. On sides 8, 12, and 16 the visited Hamming distance stays above 0.10 until the token count equals the number of cells, and at half that budget the distance is still about 0.14. A 16×16 binary map has state entropy 256. At a textbook capacity of 8 bits the information theoretic floor is 32 tokens, yet an orthonormal encoder reaches distance 0.04 only near 224 tokens and reaches zero only when every cell is kept. A corrupted render leaves the answer fixed at 2 tokens and moves it at 256 tokens. The floor is the least budget some encoder can meet, and the later crossing is the oversampling of one encoder class. The sketch panel agrees with the rate distortion bound in Theorem 9. An occupancy map plus independent texture is harder to reconstruct than the map alone, and the sketch curve lies below the photograph curve at every budget. Search without residual pixels therefore needs a quadratic read, because the next local rule has to see a quadratic state, which is the separation in Theorem 8. The same split reappears below, once the state is small enough for a frozen head.



(a) Colour accuracy rises with k. Connectivity stays near chance.

(b) A corrupted render flips colour answers.

(c) A token swap flips the colour answer, and a shuffle does not.

Figure 3. Frozen Qwen3-VL-2B on colour queries and on connectivity. Accuracy, a corrupted render, and a token intervention cross together when the state is one bit. Connectivity stays near the text baseline at every budget.

5.2 Frozen heads and bandwidth crossings

The frozen heads test whether a real connector shows the same crossing once the query is easy enough to solve. See2Think compares text only answers, tokens without a faithful render, the full loop, and a corrupted render (Yan et al., 2026). A further intervention replaces only the pooled tokens and leaves the pixels fixed. The read keeps an equispaced subset of the 196 tokens. A half plane and a colour quadrant are one bit summaries of the image, and maze connectivity on the same geometry is a high entropy control. As shown in Figure 3, accuracy on the colour queries rises with the budget, and connectivity stays near chance.

Table 1 reports the budget at which a solved query begins to use the image. On the 2B head a half plane is at chance up to 8. It is solved at 32, and the colour quadrant agrees. At that budget a corrupt stream and corrupt pixels both flip the answer. A shuffle leaves the colour bag intact. The rate is $1/32$ at 2B, $1/8$ at 4B, and $1/4$ at 8B. One token already suffices, so the extra tokens are a property of the head, as in Proposition 2.

Table 1. Crossings on frozen Qwen3-VL heads.

head	query	k_{emp}	window	C_q^{LM}
2B	half plane	32	$[16, 32]$	0.031
4B	half plane	8	$[4, 8]$	0.125
8B	half plane	4	$[0, 4]$	0.250
2B	colour quadrant	32	$[8, 32]$	0.031

The crossings match the decoration event in Theorem 7. Competence, image use, and the token path meet at one budget when the state is one bit. Connectivity remains near the text baseline even past the full visual token count, so that curve is a ceiling of this head. The information theoretic floor on the half plane is one token, yet the 2B head waits until 32, and the gap falls to 8 and then to 4 with scale. A linear probe on the same encoder never recovers a visited map to Hamming distance 0.05. The connector is therefore wide for a one bit colour bag and narrow for a quadratic occupancy map, which is the split in Theorem 8. An external canvas can finish a local rule at a constant read when the readout is one bit, and it should still be tested by changing the pixels. A native generator must let the token budget track the entropy of the next neighbourhood.

6 Conclusion

Visual chain of thought extends text chain of thought by a visual working tape. One Paint writes an $m \times m$ image through a local decoder, and Read_k returns at most kC_q bits. The residual bit separates a canvas that keeps its pixels from a generator that rebuilds the frame. An incompressible grid then costs a quadratic text transcript and only one paint. On an open grid the same local rule decides connectivity in $O(m)$ paints, while naming the occupancy map costs $\Omega(m^2)$ tokens, and the light cone supplies the write bound $T = \Omega(t^3/B)$. A read narrower than the next state leaves the intermediate image decorative. The information theoretic floor and the encoder class reachability point locate that event, and residual pixels keep the same task correct at a constant read. The working image at rate B_{read} is a sufficient statistic of the state. Exact rules reproduce the slopes, and a token intervention moves the frozen answer.

AI USE STATEMENT

In this work, we used generative AI tools for providing feedback on research methodology or experiments, and assisting with translation. We have not used generative AI tools for formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, implementing methods, supporting qualitative and thematic data analysis, interpreting results. Besides, generating synthetic data sets, helping develop theoretical models or conceptual frameworks, cleaning and reformatting dataset are not applicable to this work. Additionally, we used generative AI tools for modifying scientific figures or images, editing a research paper to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Eric Allender, David A. Mix Barrington, Tanmoy Chakraborty, Samir Datta, and Sambuddha Roy. Planar and grid graph reachability problems. *Theory of Computing Systems*, 45(4):675–723, 2009.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Alireza Amiri Bavandpour, Xinting Huang, Mark Rofin, and Michael Hahn. Lower bounds for chain-of-thought reasoning in hard-attention transformers. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 3274–3306, 2025.

Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. In *Advances in neural information processing systems*, volume 38, pp. 69110–69139, 2026.

Zihui Cheng, Qiguang Chen, Xiao Xu, Jiaqi Wang, Weiyun Wang, Hao Fei, Yidong Wang, Alex Jinpeng Wang, Zhi Chen, Wanxiang Che, et al. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. In *Advances in Neural Information Processing Systems*, volume 38, pp. 96084–96112, 2026.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: a theoretical perspective. In *Advances in neural information processing systems*, volume 36, pp. 70757–70798, 2023.

Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Richard Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. ReFocus: Visual editing as a chain of thought for structured image understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 17783–17805, 2025.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19520–19529. IEEE, 2025.

Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. In *International Conference on Learning Representations*, pp. 141405–141447, 2026.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 139348–139379, 2024.

Yu Huang, Zixin Wen, Aarti Singh, Yuejie Chi, and Yuxin Chen. Transformers provably learn chain-of-thought reasoning with length generalization. In *Advances in Neural Information Processing Systems*, volume 38, pp. 126815–126965, 2026.

Yekun Ke, Xiaoyu Li, Yingyu Liang, Zhenmei Shi, and Zhao Song. Circuit complexity bounds for visual autoregressive model. *arXiv preprint arXiv:2501.04299*, 2025.

Ang Li, Charles L. Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, Tom Goldstein, and Micah Goldblum. Zebra-cot: A dataset for interleaved vision-language reasoning. In *International Conference on Learning Representations*, pp. 59021–59061, 2026.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 36340–36364, 2025a.

Zejun Li, Ruipu Luo, Jiwen Zhang, Minghui Qiu, Xuan-Jing Huang, and Zhongyu Wei. Vocot: Unleashing visually grounded multi-step reasoning in large multi-modal models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3769–3798, 2025b.

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of thought empowers transformers to solve inherently serial problems. In *International Conference on Learning Representations*, pp. 11911–11943, 2024.

Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. In *International Conference on Learning Representations*, 2023.

Yiwei Lou, Yuanpeng He, Rongchao Zhang, Yongzhi Cao, Hanpin Wang, and Yu Huang. Multitasks-based deep evidential fusion network for blind image quality assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7512–7520, 2026.

Sachit Menon, Richard Zemel, and Carl Vondrick. Whiteboard-of-thought: Thinking step-by-step across modalities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20016–20031, 2024.

William Merrill and Ashish Sabharwal. A logic for expressing log-precision transformers. In *Advances in neural information processing systems*, volume 36, pp. 52453–52463, 2023a.

William Merrill and Ashish Sabharwal. The parallelism tradeoff: Limitations of log-precision transformers. *Transactions of the Association for Computational Linguistics*, 11:531–545, 2023b.

William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *International Conference on Learning Representations*, pp. 7690–7706, 2024.

Zhiwei Ning, Wenwen Tong, Xiangli Kong, Shengnan Ma, Ziyi Shang, Jingcheng Ni, Tao Hu, Yong Xien Chng, Jixuan Ying, Zehuan Wu, et al. Intersketch: An interleaved reasoning model with self-correcting visual sketch and stepwise reward. *arXiv preprint arXiv:2605.26520*, 2026.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *ICLR 2022 Deep Learning for Code Workshop (DM4C)*, 2022.

OpenAI. Introducing OpenAI o3 and o4-mini. `https://openai.com/index/introducing-o3-and-o4-mini/`, 2025.

Jorge Pérez, Javier Marinković, and Pablo Barceló. On the Turing completeness of modern neural network architectures. In *International Conference on Learning Representations*, 2019.

Omer Reingold. Undirected connectivity in log-space. *Journal of the ACM*, 55(4):1–24, 2008.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 8612–8642, 2024.

Lena Strobl, William Merrill, Gail Weiss, David Chiang, and Dana Angluin. What formal languages can transformers express? a survey. *Transactions of the Association for Computational Linguistics*, 12:543–561, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2087–2098. IEEE, 2025.

Kevin Xu and Issei Sato. A formal comparison between chain of thought and latent thought. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images. In *International Conference on Learning Representations*, volume 2026, pp. 36706–36739, 2026.

Siyu Yan, Zhuoran Yan, Haiying Xu, Panhao Zhou, Jingyu Chen, Chenhao Ji, Shuo Cao, Yongheng Zhang, Haoze Liu, Siyu Zhang, et al. See2think: Do multimodal models really use intermediate visual states? *arXiv preprint arXiv:2607.26769*, 2026.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33510–33520, 2026.

APPENDIX

A Notation 14

B Published facts and attention 14

C Protocol, embeddings, and decoder constants 15

D Simulation cost and one paint 16

E Open grid reachability 17

F Write arity and light cones 19

G Read bandwidth, decoration, and residual pairing 21

H Sketches, tools, and sandwiches 23

I Experiments 24

J Encoder class calculations 28

A NOTATION

Table 2 lists the symbols used in the theorems and in the measurements. Class names $\mathrm{TextCoT}$ and $\mathrm{VisualCoT}$ are as in Definition 4.

Table 2. Notation used in the theorems and the measurements.

symbol	meaning
m	linear grid size
n	input length, $n = m^2$ on grids unless stated
Γ	palette of the working image
$I \in \Gamma^{m \times m}$	working image
$\rho \in \{0, 1\}$	residual canvas versus native regeneration
r	decoder neighbourhood radius
k	number of tokens returned by a read
C_q	bits per token, an encoder hyperparameter
C_q^{LM}	effective capacity $I_{\mathrm{state}}/k_{\mathrm{emp}}$ of a frozen head
$B_{\mathrm{read}} = kC_q$	read bandwidth
S	task relevant intermediate state
$I_{\mathrm{state}} = H(S)$	state entropy
$R = \mathsf{Read}_k(I)$	tokens returned by a read
Txt	text transcript
A	answer
$T_{\mathrm{txt}}, t_{\mathrm{img}}$	text tokens and paint steps
B	text write arity, tokens named per step
D	diameter of the reachable component
$k^\star_{\mathrm{lb}}$	information theoretic floor $\lceil I_{\mathrm{state}}/C_q \rceil$
$k_{\mathrm{reach}} = c\varepsilon\, k^\star_{\mathrm{lb}}$	reachability point of an encoder class
$\mathsf{BLCA}[t, B]$	bandwidth limited cellular automaton of t steps

B PUBLISHED FACTS AND ATTENTION

The three facts below are used as axioms. Their proofs are in the cited papers, and they are not repeated here.

Theorem B.1 (Fact A, Merrill & Sabharwal, 2023b; Li et al., 2024)**.** *A transformer of constant depth $L = O(1)$, arithmetic precision $p = O(\log n)$, and polynomial size recognises a language in logspace uniform TC^0. At constant bit precision the same machine lies in AC^0.*

Theorem B.2 (Fact B, Li et al., 2024; Merrill & Sabharwal, 2024)**.** *Text chain of thought of T steps contains $\mathsf{SIZE}[T]$ at constant bit precision and $O(\log n)$ embeddings. Polynomial text chain of thought with generalised prenorm characterises P.*

Theorem B.3 (Fact C, Reingold, 2008; Allender et al., 2009)**.** *Undirected grid source to target reachability is in L and is NC^1 hard under AC^0 reductions.*

Transformers of infinite precision, with unbounded intermediate steps, are Turing complete (Pérez et al., 2019; Strobl et al., 2024). That regime is not used below. The survey of Strobl et al. (2024) records the precision, attention, and step hypotheses under which Facts A and B hold. The arguments in this appendix stay in the log precision, constant depth, decoder only setting of Merrill & Sabharwal (2024).

B.1 AVERAGE HARD ATTENTION AT LOG PRECISION

Constructions below use average hard attention. A query q places mass 1 on the set of keys that maximise a score, and mass 0 elsewhere. Merrill & Sabharwal (2023b; 2024) show that softmax attention with inverse temperature $\beta = \mathrm{poly}(n)$ realises average hard attention at precision $p = O(\log n)$, because a unique maximiser receives all but an inverse polynomial mass. Position embeddings are binary encodings of cell coordinates, and log precision distinguishes m^2 positions. Average hard attention is therefore treated as a realisable primitive, and the temperature argument is not repeated.

A finite lookup $\delta : \Gamma^{(2r+1)^2} \to \Gamma$ is a constant depth threshold circuit of size $O\big(|\Gamma|^{(2r+1)^2}\big)$. One hot colours make the map a finite convolution. Both realisations are logspace uniform in m.

C PROTOCOL, EMBEDDINGS, AND DECODER CONSTANTS

This section records the same machine as Section 3, with the bookkeeping written out. There are three tapes. The input tape holds the instance and is read only. The text tape is a string in V^*, written left to right, one symbol per text action. The visual tape is an image $I \in \Gamma^{m \times m}$, overwritten in place by Paint. When $\rho = 1$, the canvas keeps I between paints, so a later neighbourhood may read pixels that were never returned as tokens. When $\rho = 0$, the generator discards I after Read_k, and the next local rule is a function of (R, h_t) alone. Theorems 6 and 8 record the resource consequences of that switch on one connectivity task.

At step j the transformer $\mathcal{T}$ reads the input tape, the text tape, and the most recent k visual tokens, and emits $h_t^{(j)}$. A control head, realised as a constant depth map of $h_t^{(j)}$, selects $\alpha_j \in \{\mathsf{Text}, \mathsf{Paint}, \mathsf{Halt}\}$. A text action emits one token. A paint produces $I^{(j)}$ and returns $R^{(j)} \in \mathbb{R}^{k \times d}$. Halt is a TC^0 predicate of $h_t^{(j)}$, for example an empty frontier, a stable query cell, or an expired iteration counter. A transcript with T_{txt} text actions and t_{img} paints, each followed by a read of k tokens, on an $m \times m$ canvas of residual bit ρ and radius r, is charged

$$\big(T_{\mathrm{txt}},\, t_{\mathrm{img}},\, k,\, m,\, \rho,\, r\big).$$

The number of forward passes is $T_{\mathrm{txt}} + t_{\mathrm{img}}$. The accounting that matters is which output alphabet was used, and whether the read was wide enough to see the state.

Weights of $\mathcal{T}$, Dec, and Enc may depend on m in the nonuniform class. Every construction in this appendix is nevertheless uniform in logarithmic space. Position embeddings are binary representations of coordinates, the local rule δ does not depend on m, and the breadth first halt predicate compares an $O(\log m)$ bit counter with m^2.

Lemma C.1 (Control tokens do not store occupancy)**.** *$O(t)$ control tokens of a polynomial vocabulary carry $O(t \log n)$ bits. They cannot encode an ηm^2-bit incompressible occupancy map unless $t = \Omega(m^2/\log n)$.*

Proof. Each control token has entropy at most $\log |V| = O(\log n)$. $O(t)$ of them have entropy $O(t \log n)$. An occupancy map with minimum entropy $\eta m^2 \log |\Gamma|$ therefore cannot sit in the control stream unless the displayed lower bound holds. Definition 4 can therefore admit $O(t)$ control tokens inside $\mathrm{VisualCoT}[t, k, \rho]$ while leaving Theorem 1 intact. □

C.1 POSITION EMBEDDINGS AND NEIGHBOURHOOD SELECTION

The constructions use the average hard gadget of Merrill & Sabharwal (2024); Liu et al. (2023). The derivation below realises Paint as a transformer decoder whose local rule is explicit.

Index cells by $q = (x, y) \in \{0, \dots, m-1\}^2$, and write $x = (x_1, \dots, x_{\lceil \log m \rceil})$ for the binary expansion, and likewise y. The position embedding concatenates these bits with a constant number of one hot colour channels, using $O(\log m)$ dimensions and precision $O(\log m)$.

A query q and a key q' satisfy

$$\langle e_q, e_{q'} \rangle \;=\; 2\lceil \log m \rceil \quad \text{iff} \quad q = q',$$

and the inner product is at most $2\lceil \log m \rceil - 1$ otherwise, because a mismatched bit contributes 0 where a matched bit contributes 1. Average hard attention with this score puts mass 1 on the unique maximiser.

For each offset $\Delta \in [-r, r]^2$, one head uses the query $e_q + b_\Delta$, where b_Δ is a fixed bias that recentres the inner product onto $q + \Delta$. Positions outside the grid are mapped to a dead cell embedding, carrying a wall colour or a quiescent automaton state. The value vector is the one hot colour of the selected neighbour. Concatenating $(2r+1)^2$ such colours yields a vector in $\{0,1\}^{|\Gamma|(2r+1)^2}$.

A single hidden layer of width $|\Gamma|^{(2r+1)^2}$, with one hot indicators of neighbourhood configurations and a $|\Gamma|$ way output, realises an arbitrary δ. The width does not grow with m, because r and $|\Gamma|$ are constants of the rule family. Circuit size per pixel is therefore $O_r(1)$, and the total size is $O(m^2)$.

A head with uniform attention, followed by a threshold on the averaged value, implements disjunction, conjunction, and counting on a marked subset, and therefore the TC^0 readout that asks whether

a known cell has been visited. Softmax at inverse temperature $\beta = \mathrm{poly}(n)$ concentrates on the same argmax up to total variation $n^{-\Omega(1)}$, which is absorbed by the $O(\log n)$ precision already assumed for Fact A. The constructions do not require a Lipschitz analysis of softmax.

Assumption A0 is implemented by projecting h_t onto $O(\log n)$ bits before the broadcast. Those bits may select a rule identity or a paint colour. They may not carry an occupancy map. Lemma C.1 is the matching statement on the text side.

C.2 Constants in the local decoder

Radius $r = 1$ and $|\Gamma| = 4$ give a neighbourhood of nine cells and a lookup table of size $4^9 = 262144$. A one hot hidden layer of that width is finite and does not grow with m. The breadth first rule is sparse. Walls are absorbing, visited cells stay visited, and only an empty cell next to the frontier maps to F, so the implemented perceptron has far fewer active weights than the full table. The fits in Table 9 use a 3×3 convolution on a one hot input, which is the translation equivariant form of the same sparse rule, and the F1 score already exceeds 0.98. Copy north at radius 1 is a table of size $2^9 = 512$. The write arity lower bound does not use the size of δ. It uses the sensitivity of the query to each node of the light cone, so a smaller table does not reduce $N(t)$.

Average hard position matching uses $2\lceil \log_2 m \rceil$ bits of inner product budget. At $m = 24$ this is 10 bits, which lies inside the $O(\log n)$ precision of Fact A. Softmax at $\beta = m^2$ puts mass $1 - O(m^{-2})$ on the unique maximiser when the gap is at least 1, and the leftover mass is absorbed by the same precision. The broadcast of $O(\log n)$ bits from h_t is a rule identifier. For breadth first search it is a constant, and for a family of $O(1)$ local rules it is $O(1)$ bits. The broadcast violates Assumption A0 when it names a cell set of size $\omega(\log n)$, which is the text side content of Lemma C.1. Halt counters occupy $O(\log m)$ bits in h_t and are written as $O(1)$ control tokens per round. They compare against m^2 or against t, and they do not store occupancy. Every gate in the residual breadth first machine is therefore a local lookup of constant size, a position match of logarithmic precision, or a global OR in TC^0. That is the content of Theorem 2 on the open grid instance used in the main text.

D Simulation cost and one paint

Proof of Theorem 1. Let I take values in $\Gamma^{m \times m}$ with $H_\infty(I) \geq \eta m^2 \log |\Gamma|$. A lossless text encoder is a (possibly randomised) map $f(I) = (w_1, \ldots, w_T) \in V^T$ for which there exists g with $g(f(I)) = I$ almost surely. Losslessness implies $H(I) = H(f(I)) \leq \sum_{j=1}^{T} H(w_j) \leq T \log |V|$. Combined with $H(I) \geq H_\infty(I)$ this yields $T \geq \eta m^2 \log |\Gamma| / \log |V|$.

One Paint produces any target image in a single step by taking Dec to ignore the neighbourhood and emit the target colours. This accounting applies to the cost of storing an already computed image. It is not a claim that Paint invents connectivity from nothing.

If $|\Gamma| = 2$ and Hamming distortion $\mathbb{E}[d_H(\hat{I}, I)] \leq \varepsilon m^2$ is allowed, the rate distortion function of a Bernoulli$(1/2)$ source under Hamming distortion is $R(\varepsilon) = 1 - h_2(\varepsilon)$ bits per symbol, where $h_2(\varepsilon) = -\varepsilon \log \varepsilon - (1 - \varepsilon) \log(1 - \varepsilon)$. Hence $T \log |V| \geq (1 - h_2(\varepsilon)) m^2$. For a general palette the same argument uses $R(\varepsilon) = \log |\Gamma| - h(\varepsilon)$.

For a sequence $I_1, \ldots, I_t$ with $H(I_{j+1} \mid I_1, \ldots, I_j) \geq \beta m^2$, the chain rule gives $H(I_{1:t}) \geq \beta t m^2$. A lossless text archive therefore needs $T \log |V| \geq \beta t m^2$ tokens, while t paints produce the frames. If F_δ is a bijection on configurations, as it is for a reversible block rule, then $H(F_\delta(C)) = H(C)$, so a uniform input remains at full entropy in every frame. □

If $|V| \geq |\Gamma|$, a raster scan encodes a binary grid in $T = m^2$ tokens and matches the lower bound up to the factor $\log |\Gamma| / \log |V|$. When $|V| = 2$, the gap is the constant factor 1. Figure 4 records a raster of an incompressible binary image against one paint.

D.1 One paint as a cellular automaton step

Proof of Theorem 2. Partition the $m \times m$ image into sites of size 1×1, or into patches of size $p \times p$ with $p \leq r$. For each output site $q = (x, y)$, place $O((2r + 1)^2)$ average hard heads whose queries

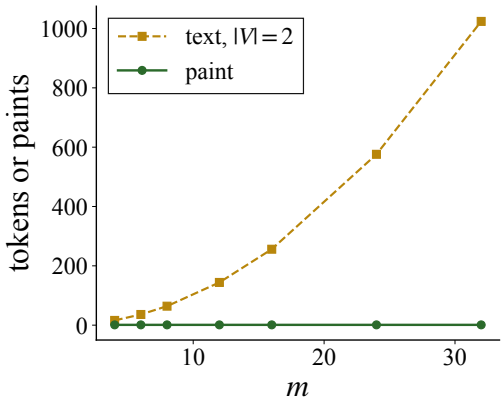


Figure 4. Binary text tokens versus one paint.

are the position embedding e_q and whose keys are the neighbour embeddings $e_{q+\Delta}$ for $\Delta \in [-r, r]^2$. On an exact position match the head assigns mass 1 to that neighbour. Value vectors carry a one hot colour in Γ. Concatenating the neighbourhood and applying a constant depth multilayer perceptron of width $O\big(|\Gamma|^{(2r+1)^2}\big)$ realises any finite table δ. Residual reading, the case $\rho = 1$, takes the neighbourhood from I^{prev}. Reading without residual pixels, with $k = m^2$ and with Enc equal to the identity, takes the neighbourhood from context patch tokens. Every pixel is written in one parallel Paint, and the output pixels inside the decoder have no serial dependence. □

A global broadcast is implemented by projecting h_t onto $O(\log n)$ bits and adding the result to the multilayer perceptron at every site. The broadcast selects a rule identity, and it does not enlarge the receptive field. Assumption A0 forbids an image summary of $\omega(\log n)$ bits inside h_t.

Justification of Remark 1. A finite δ is a finite convolution, with one hot input colours, one hot output colours, and a kernel supported on radius r. The parameter count does not grow with m, so empirical risk on a next canvas target is a standard finite dimensional problem. Spatial shuffling destroys translation equivariance, and the same convolutional class then cannot realise δ. A flattened perceptron has width growing with m^2 and does not inherit that equivariance. Appendix I reports the corresponding fits. □

E Open grid reachability

The palette is $\Gamma = \{W, E, V, F\}$, for wall, empty, visited, and frontier. The input grid is the initial canvas. Obstacles are walls, free cells are empty, and the source is painted as frontier.

Lemma E.1 (Frontier invariant)**.** *After the i-th residual paint, $V \cup F = \{u : \mathrm{dist}(s, u) \le i\}$ and $F = \{u : \mathrm{dist}(s, u) = i\}$.*

Proof. The local rule of radius 1 leaves walls as walls and turns visited or frontier cells into visited cells. An empty cell with a frontier neighbour becomes frontier, and every other colour is unchanged. The claim holds at $i = 0$ with $F = \{s\}$. One step later the new frontier is exactly the set of cells at distance $i + 1$. □

Proof of Theorem 3. On a finite grid without obstacles the Manhattan diameter satisfies $D \le 2m$. After $t_{\mathrm{img}} = D + 1$ paints the frontier is empty. The target is reachable if and only if its colour lies in $\{V, F\}$. A single Read_1, aligned to the known target coordinate, returns that colour, and the language head emits a yes or no token in one TC^0 forward pass. Parent pointers may be stored in an extra channel and followed in $O(D)$ additional paints of radius 0 when a path must be drawn. The rule δ does not depend on m. Position embeddings are binary and constructible in logarithmic space, so the construction is uniform.

For text search that writes the occupancy, a frontier list $F_i \subset [m] \times [m]$ is written explicitly. Naming one cell costs at least $2 \log m$ bits. On a yes instance at least $\mathrm{dist}(s, t) + 1$ cells enter the frontier.

On an open grid every cell is reachable, so $\sum_i |F_i| = m^2$ and $T_{\text{txt}} \geq c\, m^2 \log m / \log |V|$. Under the convention that one token names one coordinate pair, $T_{\text{txt}} \geq m^2$.

A spanning tree of parent directions is a string in $\{\uparrow, \downarrow, \leftarrow, \rightarrow, \perp\}^{m^2}$. Encoding it in a polynomial vocabulary still requires $\Omega(m^2 / \log m)$ tokens, because each token carries $O(\log m)$ bits. Search that keeps a right hand pose stores $O(\log m)$ bits. On a spiral of width one from s to t, the path has length $\Theta(m^2)$, so the text machine takes $\Theta(m^2)$ steps. Visual breadth first search on the same spiral has frontier size 1 at every layer and also takes $\Theta(m^2)$ paints. The visual saving appears when the frontier is wide. □

Table 3 records open grids and grids with mild obstacles, twelve instances per cell. On the open grids, where the obstacle rate is $p = 0$, the number of visual steps equals the diameter. At $m = 24$ the packed parent pointers use 126 tokens, against 46 paints and 576 occupancy tokens.

Table 3. Open and mildly obstructed grids.

m	p	visual steps	occupancy tokens	packed tokens	diameter
6	0.00	10.0	36	14	10
6	0.25	8.9	22	n.a.	5.8
8	0.00	14.0	64	n.a.	14
10	0.00	18.0	100	n.a.	18
12	0.00	22.0	144	n.a.	22
16	0.00	30.0	256	n.a.	30
20	0.00	38.0	400	n.a.	38
24	0.00	46.0	576	126	46
24	0.25	42.6	396	n.a.	42.3

Table 4 records the same comparison on a snake. Visual steps still equal the diameter. Write arity gives no saving, because the frontier has size one.

Table 4. Snake diameter versus open diameter.

m	snake steps	snake diameter	open steps	open diameter
4	30	30	6	6
5	48	48	8	8
6	70	70	10	10
8	126	126	14	14
10	198	198	18	18

Figure 5 plots both families together. On open grids and on snakes, the visual step count sits on the diameter.

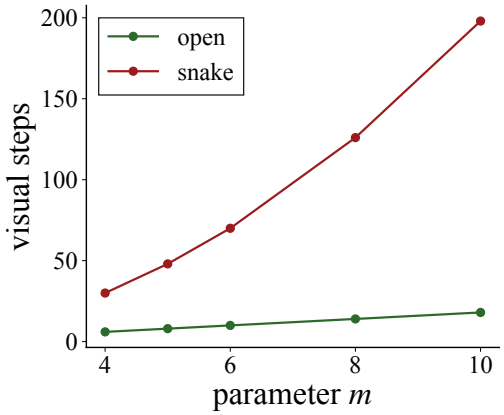


Figure 5. Visual steps track diameter on open grids and on snakes.

Remark E.1 (Conditional clause at zero steps)**.** *Under* $\mathsf{TC}^0 \neq \mathsf{NC}^1$*, Fact C implies that* $\mathrm{UGGR} \notin \mathsf{TC}^0$ *and therefore* $\mathrm{UGGR} \notin \mathrm{TextCoT}[0]$*, while Theorem 3 places open grid* UGGR *in* $\mathrm{VisualCoT}[O(m), 1, 1]$*. The clause is recorded here, and it is not an unconditional contribution.*

E.1 A WORKED BREADTH FIRST LAYER

We record one residual paint on a 3×3 open grid so that the local rule is concrete. Palette $\Gamma = \{W, E, V, F\}$. The source is the centre cell. The initial canvas is

$$\begin{bmatrix} E & E & E \\ E & F & E \\ E & E & E \end{bmatrix}.$$

The radius 1 rule of Lemma E.1 rewrites every empty four neighbour of the frontier as frontier, and rewrites the old frontier as visited. After one paint the canvas is

$$\begin{bmatrix} E & F & E \\ F & V & F \\ E & F & E \end{bmatrix}.$$

Four cells joined the frontier. A text occupancy transcript that names those cells uses four tokens when one token names one coordinate pair, and it uses $\Theta(\log m)$ bits per cell in a packed encoding. The visual machine produced all nine pixels in one forward pass and needs two further paints to cover the corners. The Manhattan diameter is 2, so $t_{\mathrm{img}} = 3$ decides every query on this grid. A parent channel may store a direction when a cell first becomes frontier. From a corner, two paints of radius 0 walk back to the source and draw a path, and those extra paints are $O(D)$.

The same layer on a snake of width one has frontier size 1, so one paint advances one cell. After D paints the visual machine and the text machine have used the same number of rounds, because there is no parallel frontier for Paint to update.

F WRITE ARITY AND LIGHT CONES

Proof of Lemma 1. A radius r local map sends information at most r cells per paint. After j paints, the value at a query cell is a function of the initial configuration inside Manhattan radius rj. If some input cell at distance rt is sensitive for P_t, as it is for a rule that copies towards the query, then $j < t$ leaves that bit unread and the output differs from P_t on that input. Hence $t_{\mathrm{img}} \geq t$. The same propagation bound holds for any local model, including a text machine that updates a sliding window. □

Lemma F.1 (Sensitive predecessors)**.** *Let the local rule be copy north, neighbourhood exclusive or, or left multiplication in the symmetric group S_5. For every light cone node $v_{q,i}$ with $i \geq 1$ there is an assignment that flips a child of $v_{q,i}$, keeps every other input fixed, and flips the query $v_{q_\star,t}$.*

Proof. Copy north is sensitive along the unique northward ray. Exclusive or is sensitive along any parity path from the node to the query. Left multiplication in S_5 is sensitive because the group is nonabelian and generated by the local factors. In each case the node is a necessary predecessor of the query. □

Proof of Theorem 4. Label dag nodes $v_{q,i}$ for the state of cell q after i updates, with edges from the radius-r neighbourhood at layer $i-1$. The number of nodes is

$$\sum_{i=0}^{t} (2ri+1)^2 \;=\; \Theta(t^3).$$

A text machine in $\mathrm{TextCoT}_B$ emits at most B tokens per step, each naming $O(1)$ pairs of a cell and a layer. By Lemma F.1 every light cone node must be written, or computed from already written children, inside a TC^0 forward pass. Naming $\Theta(t^3)$ nodes at rate B per step requires $T = \Omega(t^3/B)$.

If the machine recomputes every query from the input, the recursion tree has branching factor $\Theta(r^2)$ and depth t, hence size $\Theta((2r+1)^{2t})$, which is worse. Writing the intermediate cells is the cheaper bookkeeping inside this subclass, so the cubic bound applies to the whole subclass.

A weaker information bound uses only entropy. For a mixing or reversible rule the terminal light cone has minimum entropy $\Omega(t^2)$. The tape accumulates at most $TB \log |V|$ bits, so $T = \Omega(t^2/(B \log |V|))$ is required to store a sufficient statistic.

The case $B = O(1)$ recovers $T = \Omega(t^3)$. The case $B = \Theta(t^2)$ recovers $T = \Omega(t)$, matching t residual paints, while the total number of named cells remains $\Theta(t^3)$. The case $B = \omega(t^2)$ leaves the regime of stepwise local updates, and the theorem then claims nothing stronger than $\Omega(t^3/B)$. Linear rules over $\mathrm{GF}(2)$ may have binomial closed forms and are excluded. Residual visual chain of thought updates $\Theta(t^2)$ cells in parallel at each of t paints, and that count does not depend on the text write arity B. □

On a copy north cone of depth $t = 20$ the node count is 12,341, and $12{,}341/t^3 \approx 1.54$, in agreement with $\sum_{i=0}^{t}(2i+1)^2$. A machine with $B = 1$ uses 12,341 steps. A machine that names one layer per step uses 8 steps. Residual paints use 20 steps. Table 5 expands those counts.

Table 5. Copy north light cone. Visual steps equal t.

t	$N(t)$	visual	$B = 1$	$B = 4$	$B = 16$	$B =$ layer
4	165	4	165	42	11	3
8	969	8	969	243	61	5
12	2925	12	2925	732	183	6
16	6525	16	6525	1632	408	7
20	12341	20	12341	3086	772	8

Figure 6 plots the ratio of text steps at $B = 1$ to residual paints. The ratio grows like t^2. A naming budget of one layer per step stays near 1.

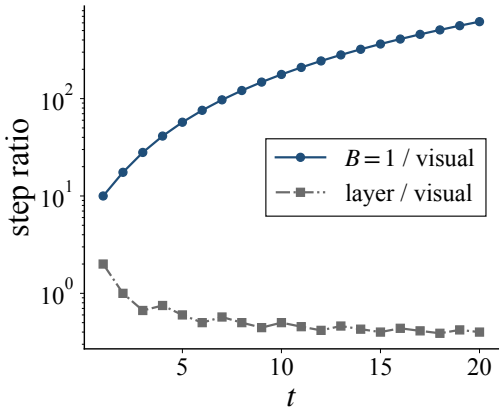


Figure 6. Ratio of text steps at $B = 1$ to residual paints.

F.1 Conditional depth hierarchy

Proposition F.1 (Conditional depth hierarchy)**.** *Let each cell store an element of the symmetric group S_5 and let δ replace the site by a noncommutative product of itself and its neighbours. Then, unless $\mathsf{TC}^0 = \mathsf{NC}^1$,*

$$P_t \in \mathrm{VisualCoT}[t, k_\star, \rho] \setminus \mathrm{VisualCoT}[t-1, k_\star, \rho]$$

for all sufficiently large $k_\star$.

Proof. Iterated noncommutative products of elements of S_5 cannot be folded into AC^0 or TC^0 unless $\mathsf{TC}^0 = \mathsf{NC}^1$. Li et al. (2024) used S_5 composition as a serial empirical task. Here the product is applied in parallel at every cell, while the serial depth at each cell remains t. A single TC^0 forward pass cannot collapse those t iterations unless the collapse holds. Visual chain of thought with $t-1$ paints realises a local circuit of depth $t-1$ and is insufficient. The bound of t paints is sufficient by Theorem 2. □

The clause has the same conjectural strength as the text step hierarchy of Merrill & Sabharwal (2024). It is recorded here, and it is not listed among the unconditional contributions.

G READ BANDWIDTH, DECORATION, AND RESIDUAL PAIRING

Proof of Theorem 5. $R = \mathrm{Enc}(I)$ is a function of I, so $S - I - R$ is a Markov chain whenever S enters I through Paint. Data processing yields $I(R;S) \le I(I;S) \le H(I)$ and $I(R;S) \le I(R;I) \le H(R)$. After quantisation to p bits, R has support size at most 2^{kdp}, so $H(R) \le kdp$. Projection onto a subspace of dimension u replaces d by u. A Gaussian bound of the same form, taken per query, gives $I(R;S) \le kC_q = B_{\mathrm{read}}$. The minimum of these upper bounds is the statement. □

Lemma G.1 (Fano's inequality)**.** *Let S be discrete and $\hat{S} = f(R, h_t)$. Then*

$$P_e \;\ge\; 1 - \frac{I(S;R,h_t)+1}{\log|\mathcal{S}|}.$$

If $I(S;R,h_t) \le B_{\mathrm{read}} + O(dp) < H(S) - 2$ and S is nearly uniform, then $P_e \ge 1/3$.

Proof. The standard Fano inequality is $H(S \mid \hat{S}) \le h_2(P_e) + P_e \log(|\mathcal{S}| - 1)$ and $H(S \mid \hat{S}) \ge H(S) - I(S;R,h_t)$. Rearrangement gives the display. The numerical form uses $H(S) \approx \log|\mathcal{S}|$ and $B_{\mathrm{read}} + O(dp) < H(S) - 2$. □

Proof of Theorem 6. When $\rho = 0$, the next Dec sees only (R, h_t). Assumption A0 confines every bit from outside the neighbourhood to those registers. For a visited map $S \in \{0,1\}^{m^2}$ that is nearly uniform on a reachable component, $H(S) = \Theta(m^2)$. Choosing k so that $kC_q + O(dp) = o(m^2)$ and applying Lemma G.1 yields $P_e \ge 1/3$. □

Proof of Theorem 7. The answer head sees (R, Txt). The chain rule and the Markov relation $A \leftarrow (S, \mathrm{Txt}) \leftarrow (R, \mathrm{Txt})$, when A is a function of S and the text, give $I(A; R \mid \mathrm{Txt}) \le I(S; R \mid \mathrm{Txt}) \le I(S;R) \le B_{\mathrm{read}}$. If $A \in \{0,1\}$ and $H(A \mid \mathrm{Txt}) = h > B_{\mathrm{read}} + \Delta$, then $H(A \mid R, \mathrm{Txt}) \ge h - B_{\mathrm{read}} > \Delta$. Binary Fano converts the residual entropy into an $\Omega(\Delta)$ error for any correction of the text posterior that uses only R.

If the model already solves A from Txt, the image is decorative. If B_{read} is too small to carry the remaining entropy, the image cannot help. A corrupted render I' produces tokens R' that still obey $I(A; R' \mid \mathrm{Txt}) \le B_{\mathrm{read}}$. Changing the answer therefore requires that $I(A; R \mid \mathrm{Txt})$ was already large, which is the event that the head uses R. □

The identity $k^\star_{\mathrm{lb}} = \lceil (I_{\mathrm{state}} - \varepsilon)/C_q \rceil$ is necessary for every encoder, because $I(R;S) \le kC_q$. Sufficiency holds when Enc is an optimal code of S on k tokens, as in Theorem 9, and $kC_q \ge H(S)$. Equispaced retention, frozen heads, and linear probes are weaker codes, so a measured crossing may sit at $c_{\mathcal{E}}\, k^\star_{\mathrm{lb}}$ with $c_{\mathcal{E}} > 1$.

Proof of Proposition 1. The floor $k^\star_{\mathrm{lb}}$ is necessary for an arbitrary Enc by data processing and Fano's inequality. If every Enc_k in $\mathcal{E}$ fails to be an optimal code of S, there exists $c_{\mathcal{E}} \ge 1$ such that $I(\mathrm{Enc}_k(I); S) < I_{\mathrm{state}} - 2$ for all $k < c_{\mathcal{E}}\, k^\star_{\mathrm{lb}}$, and Lemma G.1 gives $P_e \ge 1/3$. The multiplier is a property of the pair $(\mathcal{E}, \mathcal{G})$. Orthonormal projection with least squares recovery, on a 256 bit occupancy map and with $C^{\mathrm{tb}}_q = 8$, has $k^\star_{\mathrm{lb}} = 32$ and a Hamming drop only near 96 to 128, hence $c_{\mathcal{E}} \in [3, 4]$. Frozen Qwen3-VL heads on a half plane query have $k^\star_{\mathrm{lb}} = 1$ and $k_{\mathrm{emp}} \in \{32, 8, 4\}$, hence $c_{\mathcal{E}} \in \{32, 8, 4\}$. If $\mathcal{G}$ cannot represent the task, P_e tracks the text baseline for every finite k and $c_{\mathcal{E}} = \infty$. □

Proof of Proposition 2. Let $N = N_v$ and retain equispaced indices $i_j = \mathrm{round}(j(N-1)/(k-1))$, with the midpoint when $k = 1$.

For support hitting, let $U \subset [N]$ be a sufficient index set with $|U| = \alpha N$. Without interpolation, R carries information about S only when the retained set meets U. The expected number of hits is about αk, so a constant expected hit requires $k = \Omega(1/\alpha)$. Dividing by $k^\star_{\mathrm{lb}} = \lceil I_{\mathrm{state}}/C_q \rceil$ produces a lower bound on $c_{\mathcal{E}}$. A half plane may take $\alpha = 1/2$, and a single cell has $\alpha = 1/N$.

For linear means, write the retained tokens as $z_j = s \cdot \mathbf{1}_{j \in U} + \xi_j$ with $\mathrm{Var}(\xi_j) = \sigma^2$. The mean $\bar{z}$ has signal αs and variance σ^2/k, hence signal to noise ratio $\mathrm{SNR} = \alpha |s| \sqrt{k}/\sigma$. Crossing a threshold τ

requires $k = \Omega(1/\alpha^2)$. For a half plane this is still $O(1)$. A measured $k_{\text{emp}} = 32$ on a 2B half plane query is therefore a property of the head.

For independent patches, if $S = (S_1, \ldots, S_N)$ is nearly independent, then after the unselected tokens are set to zero one has $I(R; S) \le (k/N)H(S)$. Recovering I_{state} needs $k \ge N \cdot I_{\text{state}}/H(S)$. A visited map with $H(S) = \Theta(N)$ and a one bit answer still needs $k = \Omega(N)$ when $\rho = 0$, because the next local rule must see the map. □

Proof of Theorem 8. In the residual case, Theorem 3 places the problem in $\text{VisualCoT}[O(D), O(1), 1]$. After the canvas has finished the search, the language side state has entropy one bit.

Without residual pixels, the j-th paint must reconstruct the previous frontier from $R^{(j-1)}$. On an open grid, $H(S^{(j-1)}) = \Theta(m^2)$ from some layer onward. Theorem 6 then forbids $k = o(m^2/C_q)$. An error on the frontier is amplified along the light cone, and later local updates do not repair it.

The two sides use the same δ and the same $t = O(D)$. The gap is only the residual bit. □

Two causal paths are distinct when $\rho = 1$. A pixel intervention changes the canvas and therefore every later neighbourhood, even at $k = 0$. A token intervention changes only R, and it moves the answer only when the head uses R. A corrupted render in the sense of See2Think typically hits both paths at once. The measurements in Section 5.2 separate them.

Table 6 is an orthonormal readout of a 16×16 visited map. The textbook capacity $C_q^{\text{tb}} = 8$ gives the floor $k_{\text{lb}}^{\star} = 32$. Hamming distance stays large past that floor and vanishes only at $k = m^2$, which is the reachability point of this encoder class.

Table 6. Orthonormal readout of a 16×16 visited map.

k	B_{read} (textbook bits)	Hamming	reconstruct acc.
4	32	0.51	0.50
8	64	0.45	0.65
16	128	0.51	0.50
32	256	0.44	0.65
64	512	0.45	0.55
96	768	0.35	0.65
128	1024	0.27	0.80
192	1536	0.11	0.90
224	1792	0.041	0.95
256	2048	0.00	1.00

Figure 7 plots the same Hamming curve. The textbook floor is the first k at which kC_q^{tb} meets 256 bits. Least squares on orthonormal rows is a weaker code of the visited map, and the curve falls only well above that floor.

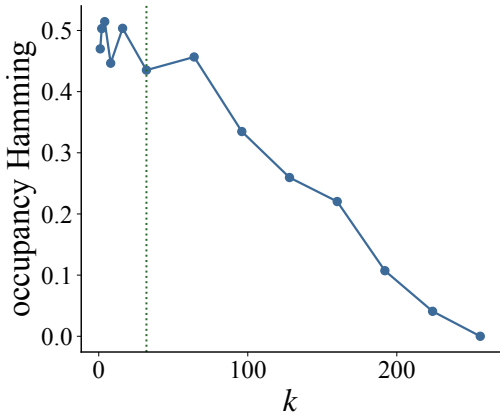


Figure 7. Occupancy Hamming against k on an orthonormal encoder.

Table 7 is the residual pairing of Theorem 8 on the same breadth first rule. Residual readout of a known target is already correct at $k = 1$. Without residual pixels, visited Hamming stays above 0.10 until $k = m^2$.

Table 7. Residual pairing on the same breadth first rule.

m	residual k for acc. 1	nonresidual k for Hamming ≤ 0.10	k/m^2	Hamming at $k = m^2/2$
8	1	64	1.0	0.14
12	1	144	1.0	0.14
16	1	256	1.0	0.14

H SKETCHES, TOOLS, AND SANDWICHES

Proof of Theorem 9. Let $I = (I_{\mathrm{sk}}(S), \xi)$ with ξ independent of S. If I_{sk} is a deterministic sufficient statistic of S, then $H(I \mid S) = H(\xi) = \Psi$ and $I(I; S) = H(I) - \Psi$. Data processing gives $I(R; S) \leq I(I; S) = H(I) - \Psi$. A sketch has $\Psi = 0$ and $I(I; S) = H(S)$. Taking the minimum with B_{read} recovers Theorem 5.

Maximising $I(R; S)$ subject to $I(R; I) \leq B_{\mathrm{read}}$ on the Markov chain $S - I - R$ is an information bottleneck. The optimal R is a compression of S, and the optimal I is any rendering from which Enc can read that compression. Occupancy marks and frontiers are discrete bottleneck codes.

A Gaussian heuristic makes the oversampling explicit. Suppose the first q principal components each contribute c nats to S and the remainder is isotropic texture of entropy Ψ. A connector that keeps k components then needs $k_{\mathrm{sk}} = \lceil H_0/c \rceil$ on a sketch and $k_{\mathrm{ph}} = k_{\mathrm{sk}}(1 + \Psi/H_0)$ on a photograph. Isotropy is a heuristic. The distribution free statement is the bound in the theorem, namely $I(R; S) \leq H(I) - \Psi$ whenever the encoder caps $H(I)$. □

Table 8 records occupancy Hamming at $k = 128$. A sketch is a sufficient statistic of the state. A photograph adds independent texture of entropy Ψ, and the Hamming distance rises with that texture.

Table 8. Sketch versus photograph occupancy Hamming at $k = 128$.

rendering	k	mean Hamming
sketch, $\Psi = 0$	128	0.23
photograph, $\Psi = 0.4$	128	0.24
photograph, $\Psi = 0.8$	128	0.32

H.1 TOOLS, POINTERS, AND INHERENT SERIALITY

Proof of Theorem 10. A tool returns $b = O(\mathrm{polylog}\, n)$ bits per call. After t_{tool} calls the total information is $t_{\mathrm{tool}} b$. Writing an incompressible $m \times m$ configuration therefore needs $t_{\mathrm{tool}} = \Omega(m^2/b)$ calls, while one Paint produces $m^2 \log |\Gamma|$ bits. If the tool internally implements grid routing, connectivity is solved in one call and the expressivity lives in the tool. Assumption A2 forbids that implementation inside the canvas.

Iterated modular squaring $x_{i+1} = x_i^2 \bmod M$ for t rounds, with n bit inputs x_0 and M, is a standard serial problem (Li et al., 2024). A local rule on the plane propagates information at r cells per paint. Placing the integer along a row, one multiplication has a carry of depth $\Omega(n)$ and needs $\Omega(n/r)$ paints, so t squarings need $\Omega(tn/r)$ paints. A tool that implements one squaring per call uses t calls. Text chain of thought that emits the n bits uses $\Omega(tn/\log |V|)$ tokens.

A paint of radius r advances a mark by one cell, by Lemma E.1. Reaching distance D therefore needs $t_{\mathrm{img}} \geq \lceil D/r \rceil$, independently of B_{read}, by Lemma 1. Extra read tokens do not shorten the diameter. □

H.2 A BANDWIDTH LIMITED SANDWICH

Definition H.1 (Bandwidth limited cellular automata)**.** $\mathsf{BLCA}[t, B_{\mathrm{read}}]$ *consists of* t *layers. Each layer is a local threshold map of radius* r*, followed by a* TC^0 *map that compresses* m^2 *pixels to at most* B_{read} *bits.*

Proposition H.1 (Circuit simulation)**.** *Under Assumptions A0 and A1, every* $\mathrm{VisualCoT}[t, k, \rho]$ *computation is simulated by a circuit in* $\mathsf{BLCA}[t, kC_q]$*. If* $B_{\mathrm{read}} = O(\log n)$ *and* $\rho = 0$*, then* $\mathsf{BLCA}[t, B_{\mathrm{read}}] \subseteq \mathrm{TextCoT}[O(t)]$*. If* $B_{\mathrm{read}} \geq m^2 \log |\Gamma|$ *or* $\rho = 1$*, the sandwich contains* t *steps of a radius* r *cellular automaton on the plane, and it is contained in a local circuit of depth* $O(t)$ *and size* $O(tm^2)$*.*

Proof. Each Paint is a local map in NC^0 or AC^0, of constant depth and bounded fan in, with size $O\big(m^2 \cdot \mathrm{poly}(|\Gamma|, 2^{O(r^2)})\big)$ by Assumption A0. Each Read_k, followed by one transformer forward pass at logarithmic precision, lies in TC^0 by Fact A, and emits at most $B_{\mathrm{read}} + O(dp)$ bits. Stacking t such layers is the sandwich.

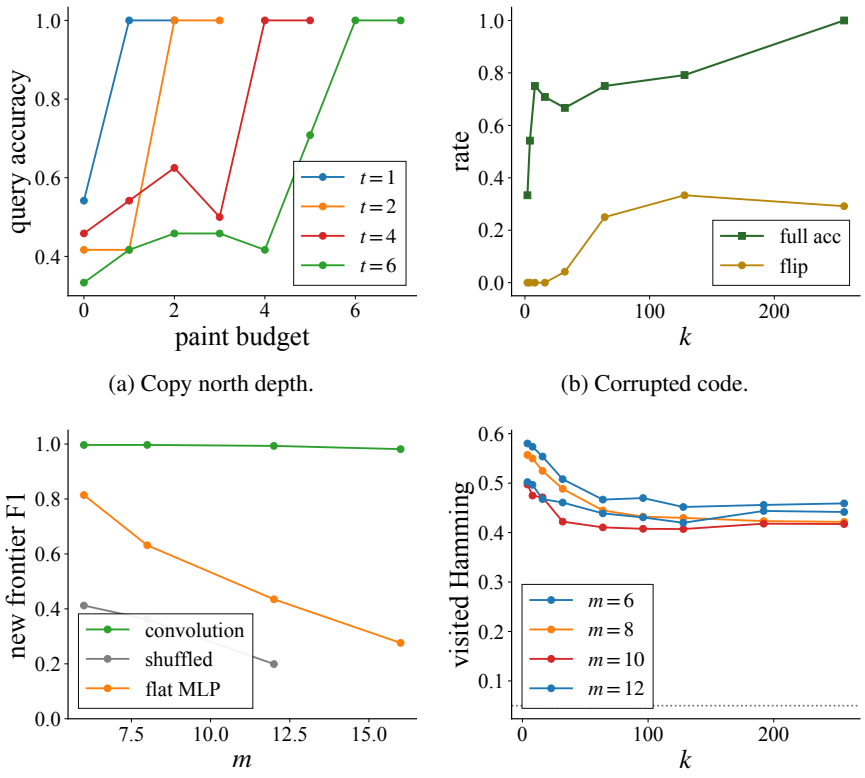
If the pipe is $O(\log n)$ bits and pixels are not kept, the pipe contents can be written as $O(1)$ text tokens and the computation is text chain of thought. If the pipe is the identity, each layer is one automaton step. The natural circuit has depth $O(t)$ and size $O(tm^2)$. For $t = O(\log m)$ the depth is $O(\log m)$, which sits near NC^1 after the threshold gates are expanded to bounded fan in. The class is not named more tightly than local circuits of depth $O(t)$. □

Xu & Sato (2026) compare a looped latent thought, which reaches TC^k in $\mathrm{polylog}$ rounds, with text chain of thought, which drops one level because tokens are emitted serially. A residual visual tape is a spatially parallel loop, and one round updates m^2 cells where a latent loop updates one hidden vector. The construction does not automatically yield TC^k, because a local automaton is not a global threshold gate. Theorem 10 shows that none of the three geometries dominates the other two on every task.

I EXPERIMENTS

Synthetic machines implement the local rules of Theorems 2, 3, and 4 exactly. Text baselines name cells, or pairs of a cell and a layer, and they do not call an external solver. Frozen measurements use the public Qwen3-VL Instruct weights at three scales (Bai et al., 2025), a side length of 448 pixels, a single tile, and $N_v = 196$ visual tokens. Read_k retains an equispaced subset and sets the complementary tokens to zero. Each reported cell uses fifty instances, unless a caption states forty. Answers are binary decisions at the last position, and the hardware is a single A100 with forty gigabytes of memory. The quantities below are slopes, crossings, and interventions.

Figure 8 records four supplementary slopes. Copy north accuracy is a cliff at the light cone budget t, and it is chance below that budget. A synthetic corrupted render, read by an orthonormal encoder, leaves a one bit answer unmoved at small k and moves it once k passes the reachability point of that encoder class. Linear probes on frozen vision transformer tokens never push visited Hamming below 0.05, so the linear capacity of that encoder class lies well below the textbook value of eight bits.



(a) Copy north depth.

(b) Corrupted code.

(c) Frontier fit.

(d) Linear probe.

Figure 8. Supplementary synthetic measurements.

A 3×3 convolution trained on the new frontier of breadth first search attains the F1 scores in Table 9. Spatial shuffling of the target, and a flattened multilayer perceptron, both degrade as m grows. The convolutional class is the construction of Remark 1.

Table 9. New frontier F1 score for a local convolution.

m	convolution	shuffled	perceptron
6	0.997	0.412	0.814
8	0.997	0.362	0.632
12	0.993	0.199	0.435
16	0.982	0.200	0.276

The same slopes reappear on the instances already used in the proofs. A raster of an incompressible binary image at $m = 32$ uses 1024 text tokens and one paint, as plotted in Figure 4. An open grid at $m = 24$ uses 46 paints, 576 occupancy tokens, and 126 packed tokens, as listed in Table 3. Light cone memoisation at $t = 20$ names 12,341 cells against 20 paints, as listed in Table 5. Iterated squaring for $t = 16$ is 16 tool calls. The visual lower bound from the carry is 128 local steps at $n = 8$ bits and 1024 steps at $n = 64$ bits. Copy north query accuracy is a cliff at budget t for every $t \in \{1, 2, 4, 6, 8\}$ on a 17×17 grid. At $t = 8$, accuracy is 0.33 at budget 4 and 1.00 at budget 8. Doubling the radius at a fixed budget of 4 raises accuracy from 0.50 to 1.00, which is Lemma 1 on a short instance.

When $|V| = 2$ and $\Gamma = \{0, 1\}$, a raster encodes m^2 tokens and Theorem 1 requires $T \geq m^2$ in the uniform case $\eta = 1$, so the gap is a factor 1. When $|V| \geq m^2$, a packed raster encodes a binary grid in $O(m^2 / \log m)$ tokens, and the gap is the same logarithmic factor as the parent pointer baseline of Theorem 3. One paint remains one step for every vocabulary. For a Bernoulli half source and Hamming distortion $\varepsilon = 0.11$, one has $h_2(\varepsilon) \approx 0.50$, so $T \log |V| \geq 0.50\, m^2$. At $m = 16$ and $|V| = 2$ this is 128 tokens for an 11% reconstruction, against one paint. The main text reports the lossless clause, because the occupancy maps used as working images are already discrete and cheap to encode exactly.

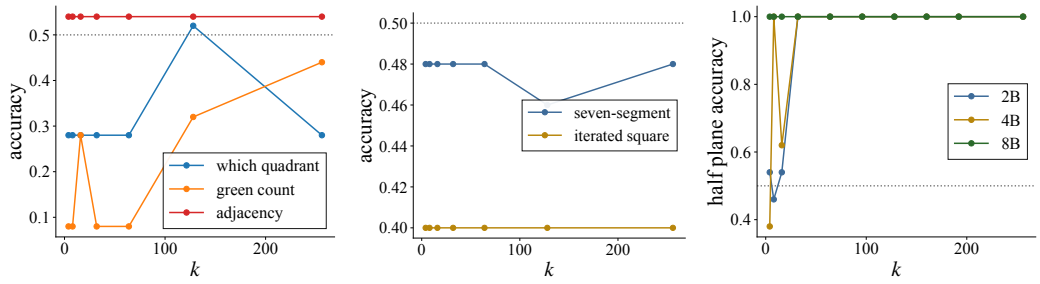
The light cone count has the closed form

$$\sum_{i=0}^{t}(2i+1)^2 = \sum_{i=0}^{t}(4i^2+4i+1)$$
$$= 4\cdot\frac{t(t+1)(2t+1)}{6} + 4\cdot\frac{t(t+1)}{2} + (t+1)$$
$$= \frac{(2t+1)(2t+2)(2t+3)}{3}.$$

The leading term is $\frac{8}{3}t^3$, so $N(t)/t^3$ tends to $8/3$. The measured ratio 1.54 at $t = 20$ is the still transient value of that rational.

I.1 Frozen heads and representation ceilings

Figure 9 reports three further scans of frozen heads. Two bit variants of the colour geometry, namely the quadrant, the number of green cells among four, and adjacency, never cross accuracy 0.8 on Qwen3-VL-2B. The same geometry with a one bit colour query does cross, as shown in Figure 3. The gap is a limit of what the frozen head represents. Seven segment parity and iterated visual squaring remain near chance at every k, in agreement with Theorem 10. Half plane accuracy moves left as the head grows from 2B to 8B, which is the decline of $c_{\mathcal{E}}$ in Table 1.



(a) Two bit colour queries. (b) Serial visual tasks. (c) Half plane across scale.

Figure 9. Additional scans of frozen heads.

A scripted residual flood fill produces a correct visited canvas, while the 2B head answers yes almost independently of k. The closed loop is therefore the local decoder of Remark 1 under the residual regime of Theorem 8. Green fraction estimation on the 8B head reaches at most 0.54, and the main text leaves it out of the decoration claim.

Table 10 records the effective capacity $C_q^{\mathrm{LM}} = I_{\mathrm{state}}/k_{\mathrm{emp}}$ of the language head, on queries the head solves, together with the equispaced first hit on a 14×14 token grid. A random cell has first hit $k = 128$ and is not a solved query, so C_q^{LM} is left undefined there.

Table 10. Effective capacity of the language head on solved queries.

head and query	k_{emp}	C_q^{LM}	first hit
2B half plane	32	0.031	1
4B half plane	8	0.125	1
8B half plane	4	0.250	1
2B colour quadrant	32	0.031	1
random cell, $p \geq 1/2$	128	n.a.	128

Table 11 expands the half plane windows at three scales. The window runs from the last k at which the answer is still chance to the first k at which full accuracy and the corrupted render flip both reach one. The decline of $c_{\mathcal{E}}$ with scale is the head factor of Proposition 2. Equispaced retention already hits a sufficient token at $k = 1$ on this query, so the table is a measurement of the head.

Table 12 lists queries that stay below accuracy 0.8 for the 2B head, on every tested k. These rows are a ceiling on the representation. The one bit colour quadrant on the same geometry does cross in Table 13, so the ceiling is a property of the head.

Table 11. Half plane crossings at three scales.

head	k_{emp}	window	C_q^{LM}	$c_{\mathcal{E}}$ if $k_{\mathrm{lb}}^{\star} = 1$
2B	32	$[16, 32]$	0.031	32
4B	8	$[4, 8]$	0.125	8
8B	4	$[0, 4]$	0.250	4

Table 12. Queries the 2B head does not represent.

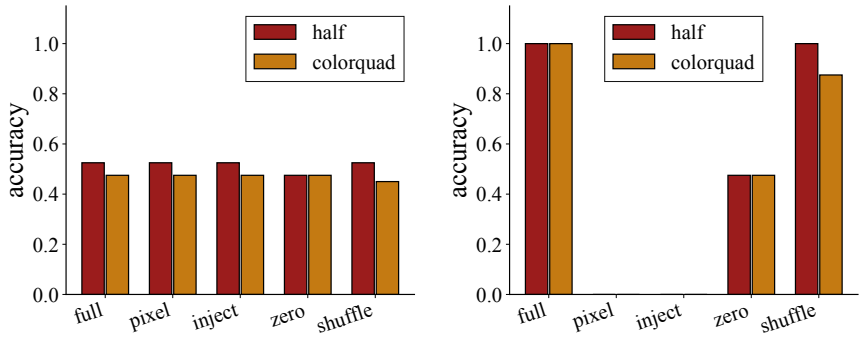
query	I_{state} (bits)	best full accuracy	flip at that k
which quadrant	2	0.52 at $k = 128$	0.70
how many green among four	2.3	0.44 at $k = 256$	0.36
adjacency	1	0.54 at every k	0
seven segment parity	1	0.48 at every k	0
iterated visual square	serial	0.40 at every k	0

The four See2Think settings, together with a replacement that touches only tokens, are the causal diagnostics of Theorem 7. Table 13 reports Qwen3-VL-2B on forty instances. At $k = 8$ every setting is chance. At $k = 32$ the full render is perfect, a render corrupted pixel by pixel is 0, and replacing only the pooled tokens by the corrupted stream is 0. Setting the tokens to zero returns chance. Shuffling the tokens leaves a one bit colour bag intact.

Table 13. Token path intervention on Qwen3-VL-2B.

query	k	full	pixel	inject	zero	shuffle
half plane	8	0.50	0.50	0.50	0.47	0.50
half plane	32	1.00	0.00	0.00	0.48	1.00
half plane	256	1.00	0.00	0.00	0.48	1.00
colour quadrant	8	0.48	0.48	0.48	0.48	0.50
colour quadrant	32	1.00	0.00	0.00	0.48	0.88
colour quadrant	256	1.00	0.00	0.00	0.48	0.88

Figure 10 shows the same intervention at two token budgets. At $k = 8$ every bar is chance. At $k = 32$ the full render is perfect, and both the pixel replacement and the token replacement collapse.



(a) Token path at $k = 8$. (b) Token path at $k = 32$.

Figure 10. The same intervention at two token budgets.

On the orthonormal encoder, a synthetic corrupted render at $m = 16$ uses twenty four instances. The flip is 0 at $k = 2$ and 0.29 at $k = 256$, with full accuracy 0.33 and 1.00 respectively. A flip of 0 at small k is the decoration regime, because the policy does not use R, and corrupting R or I cannot move A. Full accuracy at small k on easy open instances is a text heuristic. Fifty instances of a binary decision have a standard error near 0.07 at chance and near 0 at accuracy 1. The crossings reported above are windows, and they are identical across $m \in \{6, 8, 10\}$ on the 2B half plane. Forty instances in Table 13 separate 1.00 from 0.00 at $k = 32$. They do not resolve an effect of size 0.02.

J ENCODER CLASS CALCULATIONS

The arithmetic below is the content of Propositions 1 and 2 on the instances used above. A 16×16 binary map has $I_{\text{state}} = 256$. Uniform tokens of eight bits give $C_q^{\text{tb}} = 8$ and $k_{\text{lb}}^{\star} = \lceil 256/8 \rceil = 32$, while a one bit colour query has $k_{\text{lb}}^{\star} = 1$ for any $C_q \geq 1$. On the frozen models, $N_v = 196$ tokens lie on a 14×14 grid, and equispaced retention of k indices is $i_j = \text{round}(j(N_v - 1)/(k - 1))$. A half plane occupies a fraction $\alpha = 1/2$ of the tokens, so a hit is expected at $k = O(1)$ and is observed at $k = 1$. A colour quadrant is the same order. A uniformly random single cell has $\alpha = 1/196$, and the first k whose hit probability is at least $1/2$ is $k = 128$.

Retained tokens of the form $z_j = s \cdot \mathbf{1}_{j \in U} + \xi_j$, with variance σ^2, have mean signal αs and variance σ^2/k, hence signal to noise ratio $\text{SNR} = \alpha |s| \sqrt{k}/\sigma$. Crossing a fixed threshold on a half plane is $k = O(1)$. The measured $k_{\text{emp}} = 32$ on the 2B half plane query is therefore a property of the head, and it shrinks to 8 and then to 4 at 4B and 8B. If the coordinates $S = (S_1, \ldots, S_{N_v})$ are nearly independent and the unselected tokens are set to zero, then $I(R; S) \leq (k/N_v) H(S)$. Recovering a visited map of $\Theta(N_v)$ bits at $\rho = 0$ still needs $k = \Omega(N_v)$, even when the answer is one bit, because the next local rule must see the map. Linear probes on frozen vision transformer tokens never reach Hamming 0.05, which is the same statement for a different decoder class $\mathcal{G}$.

Effective capacity of the language head is defined only on queries the head solves, by $C_q^{\text{LM}} = I_{\text{state}}/k_{\text{emp}}$. On the half plane this is $1/32$, $1/8$, and $1/4$ at the three scales. The quantity is the number of bits per token that this head used on the queries it solves.

A 3×3 rule on a four colour one hot input has $4 \cdot 9 \cdot 4 = 144$ kernel weights when the output is one hot over the same palette, and the count does not grow with m. Supervised training on the next frontier attains the F1 scores in Table 9. Spatial shuffling of the target drops the score from 0.41 to 0.20, and a flattened perceptron drops it from 0.81 to 0.28. The construction of Theorem 2 therefore lies in a standard function class whose size does not depend on m. That class is the local decoder used by the theorem.

Fano's inequality on a visited map with $H(S) = m^2$ and bandwidth $B = kC_q$ gives $P_e \geq 1 - (kC_q + 1)/m^2$. For $m = 16$ and $C_q = 8$, the bound exceeds $1/3$ whenever $k < 21$. The orthonormal encoder in Table 6 remains near Hamming 0.45 through $k = 64$, which is a weaker guarantee than Fano. Fano's inequality lower bounds the error of an optimal decoder of S. Hamming distance under least squares is a different quantity, and the two statements are compatible. The inequality forbids success below the floor, while a particular encoder may still fail well above the floor.

A parent pointer field in $\{\uparrow, \downarrow, \leftarrow, \rightarrow, \perp\}^{m^2}$ has entropy at most $m^2 \log 5$ bits. A polynomial vocabulary carries $O(\log m)$ bits per token, so at least $\Omega(m^2/\log m)$ tokens are required. At $m = 24$ this agrees with the measured 126 packed tokens against 576 occupancy tokens in Table 3. An n bit schoolbook square has a carry that travels $\Omega(n)$ positions. A paint of radius r advances the carry by r cells, so one square is $\Omega(n/r)$ paints and t squares are $\Omega(tn/r)$ paints. At $n = 8$ and $r = 1$ this is 128 local steps for $t = 16$, and at $n = 64$ it is 1024. A tool that implements one square per call uses t calls.

After j residual paints the visited set is a deterministic function of the input and of j, so the language head needs only the colour of a known target and $k = O(1)$ suffices. After j paints without residual pixels, the next decoder must recover that visited set from $R^{(j)}$. On an open grid, from some layer onward, a constant fraction of cells remains undetermined given the text, so $H(S^{(j)}) = \Theta(m^2)$. Theorem 6 then applies at every later layer. An error on the frontier is copied outward by δ, and later local updates do not repair it. A half plane answer, and a colour quadrant answer, depend on the bag of patch colours once a sufficient patch has been retained. They do not depend on the spatial order of those patches. Permuting the k tokens leaves the bag unchanged, setting the tokens to zero removes the bag, and injecting a corrupted bag replaces the colour. These three interventions therefore probe the linear mean in Proposition 2 on a colour bag.