TEMPORAL-ATTENTION HEAD SPECIALIZATION DURING VIDEO DIFFUSION TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video diffusion transformers depend on temporal attention to coordinate information across frames, yet existing analyses of this mechanism examine only trained models. Such final-checkpoint analyses show what temporal structure exists but not when or where it forms *during* training, which is the information needed to monitor training and to target head-level interventions such as pruning. Population averages can also hide this structure, because the increase in a few heads and the decrease in most other heads offset each other in the mean. We therefore conduct a checkpoint-resolved census of every temporal-attention head across nine Open-Sora STDiT training runs spanning three model scales (306M to 1.03B parameters). Each head is scored by its cross-frame attention concentration (CFAC), the entropy-normalized concentration of its attention over the 16 frames. CFAC is 0 when a head spreads attention evenly across all frames and approaches 1 when it routes each frame's attention to a few specific frames. Heads are then selected under a preregistered change-point and effect-size rule. Per-head analysis shows that a small minority of heads, roughly 4–13% in full-grid runs, becomes strongly concentrated during training, while aggregate CFAC stays flat or decreases in every run. Across seeds, selected heads repeatedly appear in the first temporal block, but the specific heads selected within that block differ across seeds. In the 760M runs, selected heads show two main recurring frame-to-frame attention patterns. In the first, a self-frame diagonal, each frame attends mostly to itself. In the second, an adjacent-frame band, each frame attends mostly to its neighbors. These patterns recur across runs even when they emerge in different heads (i.e., at different layer and head indices). Our correlation and ablation experiments do not establish that these heads affect generated video quality. The census nevertheless identifies when and where temporal specialization forms during training. The same checkpoint-resolved, per-head analysis under fixed selection rules can provide this information for other factorized video diffusion transformers and, with adapted routing metrics, for joint spatio-temporal architectures.

1 INTRODUCTION

Video diffusion models combine scalable denoising objectives with transformer backbones and flexible spatial–temporal computation (Ho et al., 2020; Rombach et al., 2022; Peebles & Xie, 2023; Ho et al., 2022). In factorized architectures, spatial attention processes tokens within a frame while temporal attention exchanges information across frames at corresponding spatial locations (Bertasius et al., 2021; Arnab et al., 2021; Zheng et al., 2024). Because temporal attention is the only operation in such a model that exchanges information across frames, its heads are the natural unit for studying how temporal structure develops.

Existing analyses of video diffusion attention cover recurrent patterns, sparsity, sinks, temporal correspondence, and the functional split between spatial and temporal modules (Wen et al., 2025; Xi et al., 2025; Liu et al., 2025; Nam et al., 2025), but they examine final or nearly trained checkpoints, leaving unclear how temporal-attention structure develops during training. Studying this process is important for two reasons. First, specialization may occur in only a small subset of heads and therefore be obscured by averages over all heads. Second, identifying when and where such specialization emerges may provide useful signals for training diagnostics, pruning, and studying its relationship to generation quality.

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.

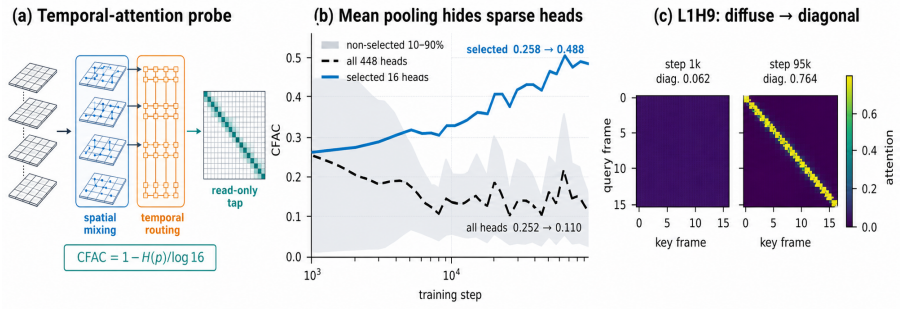


Figure 1: **Sparse temporal-attention specialization is not detectable in the population mean.** (a) Factorized STDiT and the read-only temporal probe, with CFAC summarizing each head's cross-frame attention concentration. (b) Population-mean CFAC is flat or decreasing while a small selected minority rises sharply. (c) A representative selected head transitions from uniform attention to a self-frame diagonal over training.

Studying how temporal attention develops during training requires an architecture in which temporal interactions can be measured directly and a training pipeline that can be rerun with frequent checkpointing. Open-Sora STDiT (Zheng et al., 2024) satisfies both requirements. Its factorized spatial–temporal design gives each temporal-attention head a well-defined attention distribution over key frames, so each head can be tracked individually during training. Its open training pipeline also allows us to train multiple runs under matched configurations and analyze their trajectories across checkpoints. The same analysis can be applied directly to other factorized video diffusion transformers, such as Latte and VDT (Ma et al., 2025; Lu et al., 2023). For joint spatio-temporal architectures such as Wan, LTX-Video, CogVideoX, and HunyuanVideo (Team Wan, 2025; HaCohen et al., 2024; Yang et al., 2025; Kong et al., 2024), an additional aggregation step is required to convert joint attention over spatial–temporal tokens into a frame-level attention distribution.

Figure 1 summarizes the central finding. Panel (a) illustrates the measurement setup, in which a read-only probe attached to each STDiT block records every temporal head's 16×16 query–key frame attention distribution and summarizes it with cross-frame attention concentration (CFAC, defined in Section 3). Panel (b) plots this quantity over training for 760M Run A. The mean over all heads (dashed) decreases from 0.252 to 0.110, whereas the sixteen heads later identified by our census (blue) increase from 0.258 to 0.488. The population mean therefore does not reflect the increase in the selected heads. Panel (c) shows the attention map of one such head, which is indistinguishable from uniform at step 1k and concentrates 0.764 of its mass on the self-frame diagonal by step 95k.

These findings hold across scales, but they do not support emergence at specific (layer, head) coordinates. Aggregate CFAC shows no clear transition. The selected subset is sparse in every full-grid run, though its size varies across seeds. Reproduction across runs is concentrated in the first temporal block (L0). In every seed, L0 contains selected heads and the population-level rank statistics reproduce. The specific heads selected within L0 differ across seeds. Finally, in the analyzed 760M subset, selected heads develop recurring self-frame diagonals and adjacent-frame bands even when coordinates and directions differ across runs. Our contributions are as follows.

- a checkpoint-resolved, per-head analysis of temporal-attention development across nine STDiT training runs and three model scales, using fixed selection criteria and cross-seed reproduction tests.
- evidence that specialization occurs in only a small subset of temporal-attention heads and can therefore be obscured by population-level averages. Across runs, this specialization is reproducible primarily at the block level rather than at specific head indices. Selected heads develop recurring frame-to-frame attention patterns, and negative controls separate this description from claims about generation quality or causality.
- an analysis framework consisting of the attention-concentration metric, the head-selection procedure, and the cross-seed localization tests, which applies directly to other factorized video diffusion transformers and extends to joint spatio-temporal architectures with an adapted frame-level routing metric.

2 Related Work

Diffusion models, latent diffusion, and diffusion transformers underpin modern video systems (Ho et al., 2020; Rombach et al., 2022; Peebles & Xie, 2023), with video architectures adding temporal operators or factorized spatial–temporal attention (Ho et al., 2022; Singer et al., 2023; Blattmann et al., 2023; Bertasius et al., 2021). Open-Sora STDiT is well suited for developmental analysis because its temporal-attention module is explicitly separated and its training pipeline is publicly available (Zheng et al., 2024). VDT and Latte also share the factorized design (Lu et al., 2023; Ma et al., 2025). Prior studies have analyzed attention in trained video diffusion models from several perspectives, including common attention structures, sparsity, and attention sinks (Wen et al., 2025), attention patterns for efficient inference (Xi et al., 2025), the functional roles of spatial and temporal modules (Liu et al., 2025), and temporal correspondence across frames (Nam et al., 2025). In contrast, our work focuses on how temporal attention develops during training. We track every temporal-attention head across checkpoints using fixed selection criteria and compare the resulting patterns across random seeds.

In language transformers, a minority of heads often carries interpretable structure while many are redundant (Clark et al., 2019; Voita et al., 2019; Michel et al., 2019; Kovaleva et al., 2019). Two lines of developmental work are closest to ours. Olsson et al. (2022) track checkpoints, observe induction heads forming abruptly, and connect that formation to in-context learning with causal head-level interventions in small models. Bietti et al. (2023) analyze the "birth" of the induction mechanism in a simplified two-layer transformer, including weight dynamics and supporting theory. Our study makes the opposite choice. We analyze a real, substantially larger video diffusion model at the cost of mechanistic depth. We characterize where attention concentrates and when, but we do not identify the implemented algorithm or, given the null intervention reported here, a link to capability. Complementary developmental tools include progress measures for grokking (Nanda et al., 2023), refined local learning coefficients (Wang et al., 2025), and checkpoint studies of capability-specific heads and sink formation in 1B-class language models (Xu, 2026).

Attention weights expose routing but are not, by themselves, faithful causal explanations of model output (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019; Abnar & Zuidema, 2020), and high concentration can reflect sink-like behavior rather than a meaningful relation (Xiao et al., 2024; Guo et al., 2024). We therefore use attention maps descriptively, screen for single-frame sinks, and reserve causal claims for interventions. The intervention reported here is null under the preregistered metric.

3 Experimental Setting and Analysis

3.1 Models, data, and training runs

We analyze three Open-Sora STDiT configurations trained on UCF-101 (Soomro et al., 2012) at 16 frames and 256×256 resolution. Within each STDiT transformer block, spatial attention operates over tokens within a frame, whereas temporal attention operates across frames at corresponding spatial locations (Figure 1a). With 16-frame inputs, each temporal head therefore yields a 16×16 query–key frame attention matrix per spatial location. We refer to the models by rounded size (306M, 760M, 1.03B), with 16/28/28 blocks and 16/16/21 temporal heads per block, respectively. Exact parameter counts are given in Table 5 (Appendix C). All three STDiT configurations (306M, 760M, and 1.03B) were trained from scratch. Differences in architecture, checkpoint grids, and batch size prevent interpreting cross-scale differences as pure effects of parameter count.

At each scale, Runs A, B, and C denote distinct random-seed training trajectories. The primary 760M and 1.03B analyses use Runs A and B, and the three-run extensions and the 306M analysis use all three runs. The 306M and 1.03B models have 60 log-spaced checkpoints from 1k to 50k steps. The primary 760M grids have 27 checkpoints from 1k to 95k. Their three-seed intersection has 16 checkpoints from 1k to 38k and is diagnostic because it falls below the 20-point minimum required for a registered main-analysis verdict.

All runs use bf16, HybridAdam in AdamW mode, a learning rate of 2×10^{-5}, zero weight decay, gradient clipping at 1.0, and a 50k-step cap except for the longer 760M trajectory. The configured

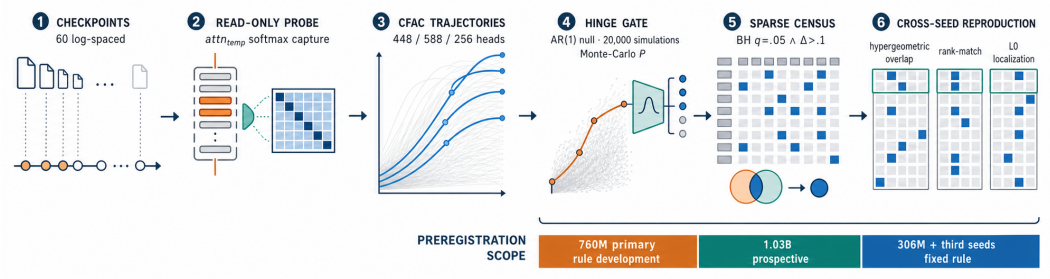


Figure 2: The preregistered census pipeline, from checkpoint probe to cross-seed localization tests. The 760M primary runs served as rule development. The fixed rule was then applied prospectively at 1.03B and unchanged in the three-seed extensions.

per-process batch size is 8 for 306M and 760M and 4 for 1.03B. Seeds and output paths differ within a scale, and the remaining recorded training configuration is held fixed.

3.2 Checkpoint probe and CFAC

We write `LxHy` for head y in temporal block x, using zero-based indices for both. For example, `L1H9` denotes head 9 in the second temporal block. The read-only probe patches `STDiTBlock.attn_temp` and captures the softmax temporal attention. Within temporal attention, the query–key logits are computed in float32 and softmax is applied over the key-frame axis. For each layer, the saved tensor has shape $(\text{groups}, \text{heads}, 16, 16)$, where groups are the video-batch times spatial positions and the final axes are query and key frames. The extractor renormalizes every row, clips probabilities at 10^{-10}, computes row entropy, and averages over all groups and all 16 query frames. Thus, for head h at checkpoint t,

$$\mathrm{CFAC}_h(t) = 1 - \frac{1}{|G|\,N} \sum_{g \in G} \sum_{q=1}^{N} \frac{H\left(p^{(t)}_{h,g,q}\right)}{\log N}, \qquad H(p) = -\sum_{i=1}^{N} p_i \log p_i, \tag{1}$$

where $p^{(t)}_{h,g,q}$ is head h's attention row for query frame q in group g of the probe batch G, $H(p)$ is the Shannon entropy of a normalized attention row p, p_i is its weight on key frame i, and $N = 16$ is the number of frame positions. CFAC is 0 for uniform temporal attention and approaches 1 as probability mass concentrates on fewer frame positions (Shannon, 1948). CFAC captures only how concentrated a head's routing across frames is, not which frames receive the mass. A head that attends to its own frame and a head that attends to a distant frame can have the same CFAC, which is why the motif analysis below inspects the attention maps directly. The unit of analysis is the head trajectory $\{\mathrm{CFAC}_h(t)\}_t$, not a mean pooled over heads. At each checkpoint, rank 0 makes one probe call per layer on a training batch of 8 with the training text conditioning, no classifier-free guidance, and a randomly drawn diffusion timestep, while a fixed-timestep re-probe uses $\tau \in \{50, 250, 500, 750, 950\}$. Spatial probes capture `STDiTBlock.attn` over $S = 256$ tokens per frame, using $1 - \langle H(p) \rangle / \log S$ per head; averaging and onset selection are detailed with Table 8 (Appendix C).

3.3 Operational birth census

Figure 2 shows the pipeline, which produces one CFAC trajectory per head (448/588/256 at the three scales). Each head trajectory is screened with a two-segment hinge model (Muggeo, 2003; Truong et al., 2020). Improvement over a single-trend model is compared with $n_{\text{null}} = 20000$ simulated trajectories from a fitted linear-trend plus AR(1) Gaussian null. The gate reports an empirical Monte Carlo p-value, direction, effect size, and hinge and half-maximum onset estimators. The main census rule is

$$\mathrm{BH}(q = 0.05) \;\wedge\; \Delta_h = \mathrm{CFAC}_h(t_{\text{last}}) - \mathrm{CFAC}_h(t_{\text{first}}) > 0.1, \tag{2}$$

where BH denotes Benjamini–Hochberg correction across heads within a run (Benjamini & Hochberg, 1995). We call heads satisfying this rule *birth heads*. The term follows earlier transformer training analyses (Bietti et al., 2023; Olsson et al., 2022), but we use it only as shorthand for

a detected concentration increase. It does not denote the onset of model-level temporal capability, a causal role in quality, or a circuit. Elsewhere we use the plainer term *selected heads*. Direction-hardened and alternative effect-size thresholds are sensitivity analyses, not substitutes for this rule. The effect-size filter creates most of the sparsity. In the four primary 760M/1.03B runs, BH alone retains 226–496 heads, whereas the conjunction retains 16–75 (Table 3, Appendix B).

The 760M primary criteria were codified after those runs had been observed and therefore constitute rule-development evidence, not prospective confirmation. The 1.03B application was prospective and confirmatory, and the 306M and Run C extensions then applied the already-fixed rule unchanged. This distinction follows the confirmatory versus exploratory role of preregistration (Nosek et al., 2018). The anonymized supplementary archive contains both preregistration documents, the metric and probe implementations, and the scripts and result files from which every reported number is recomputed, and Appendix B records the registration timeline.

We use this change-point gate as a screening step for selecting candidate heads rather than as a calibrated hypothesis test. Its AR(1) null indexes checkpoints by number even though they are log-spaced in training steps, and step-indexed surrogate checks suggest that the gate rejects more often than its nominal level. Our conclusions therefore rest on the effect-size threshold, repetition across runs, the sensitivity grid, and positional reproduction rather than on individual gate p-values.

3.4 Cross-run reproduction and motif analysis

We evaluate reproduction at two levels. A hypergeometric test compares the observed overlap between two census sets with chance among all head coordinates, and an initialization-controlled rank-match test residualizes each run's per-head net increase on the step-1,000 CFAC values of both runs and tests whether high residual ranks align (defined on the broader $\Delta_h > 0.1$ population without the BH gate, so its p-value is not directly comparable with the overlap test). To localize the signal, we repeat the rank-match test excluding L0 and test coordinate overlap conditional on both runs selecting L0 heads. Frame-attention dumps from eight 760M checkpoints provide 16×16 maps for blocks L0–L10, and we classify final maps into self-frame diagonal, adjacent-frame band, or diffuse patterns and screen for previous- or next-frame asymmetry. These are descriptive routing measurements, not causal attributions.

4 Results

4.1 Population averages obscure sparse head trajectories

Figure 3 shows why mean pooling misses the effect. The blue selected trajectories fall into two groups. One group already sits near CFAC 0.4 at the first probed checkpoint and increases steadily. A second group, including the highlighted L1H9 (red), starts inside the gray envelope of non-selected heads and rises above its upper edge only after its hinge onset, saturating above 0.6. Meanwhile the gray 10–90% band drifts downward, and the all-head mean (black dashed) follows the majority. The mean *decreases* from 0.252 to 0.110 while the 16 selected heads rise from 0.258 to 0.488 (non-selected 0.252 to 0.096). The mean curve alone gives no indication that the selected subset exists. The same decomposition appears at 1.03B Run A (mean 0.183 to 0.180, selected 0.161 to 0.313, rest 0.186 to 0.160, Appendix A), and all nine runs are collected in Figure 7 (Appendix D). This cancellation explains why the preregistered mean-level verdict is mixed while the head-level census is non-empty in every run. The two analyses measure different populations. The head-level result does not support a model-level phase transition. It shows that the model-level average is not an adequate statistic for sparse specialization.

4.2 A sparse census appears across scales, but its size is not stable

Table 1 summarizes the fixed-rule birth census. On the main analysis grids the selected population is a minority in every run, with 17/10/21 of 256 heads at 306M, 16/23 of 448 in the primary 760M runs, and 75/26 of 588 in the primary 1.03B runs (23 in Run C), a share of 3.6–12.8% in the four primary runs. The exact size is not stable. The two 1.03B runs differ by almost a factor of three, and the 760M three-seed 16-point diagnostic triplet is 15/81/16, reaching 18.1%.

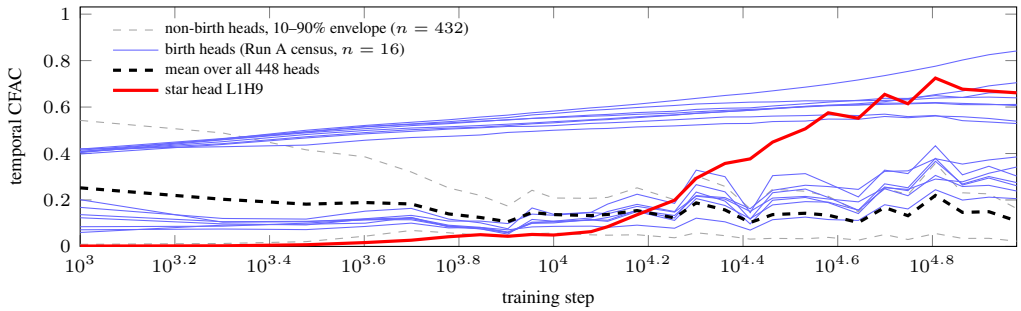


Figure 3: Per-head CFAC trajectories for 760M Run A. Blue curves are the 16 heads selected by the birth census, gray dashed curves show the 10–90% envelope of the remaining 432 heads, and the black dashed curve is the all-head mean. L1H9 (red) is a representative large transition, not a coordinate that reproduces across seeds.

By contrast, non-emptiness is stable across the sensitivity grid $q \in \{0.01, 0.05\}$ and $\delta_{\min} \in \{0.05, 0.1, 0.2\}$, and all registered rule variants retain at least three heads per run. The evidence therefore supports a sparse minority with substantial increases, not a universal percentage. The 75-versus-26 gap between the two primary 1.03B runs is the clearest instance. Many heads lie near the $\Delta_h > 0.1$ boundary, so seed-level variation in how far borderline heads rise produces large changes in census size even though sparsity, block enrichment, and population-level rank statistics reproduce (Section 4.3). Table 3 (Appendix B) decomposes the gate, and Figure 8 (Appendix D) shows census shares by scale and run.

Onsets are also spread across training rather than synchronized. In the onset raster (Figure 9, Appendix D), each 760M run's census onsets span more than an order of magnitude, from roughly 2k steps to past 60k, so the census identifies no single model-wide transition, and heads join the selected population throughout training. The five coordinates selected in both runs all lie in L0, which is consistent with the localization result of Section 4.3.

Table 1: Birth census under the fixed main rule (BH $q = 0.05$ and $\Delta_h > 0.1$).

Model	Run	Birth heads	% of temporal heads	in L0
STDiT-XL/2 (760M)	Run A	16	3.6%	8
STDiT-XL/2 (760M)	Run B	23	5.1%	8
STDiT-1B/2 (1.03B)	Run A	75	12.8%	10
STDiT-1B/2 (1.03B)	Run B	26	4.4%	12

4.3 Reproduction is block-level, not coordinate-level

The two primary 760M censuses share 5 coordinates versus 0.82 expected under a hypergeometric null ($P = 6.83 \times 10^{-4}$), and the two primary 1.03B censuses share 9 versus 3.32 expected ($P = 2.85 \times 10^{-3}$). The initialization-controlled rank-match tests are also significant ($P = 1.37 \times 10^{-8}$ at 760M and $P = 1.45 \times 10^{-10}$ at 1.03B). In the matched three-seed extension all nine within-scale pairs pass $P < 0.001$, with the 760M pairs evaluated on the diagnostic 16-point grid (Table 6, Appendix C).

Where this signal is located determines how it should be interpreted. All five shared 760M coordinates and six of the nine shared 1.03B coordinates lie in L0. When L0 is excluded, the primary initialization-controlled tests fail at 760M ($P = 0.147$) and do not meet the registered 0.001 threshold at 1.03B ($P = 0.02$). None of the nine pairwise tests in the three-seed extension passes after L0 exclusion. Within L0, coordinate overlap conditional on both runs selecting L0 heads is consistent with chance ($P = 0.310/0.575$ at 760M/1.03B).

Figure 4 shows the same result spatially. Each grid places census heads at their (layer, head) coordinates with L0 as the top row, and the top row is where the two runs agree. Every red (both-runs) 760M cell and six of the nine red 1.03B cells lie in L0. Below L0, the two runs select heads across

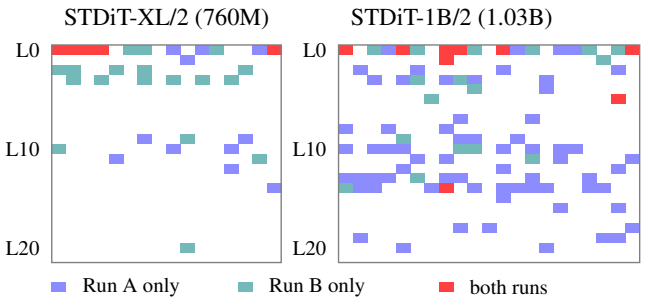


Figure 4: Census atlas of birth-head positions by layer (rows, L0 at top) and head (columns) for both scales. Red cells are in both runs' censuses, and the overlap concentrates in L0.

many layers but rarely the same coordinates, which is consistent with deeper selections being exchangeable. The supported interpretation is therefore restricted. The first temporal block repeatedly contains selected heads, but we find no evidence that training assigns stable roles to the same head indices. What reproduces is the block and the population-level ranking, not a specific coordinate such as L1H9.

4.4 Selected heads acquire recurrent frame-attention motifs

The 760M frame-attention archives cover 13 of 16 Run A and 21 of 23 Run B census heads. Every analyzed head that leaves the diffuse state ends in one of two pattern families, self-frame diagonal or adjacent-frame band. At the final checkpoint, Run A has 6 self-frame and 7 neighbor-band heads, and Run B has 14 self-frame, 6 neighbor-band, and 1 still-diffuse head (Table 9, Appendix C). None of these 34 analyzed census heads is classified as a single-frame sink at its final checkpoint.

The representative L1H9 transition is a diffuse-to-diagonal change, with diagonal mass 0.062 at step 1k, 0.070 at 7k, 0.178 at 13k, 0.517 at 26k, and 0.764 at 95k. Other selected heads instead form bands one frame behind or ahead of the query frame, and a directional screen identifies seven strong previous- or next-frame cases across the two runs, for example L0H11 in Run B (previous-frame mass 0.508) and L0H15 in Run A (next-frame mass 0.548). We describe these as copy-like motifs only at the level of the attention map, since the maps do not establish contextual matching, copying into the model output, or an induction circuit.

Figure 5 displays each motif family's endpoint as a query×key frame map. The checkpoint-resolved evolution of both families and the directional statistics are collected in Figure 10 (Appendix D) and Table 10 (Appendix C).

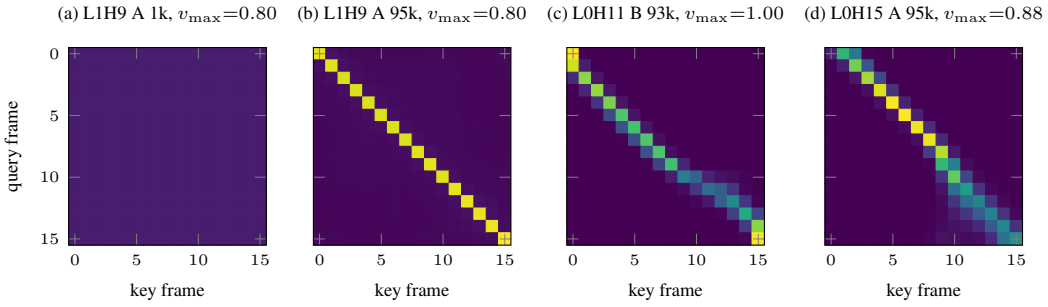


Figure 5: Mean temporal attention (query × key frame) of representative census heads, rendered from the committed heatmap archives. (a) Representative head L1H9 (Run A) at step 1000, indistinguishable from uniform (shown on the same color scale as (b)). (b) The same head at step 95000, a self-frame diagonal of mass 0.764. (c) Previous-frame copy head L0H11 (Run B) at step 93000, a band one frame behind the diagonal, mass 0.508. (d) Next-frame head L0H15 (Run A) at step 95000, a band one frame ahead, mass 0.548. Color spans 0 to the per-head maximum $v_{\max}$.

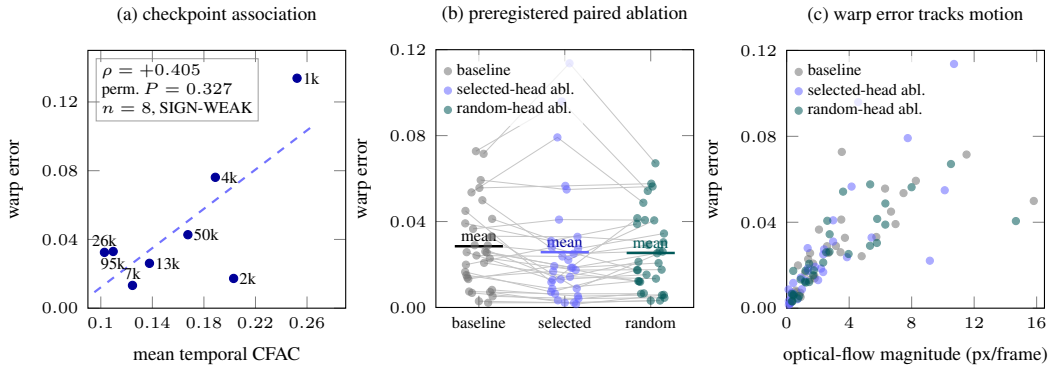


Figure 6: Controls that bound the interpretation. (a) The checkpoint-level CFAC–warp-error association is sign-consistent but not significant; $n = 8$ is the number of evaluated checkpoints (1k, 2k, 4k, 7k, 13k, 26k, 50k, 95k), not a clip count. (b) The preregistered 10-head ablation is null under warp error. (c) Warp error is strongly motion-confounded, which explains why this metric cannot establish or refute a quality link.

Motif families reproduce more reliably than coordinates. Self-frame, neighbor-band, previous-frame, and next-frame patterns all occur in both runs. The same coordinate can even reverse direction. L0H15 is next-frame-dominant in Run A and previous-frame-dominant in Run B. What is stable across runs is therefore the set of local temporal motifs, not a one-to-one mapping from coordinate to function.

4.5 Controls that bound the interpretation

In the 760M probe, the L0 CFAC maximum is approximately 0.013 at step 0 versus 0.424 at step 1000 (L1H9 has 0.00053 at step 0 and 0.6613 at the end), so the L0 pattern is absent at step 0 and present by step 1000. This control is weak because STDiT zero-initializes the temporal-attention output projection, so the temporal-attention branch initially contributes nothing through that projection. The near-uniform attention observed at step 0 is therefore an empirical property of the initialization, not a direct consequence of the zero output projection. The registered temporal-versus-spatial onset comparison is not consistent across scales. Both 760M runs show earlier temporal onsets under both estimators, and 1.03B Run A agrees, but 1.03B Run B shows a significant reversal (Table 8, Appendix C). We therefore do not infer a scale-general temporal-before-spatial ordering.

The preregistered checkpoint-level association between mean temporal CFAC and warp error is **SIGN-WEAK**, with $\rho = +0.405$ and exact permutation $P = 0.327$. The secondary selected-head series gives $\rho = +0.191$, $P = 0.665$. Warp error is also strongly coupled to motion magnitude in the intervention data (Spearman 0.909–0.914). A preregistered ablation of ten direction-hardened selected heads is NULL under both registered comparisons, with $P = 0.983$ (P1, born vs. baseline) and $P = 0.939$ (P2, born vs. random). An exploratory temporal-LPIPS comparison on the same generated clips reaches the same verdict. Both ablations reduce frame-to-frame perceptual change together with motion, and the born-versus-random contrast is indistinguishable. These analyses neither establish nor refute a causal role in temporal coherence. They show that the available motion-confounded quality metrics cannot support that claim (Appendix E). Figure 6 presents these negative results. In panel (a) the positive rank correlation reflects the joint downward drift of both quantities over training. In panel (b) the three condition means (0.0285, 0.0257, 0.0254) are nearly identical. Panel (c) shows why the metric cannot answer the question. Within every condition, warp error rises almost monotonically with optical-flow magnitude, so any intervention that reduces motion also reduces warp error, independently of temporal coherence.

5 Discussion

The main empirical result is a separation between population-level and head-level dynamics. Temporal attention shows no robust mean transition, yet a small subset of heads develops large, struc-

tured concentration increases. This extends to video diffusion the observation that specialized heads carry a disproportionate share of transformer structure (Voita et al., 2019; Michel et al., 2019). The positional result refines this. Reproduction at the level of L0 combined with chance-level coordinate matching within L0 suggests that the first temporal block is a location where selected heads consistently appear, while head identities are consistent with exchangeability at the resolution of these tests. We find no evidence that training assigns the same role to the same index across runs, and explaining the block-level preference would require interventions on optimization, residual pathways, or block placement. The motifs themselves remain hypotheses about function. Self-frame diagonals could preserve frame-local information, and adjacent bands could route local motion, but attention probabilities do not reveal how downstream computation transforms the routed information (Jain & Wallace, 2019; Abnar & Zuidema, 2020). The null ablation and the weak quality correlation also argue against a capability claim.

We now consider how far these observations extend. Three nested hypotheses are consistent with our evidence. The dynamics could be specific to Open-Sora, general to factorized architectures, or a property of video DiT training in general. We claim only the first, but the methodology is designed to test the other two. For Latte and VDT (Ma et al., 2025; Lu et al., 2023) the probe, CFAC, and census rule apply unchanged, and for joint architectures (Team Wan, 2025; HaCohen et al., 2024; Yang et al., 2025; Kong et al., 2024) marginalizing full attention over spatial positions yields the required per-head routing distribution. If sparse, block-localized onsets appear there, they would indicate a general feature of video diffusion training. If they do not, the effect would depend on factorization.

6 Limitations

All runs use UCF-101 and one Open-Sora STDiT family. Because the experiments cover a single family, we cannot determine whether the observed dynamics are Open-Sora-specific, factorized-general, or general to video DiTs, and we accordingly position the contribution as a preregistered methodology plus a systematic empirical case study rather than an architecture-general claim. The analysis transfers to Latte and VDT (Ma et al., 2025; Lu et al., 2023) but is untested there, and joint architectures require a derived routing statistic first. Three seeds per scale support repeatability, not the full variability over data orderings or initialization (Bouthillier et al., 2021), and the scales differ in architecture, grids, and batch size, so no result is a controlled scaling law.

The change-point null is imperfect for log-spaced checkpoints, and head-level tests are dependent. Benjamini–Yekutieli correction preserves the non-empty census and positional-overlap conclusions with reduced counts (Benjamini & Yekutieli, 2001), and reported BH values remain nominal (Appendix B). Each checkpoint uses one probe pass, and CFAC is sensitive to the scale of attention logits. A post-hoc fixed-timestep re-probe found the per-head Δ_h ordering stable across timesteps (median pairwise Spearman ρ 0.71–0.93 per run), but uniform logit sharpening can raise concentration without reordering attended frames, so scale-invariant routing statistics are needed to separate these effects. Attention maps describe routing rather than causal computation. The available warp metric is motion-confounded, and the ablation is underpowered to identify subtle or compensatory functions. No causal or perceptual-quality conclusion is supported. Table 11 (Appendix C) summarizes the claim boundary.

7 Conclusion

We studied how temporal-attention structure arises during video diffusion training, using a checkpoint-resolved census of every temporal head across nine Open-Sora STDiT runs at three scales. Temporal-attention organization is better described as sparse head-level specialization than as a model-wide transition. A minority of heads gains substantial concentration while population averages stay flat or decline. The reproducible positional signal is the first temporal block rather than particular coordinates. Attention maps among the analyzed 760M selected heads converge to a small set of local frame-routing motifs. Correlation and ablation analyses do not establish a quality or causal link. The preregistered methodology transfers directly to other factorized video DiTs and, with an adapted routing statistic, to joint architectures. Applying it there, with interventions that are robust to motion confounds, is the natural next step.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020. doi: 10.18653/v1/2020.acl-main.385.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. ViViT: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6816–6826, 2021. doi: 10.1109/ICCV48922.2021.00676.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001. doi: 10.1214/aos/1013699998.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 813–824, 2021.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. In *Advances in Neural Information Processing Systems*, volume 36, pp. 1560–1588, 2023. doi: 10.52202/075280-0077.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22563–22575, 2023. doi: 10.1109/CVPR52729.2023.02161.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3, 2021.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP*, pp. 276–286, 2019. doi: 10.18653/v1/W19-4828.

Tianyu Guo et al. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in LLMs. *arXiv preprint arXiv:2410.13835*, 2024.

Yoav HaCohen et al. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646, 2022. doi: 10.52202/068431-0628.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of NAACL-HLT*, pp. 3543–3556, 2019. doi: 10.18653/v1/N19-1357.

Weijie Kong et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of BERT. In *Proceedings of EMNLP-IJCNLP*, pp. 4364–4373, 2019. doi: 10.18653/v1/D19-1445.

Bingyan Liu, Chengyu Wang, Tongtong Su, Huan Ten, Jun Huang, Kailing Guo, and Kui Jia. Understanding attention mechanism in video diffusion models. *arXiv preprint arXiv:2504.12027*, 2025. doi: 10.48550/arXiv.2504.12027.

Haoyu Lu et al. VDT: General-purpose video diffusion transformers via mask modeling. *arXiv preprint arXiv:2305.13311*, 2023.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *Transactions on Machine Learning Research*, 2025.

Henry B. Mann and Donald R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947. doi: 10.1214/aoms/1177730491.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Vito M. R. Muggeo. Estimating regression models with unknown break-points. *Statistics in Medicine*, 22(19):3055–3071, 2003. doi: 10.1002/sim.1545.

Jisu Nam, Soowon Son, Dahyun Chung, Jiyoung Kim, Siyoon Jin, Junhwa Hur, and Seungryong Kim. Emergent temporal correspondences from video diffusion transformers. *arXiv preprint arXiv:2506.17220*, 2025. doi: 10.48550/arXiv.2506.17220.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023. Oral presentation; arXiv:2301.05217.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018. doi: 10.1073/pnas.1708274114.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4172–4182, 2023. doi: 10.1109/ICCV51070.2023.00387.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27 (3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-A-Video: Text-to-video generation without text-video data. In *International Conference on Learning Representations*, 2023.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. In *CRCV-TR-12-01*, 2012.

Team Wan. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020. doi: 10.1016/j.sigpro.2019.107299.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, 2019. doi: 10.18653/v1/P19-1580.

George Wang, Jesse Hoogland, Stan van Wingerden, Zach Furman, and Daniel Murfet. Differentiation and specialization of attention heads via the refined local learning coefficient. In *International Conference on Learning Representations*, 2025.

Yuxin Wen, Jim Wu, Ajay Jain, Tom Goldstein, and Ashwinee Panda. Analysis of attention in video diffusion transformers. *arXiv preprint arXiv:2504.10317*, 2025. doi: 10.48550/arXiv.2504.10317.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of EMNLP-IJCNLP*, pp. 11–20, 2019. doi: 10.18653/v1/D19-1002.

Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, Jianfei Chen, Ion Stoica, Kurt Keutzer, and Song Han. Sparse VideoGen: Accelerating video diffusion transformers with spatial-temporal sparsity. In *International Conference on Machine Learning*, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024.

Yongzhong Xu. When do attention circuits form? developmental trajectories of capability and attention-sink emergence across three 1b-class architectures. *arXiv preprint arXiv:2606.02378*, 2026. doi: 10.48550/arXiv.2606.02378.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

AI USE STATEMENT

Generative AI tools were used to support literature search, manuscript restructuring, language editing, and critical review of methodological presentation and result interpretation. They were not used to generate experimental data or execute the reported training runs. The authors reviewed all numerical claims, statistical conclusions, citations, and interpretations and take responsibility for the manuscript.

ETHICS STATEMENT

UCF-101 is a public academic action-recognition dataset (Soomro et al., 2012). We used it to train small research models for analysis, not to deploy a video-generation system. Nevertheless, video-generation methods can be misused to create deceptive or non-consensual media. This study does not remove that risk, and downstream use requires appropriate safeguards and compliance with the dataset's terms.

REPRODUCIBILITY STATEMENT

The Methods section specifies the models, checkpoint grids, read-only probe, change-point screening gate, preregistered decision rules, and seed comparisons. The appendices report robustness checks and full supporting tables, and Appendix B records the registration timeline with the pre-observation commit hashes. The anonymized supplementary archive contains the preregistration documents, the metric and probe implementations, the analysis script behind each reported result, and the result files those scripts consume. Table 4 (Appendix C) records the available compute metadata. Its times are Slurm requested wall-time ceilings, not measured elapsed times. Because per-run probe limits were not archived uniformly, the documented minimum–maximum range is reported explicitly.

A Why Mean-Pooled Metrics Miss Sparse Specialization

Table 2 decomposes the first-to-last mean CFAC change (Run A, both scales) into census and non-census heads. The census group rises substantially at both scales, but it is 3.6–12.8% of heads, and the non-census majority is flat to falling (entropy rising) over the same interval, so the all-head mean moves little and can even fall while the census heads are rising. This is the arithmetic reason the mean-level D1 verdict (MIXED) and the head-level test passing at both scales coexist without contradiction. They measure different populations.

Table 2: Mean CFAC at the first and last checkpoints, by head group (Run A, with first/last = 1k/95k at 760M and 1k/50k at 1B). Computed from the committed trajectory files.

Model (Run A)	Head group	n	CFAC first	CFAC last	delta
STDiT-XL/2	all	448	0.252	0.110	-0.143
STDiT-XL/2	birth (census)	16	0.258	0.488	+0.230
STDiT-XL/2	non-birth	432	0.252	0.096	-0.156
STDiT-1B/2	all	588	0.183	0.180	-0.003
STDiT-1B/2	birth (census)	75	0.161	0.313	+0.152
STDiT-1B/2	non-birth	513	0.186	0.160	-0.025

B Exploratory Robustness of the Census Rule

The 760M criteria were codified after those runs were observed, and the same rule was fixed prospectively for the confirmatory 1B application and then applied unchanged to the 306M and Run C extensions. The checks below are *exploratory* under the registration's change-control clause. They are reported alongside, and do not replace, any registered verdict.

The registration timeline is as follows. The decision criteria were recorded in `analysis/dense_reprobe_760m/PREREGISTRATION_D1D2D3.md` on 2026-07-08, before any 1B probe or sampling observation. The initial registration and the pre-observation Amendment 1 correspond to commits `35f1a12` and `07f670f`, and the file later moved in commit `e92c8c5`, so its history requires `git log -follow`. Because the repository is private, reviewers cannot resolve these hashes from the submission alone, and the anonymized supplementary archive therefore carries the referenced material directly. Post-observation criteria changes are prohibited by the document's change-control clause, and post-hoc tests may only be labeled exploratory.

The main rule is a conjunction of a BH-adjusted gate and a net-rise filter $\Delta_h > 0.1$, and the two components contribute very unequally to the selection. Applied alone to the committed per-head gate files, BH at $q = 0.05$ retains a majority of heads, while $\Delta_h > 0.1$ retains a small minority, and the census is their intersection (Table 3). The change-point gate p-values are not established as calibrated, so the $q = 0.05$ BH adjustment is treated here as nominal rather than as a formal false-discovery-rate guarantee. The reported census is a further data-dependent restriction of the BH-adjusted set with no separately stated error rate. Sparsity is a property of the net-rise threshold, not of the multiplicity correction.

Head-level tests within a run share batch order, learning-rate schedule, and layer inputs, so BH's dependence conditions are an assumption (Section 6). Replacing BH with the Benjamini–Yekutieli procedure, which addresses arbitrary dependence under standard p-value validity assumptions (Benjamini & Yekutieli, 2001), shrinks every census but leaves both conclusions intact. The census stays non-empty at both scales, and cross-run positional overlap stays above chance, with $P = 2.59 \times 10^{-4}$ at 760M (4 shared of 12/14, 0.375 expected) and $P = 3.46 \times 10^{-3}$ at 1B (5 shared of 33/20, 1.12 expected), against 6.83×10^{-4} and 2.85×10^{-3} under BH.

The gate's p-value is the exceedance fraction over 20,000 simulated trajectories and is not floored, so a head whose statistic exceeds every simulated value receives $p = 0$ exactly (138 of 588 heads at 1B Run A). Recomputing every p-value with the conventional $(r+1)/(B+1)$ estimator changes no census membership and no overlap statistic at either scale. The censuses remain $16/23$ and $75/26$ and the overlap p-values are unchanged to the reported precision. The resolution floor therefore affects only how the p-values are presented and does not change any result.

Table 3: Exploratory decomposition of the census rule and its arbitrary-dependence variant, recomputed from the committed per-head gate files. "BH" and "BY" are applied at $q = 0.05$, and "census" is the preregistered conjunction with $\Delta_h > 0.1$.

Series	heads	BH alone	$\Delta_h > 0.1$ alone	census (BH $\wedge$ Δ_h)	BY $\wedge$ Δ_h
760M Run A	448	232	28	16	12
760M Run B	448	226	69	23	14
1B Run A	588	369	153	75	33
1B Run B	588	496	44	26	20

C Detailed Statistical Results

This appendix collects the full tables behind the main-text results, each introduced by the conclusion it supports.

Table 4 records the compute allocation for the nine training runs and the checkpoint probes. Times are requested Slurm wall-time ceilings rather than measured elapsed times, as noted in the reproducibility statement.

Table 4: Recorded compute allocation. All jobs used the `h100` queue on TACC Stampede3. Times are requested Slurm wall-time ceilings, not measured elapsed time.

Workload	Runs	Run IDs	H100 nodes/job	Time limit/job
306M training	3	A/B/C	2	$\leq$ 26 h
760M training	3	A/B/C	2	$\leq$ 26–44 h
1B training	3	A/B/C	4	$\leq$ 22 h
Checkpoint attention probes	9	corresponding runs	1	$\leq$ 55 min–4 h

Table 5 gives the exact parameter counts, block counts, per-block temporal head counts, and checkpoint grids for the three model configurations. The 760M three-seed common grid has 16 checkpoints, below the 20-point minimum required for a registered main-analysis verdict, which is why its three-seed census is reported as diagnostic.

Table 5: Models and checkpoint grids. Parameter counts are exact. The 760M three-seed common grid has 16 checkpoints and is diagnostic.

Architecture	Parameters (exact)	Temporal heads (blocks x heads)	Checkpoints (range)	Runs
STDiT	306,874,592	16 x 16 = 256	60 (1k-50k)	Run A / Run B / Run C
STDiT-XL/2	759,620,768	28 x 16 = 448	27 primary (1k-95k), 16 common (1k-38k)	Run A / Run B / Run C
STDiT-1B/2	1,032,795,488	28 x 21 = 588	60 common (1k-50k)	Run A / Run B / Run C

Table 6 reports the matched three-scale, three-seed reproduction. All nine within-scale initialization-controlled pairwise tests pass $P < 0.001$, and all nine fail once L0 is excluded, which is the basis for the block-level scope of the positional claim.

Table 6: Matched three-scale, three-seed reproduction. Census counts follow Run A/B/C order, and each p-value column lists the three within-scale run-pair comparisons in the order A–B, A–C, B–C. The 760M census is diagnostic because its 16-point common grid is below the 20-point minimum required for a registered main-analysis verdict.

Scale	Grid	Census (A/B/C)	Init-controlled pairwise P	L0-excluded pairwise P
306M	60 (main)	17/10/21	4.99e-09 / 4.22e-10 / 1.38e-09	$>$ 0.99 / 0.23 / 0.136
760M	16 (diag.)	15/81/16	3.42e-05 / 7.2e-12 / 3e-05	0.0737 / 0.058 / 0.26
1B	60 (main)	75/26/23	1.45e-10 / 4.47e-09 / 1.26e-12	0.02 / 0.322 / 0.801

Table 7 lists the preregistered positional tests for the primary run pairs, covering the hypergeometric census overlap, the initialization-controlled rank-match test, and the registered L0-exclusion scope check.

Table 7: Positional-reproduction tests (preregistered tests (a) and (b), plus the registered L0-exclusion scope check). Test (a) is the hypergeometric tail probability of the census intersection over all temporal heads. Test (b) regresses each run's per-head net increase Δ_h on an intercept and the step-1k CFAC values of both runs, ranks the residuals, keeps the top k heads per run with k equal to that run's number of heads with raw $\Delta_h > 0.1$, and applies the same hypergeometric test to the two top-k sets. The L0-excluded row repeats test (b) with the regression, ranking, and k recomputed on the non-L0 heads.

Model	Positional test	Overlap	$P[X \geq k]$
STDiT-XL/2 (760M)	census overlap (a)	5 vs 0.82 expected	$P = 6.83 \times 10^{-4}$
STDiT-XL/2 (760M)	init-controlled rank-match (b)	17 vs 4.31 expected	$P = 1.37 \times 10^{-8}$
STDiT-XL/2 (760M)	init-controlled, L0 excluded	4 vs 2.1 expected	$P = 0.147$
STDiT-1B/2 (1.03B)	census overlap (a)	9 vs 3.32 expected	$P = 2.85 \times 10^{-3}$
STDiT-1B/2 (1.03B)	init-controlled rank-match (b)	31 vs 11.45 expected	$P = 1.45 \times 10^{-10}$
STDiT-1B/2 (1.03B)	init-controlled, L0 excluded	12 vs 6.77 expected	$P = 0.02$

Table 8 reports the registered temporal-versus-spatial onset comparisons. Both 760M runs and 1.03B Run A show earlier temporal onsets, while 1.03B Run B shows a significant reversal, which rules out a scale-general ordering claim. Spatial attention spans 256 tokens per frame (32×32 VAE latents with 2×2 patches). The probe saves softmax probabilities for a capped prefix of batch/frame groups in float16. The extractor renormalizes each row, clips probabilities below 10^{-10}, and averages row entropy over all query tokens and retained groups before normalizing by $\log 256$.

Table 8: Onset ordering (temporal vs. spatial series), preregistered main population and tests. n (t/s) counts the temporal and spatial heads that pass the registered onset population rule in each run, BH $q = 0.05$ on the change-point gate and $\Delta_h > 0.05$, applied separately to the temporal and spatial per-head series; this rule is looser than the $\Delta_h > 0.1$ census, so the temporal counts differ from Table 1. MWU = two-sided tie-corrected Mann–Whitney (Mann & Whitney, 1947), and perm = 20,000 rank-sum label permutations.

Model	Run	Estimator	n (t/s)	median onset (t/s)	MWU p	perm p	temporal first?
STDiT-XL/2	Run A	hinge	23/145	10000/18000	0.004849	0.0044	yes
STDiT-XL/2	Run A	half-max	23/145	18000/56000	4.723×10^{-6}	5×10^{-5}	yes
STDiT-XL/2	Run B	hinge	40/80	10000/25000	0.008603	0.0094	yes
STDiT-XL/2	Run B	half-max	40/80	6000/33000	1.116×10^{-6}	5×10^{-5}	yes
STDiT-1B/2	Run A	hinge	136/92	10183/14185	0.01009	0.00975	yes
STDiT-1B/2	Run A	half-max	136/92	19762/21116	0.004832	0.00425	yes
STDiT-1B/2	Run B	hinge	68/152	9529/4910	4.571×10^{-5}	0.0001	no (reversed)
STDiT-1B/2	Run B	half-max	68/152	5991/3299	3.004×10^{-12}	5×10^{-5}	no (reversed)

Table 9 counts the final-checkpoint motifs among the analyzed census heads, supporting the two-family description in Section 4.4. Classification uses the attention matrix averaged over batch/spatial positions. We compare mean self-frame, previous-frame, and next-frame mass with the largest off-diagonal column mean (the sink score). Previous/next means use only the 15 valid query–key pairs, without wraparound; each column mean excludes its diagonal entry and averages over the other 15 queries. The largest feature determines the label if it is at least $2/16 = 0.125$. Otherwise the label is diffuse/uniform when mean row entropy divided by $\log 16$ exceeds 0.90, and weak/mixed otherwise. Previous- and next-frame labels are grouped as adjacent-frame bands in the table.

Table 9: Final-checkpoint motif counts among analyzed census heads (760M line, where the frame-attention dumps cover temporal blocks L0–L10).

Final-checkpoint motif	Run A (13 analyzed)	Run B (21 analyzed)
self-frame (diagonal)	6	14
neighbor band, prev-dominant class	5	5
neighbor band, next-dominant class	2	1
still diffuse/uniform	0	1

Table 10 lists the seven heads passing the directional band-asymmetry screen, including the coordinate L0H15 whose preferred direction reverses between runs.

Table 10: Directional previous/next-frame-copy (induction-like) subtype, covering census heads with final-checkpoint band asymmetry $|\mathrm{asym}| > 0.3$ and dominant ± 1 band mass above 0.125 (twice the uniform per-frame mass $1/16$). Here prev and next are the mean attention mass on the key frame one step behind or ahead of the query frame, $\mathrm{asym} = (\mathrm{prev} - \mathrm{next})/(\mathrm{prev} + \mathrm{next})$, and positive values denote previous-frame dominance. This screen is applied separately from the final-motif classification of Table 9, so a head whose largest mass is on the self-frame diagonal (final motif "self-frame(diag)") can still pass it.

Run	Head	prev band	next band	asymmetry	subtype	final motif
Run A	L0H8	0.240	0.118	+0.34	prev-copy	prev-frame(copy)
Run A	L0H14	0.331	0.156	+0.36	prev-copy	self-frame(diag)
Run A	L0H15	0.107	0.548	-0.67	next-copy	next-frame
Run A	L10H12	0.304	0.138	+0.38	prev-copy	prev-frame(copy)
Run B	L0H1	0.127	0.407	-0.53	next-copy	self-frame(diag)
Run B	L0H11	0.508	0.153	+0.54	prev-copy	prev-frame(copy)
Run B	L0H15	0.428	0.080	+0.68	prev-copy	prev-frame(copy)

Table 11 summarizes the claim boundary, stating for each potential claim whether it is made and the evidence or reason.

Table 11: Claim boundary. D1, D2, and D3 are the three preregistered decisions (mean-level temporal emergence at 1.03B, per-head reproduction at 760M, and the metric–quality correlation). The mean-gate criterion applies the calibrated change-point gate to the head-averaged CFAC series and requires a significant positive break with an interior break point and a positive post-break slope; the argmax criterion requires the series maximum to fall after the first checkpoint; and the two-run rule requires both primary runs to pass, otherwise the decision is MIXED.

Claim	Status	Evidence / reason
Temporal attention forms by sparse head birth	claimed	primary census 16/23 of 448 and 75/26 of 588, and 306M three-seed census 17/10/21 of 256
Positional reproduction, restricted scope	claimed (L0 block + population statistics only)	all nine three-seed init-controlled pairwise tests have P<0.001, and all nine fail with L0 excluded
Mean-level temporal emergence	not claimed	preregistered D1 = MIXED, since 1B Run A satisfies the mean-gate criterion (U-shaped recovery) but Run B fails the argmax criterion, and the two-run rule blocks the claim
Temporal-before-spatial onset ordering as a scale-general law	not claimed	1B Run B shows a significant reversal, and the 760M observation is reported as scale-local
Cross-run reproduction of individual head coordinates outside L0	not claimed	none of the nine within-scale seed pairs passes P<0.001 after L0 exclusion
Metric-quality correlation (D3)	not claimed (SIGN-WEAK)	rho=+0.405, exact permutation P=0.327, n=8, and the CFAC-quality link is not established
Causal role of the born heads	not claimed (ablation NULL)	10-head direction-hardened set, n=32, born-vs-baseline warp P=0.983, and warp error is strongly motion-confounded

D Additional Data-Derived Figures

The census atlas and the motif endpoint maps appear in the main text (Figures 4 and 5).

Figure 7 repeats the decomposition of Figure 3 for all nine runs, one panel per run with the census size in each title. The same pattern appears in every panel. A sparse set of blue selected trajectories rises above the gray 10–90% band while the all-head mean (dashed) stays flat or declines. The panels also visualize the census-size instability discussed in Section 4.3. The 1.03B Run A panel shows its 75-head census as a dense set of trajectories, many of which lie near the effect-size boundary, whereas the neighboring 1.03B panels select sparser, more separated sets. The 760M Run C panel is computed on the diagnostic 16-point grid and is marked as such.

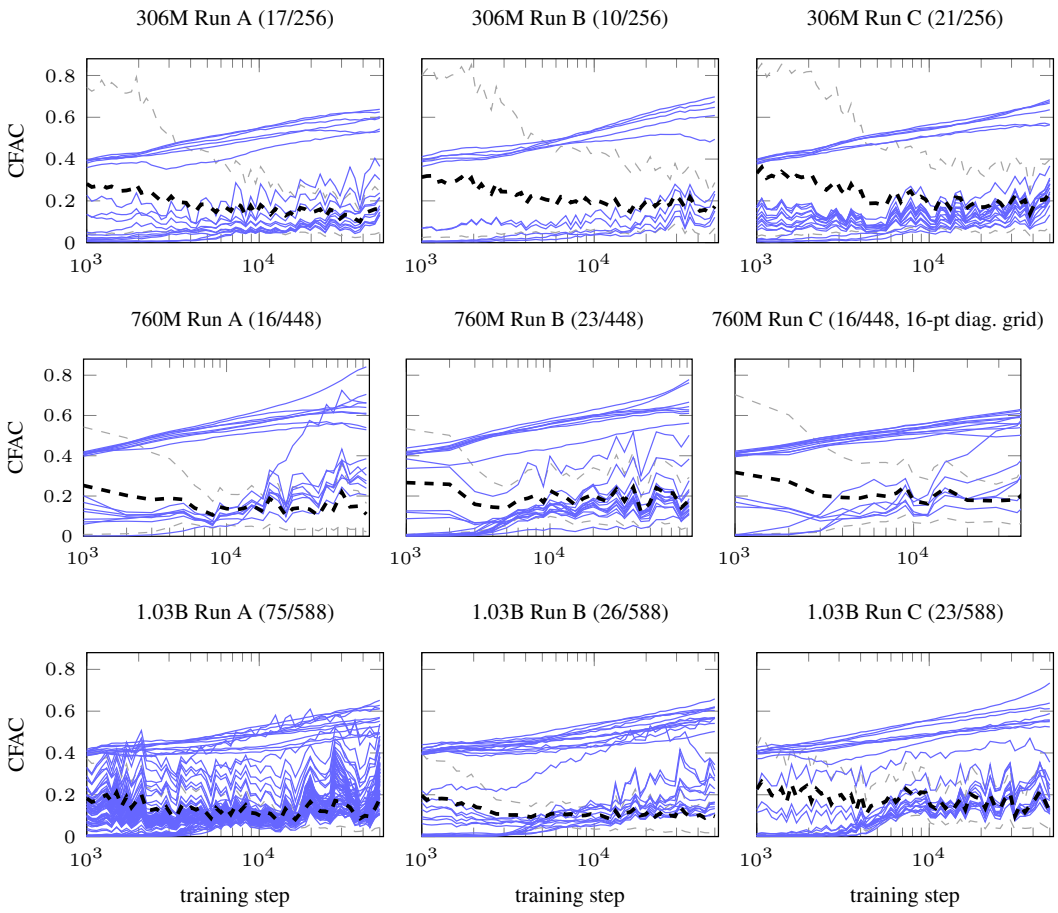


Figure 7: Per-head CFAC trajectories for all nine runs. The sparse selected-head rise against a flat or declining population reproduces in every run. Figure 3 shows the 760M Run A panel at full size.

Figure 8 plots the census share of temporal heads for the four primary runs (760M and 1.03B, Runs A and B). Sparsity holds across scales, while the share itself varies by run and scale, which is consistent with the census-size instability noted above.

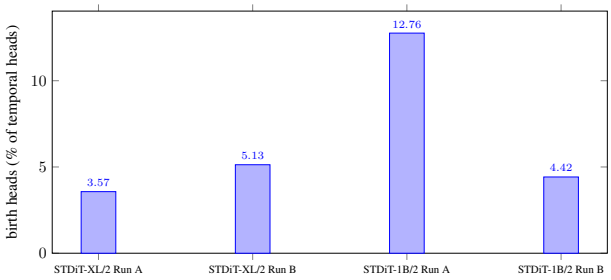


Figure 8: Birth-census share of temporal heads in the four primary runs (760M and 1.03B, Runs A and B). Sparsity holds in every run, while the share varies by run and scale.

Figure 9 shows the onset raster behind the timing discussion in the main text. Each row is one 760M census head and each mark its hinge onset on a logarithmic axis. Onsets span roughly 2k to past 60k steps in both runs, and the red rows, the five cross-run overlap coordinates, all carry L0 indices.

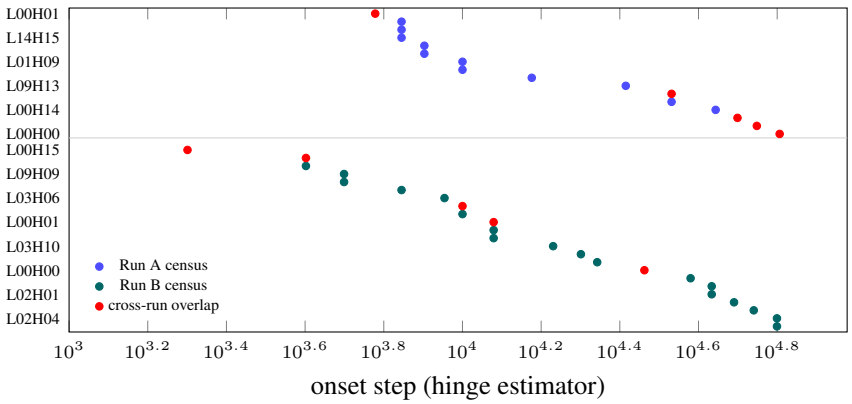


Figure 9: Birth-onset raster (hinge estimator), STDiT-XL/2 (760M), Runs A/B. Rows follow increasing onset time within each run, and head labels are subsampled. Red marks denote the five shared head coordinates, all in L0.

The clearest individual trajectory is 760M Run A head L1H9 (CFAC 0.00147 at 1k to 0.6613 at 95k, peak 0.725, hinge onset near 10k, saturation after 50k). We treat it only as an interpretable example of the selected class. Figure 10 adds the time axis that the endpoint maps of Figure 5 omit. In the top row, L1H9 is flat at steps 1k and 7k (diagonal mass 0.062 and 0.070, at the uniform baseline), a faint diagonal becomes visible at 13k (0.178), dominates the map by 26k (0.517), and is saturated by 50k (0.760) with essentially no further change to 95k (0.764). This is an abrupt transition. In the bottom row, L0H11 instead starts from an already-broad self-frame band at 1k (previous-frame mass 0.266) and sharpens gradually and monotonically toward the previous-frame copy motif (0.287 at 13k, 0.382 at 49k, 0.508 at 93k). The two motif families thus reach their final forms on different schedules. One changes abruptly and the other sharpens steadily.

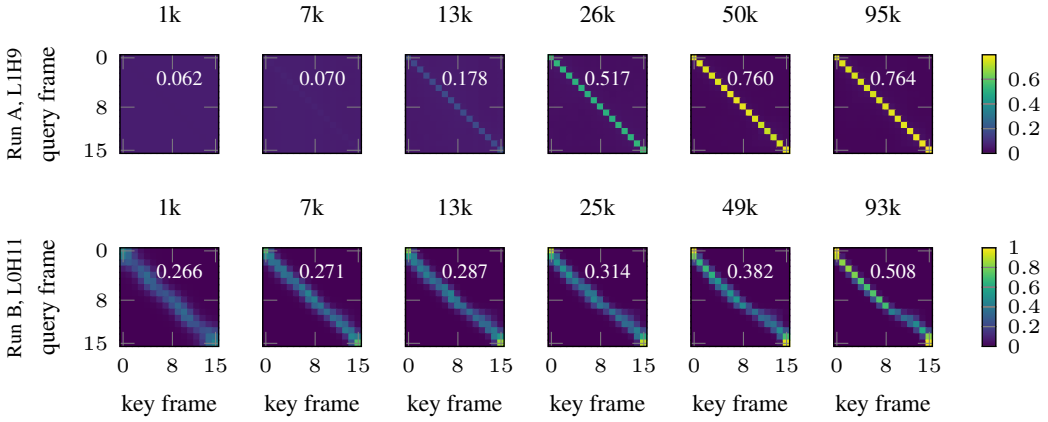


Figure 10: Checkpoint-resolved evolution of two representative census heads. In the top row, Run A L1H9 transitions from diffuse attention to a self-frame diagonal (diagonal mass 0.062 at 1k to 0.764 at 95k). In the bottom row, Run B L0H11 sharpens from a broad self-frame band toward a previous-frame copy band (previous-frame mass 0.266 at 1k to 0.508 at 93k). The relevant mass is printed in each panel.

E Causal Ablation of the Born Heads, a Preregistered Null with a Metric Confound

The main results are descriptive. To test a causal prediction, we preregistered an intervention before video generation (`analysis/ablation/PREREG_HEADS.md`). The *born* intervention

uses the 10-head direction-hardened Run A set (BH $q = 0.05$, $\Delta_h \geq 0.1$, direction $= +1$), not the 16-head primary census. Its coordinates are L0H10, L0H15, L1H9, L9H6, L9H13, L10H8, L10H12, L11H4, L12H12, and L14H15. The random control uses 10 disjoint non-census heads drawn with fixed seed 20260718: L2H15, L5H10, L12H13, L13H7, L13H15, L14H7, L19H4, L22H8, L23H0, and L25H10. Baseline, born-ablated, and random-ablated conditions generate the same 32 prompts with identical sampled noise. Ablation zeroes selected per-head outputs before the temporal-attention output projection (Michel et al., 2019; Voita et al., 2019).

The primary one-sided paired tests ask whether born-head ablation raises warp error relative to baseline (P1) and random-head ablation (P2). Neither approaches significance. Mean warp error is 0.0285 (baseline), 0.0257 (born-ablated), and 0.0254 (random-ablated), with $P = 0.983$ for P1 and $P = 0.939$ for P2. Under the fixed decision grid the verdict is NULL, and at $n = 32$ the experiment does not detect any degradation from born-head ablation. Random ablation instead lowers warp error relative to baseline (report-only two-sided $P = 0.00022$). Table 12 reports the full paired results.

As a post-hoc motion check, mean optical-flow magnitude is 3.549 px/frame at baseline, 2.654 after born-head ablation ($P = 0.00549$), and 3.227 after random-head ablation ($P = 0.00409$). The born-vs-random motion contrast is not significant ($P = 0.062$). Within each condition, motion magnitude and warp error are almost monotone (Spearman 0.909–0.914). Warp error therefore decreases whenever motion decreases, so it is confounded for this intervention.

As a post-hoc, non-preregistered complement to the warp metric, we compute per-clip temporal LPIPS (Zhang et al., 2018) (mean AlexNet LPIPS between consecutive frames, `analysis/ablation/temporal_lpips.py`) on the same 96 stored clips. Means are 0.140 (baseline), 0.119 (born-ablated), and 0.127 (random-ablated). Paired two-sided sign-flip permutation tests ($n = 32$, 200,000 resamples) give $P = 0.027$ for born-ablated versus baseline, $P = 0.00002$ for random-ablated versus baseline, and $P = 0.378$ for born-ablated versus random-ablated. Both ablations reduce frame-to-frame perceptual change relative to baseline, consistent with the motion reduction above, and the born-versus-random contrast remains indistinguishable. This exploratory check does not alter the preregistered NULL.

We retain the preregistered NULL, make no causal claim in either direction, and do not claim that the born heads uniquely drive motion. The result explains why D3 and the intervention cannot establish a perceptual-quality link with the current metric. It does not alter the descriptive birth, block-enrichment, or motif results.

Table 12: Head-ablation results over 32 paired clips. P1 (born vs. baseline) and P2 (born vs. random) are the preregistered one-sided paired tests of whether ablation raises warp error. The random-vs.-baseline warp comparison is report-only and two-sided, and the motion analyses are post hoc and two-sided.

Condition	Mean warp	p vs. base (warp)	p vs. random (warp)	Mean $\lvert\text{flow}\rvert$	p vs. base (motion)
baseline	0.0285	n/a	n/a	3.549	n/a
born-ablated	0.0257	0.983 (P1, one-sided)	0.939 (P2, one-sided)	2.654	0.00549
random-ablated	0.0254	0.00022 (two-sided)	n/a	3.227	0.00409