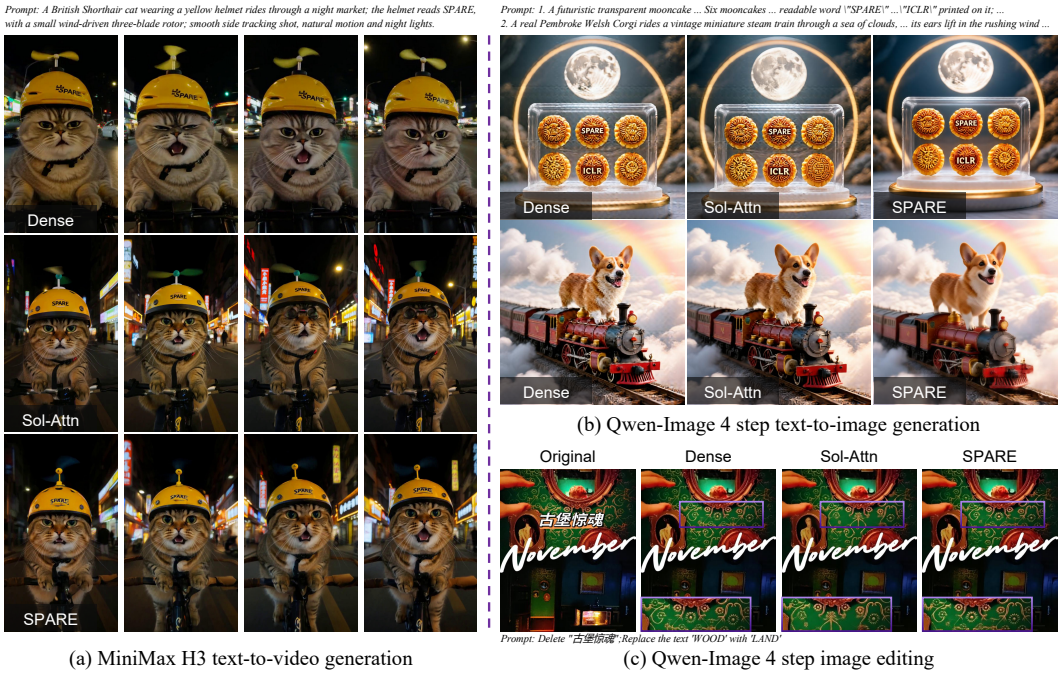
# RETHINKING SPARSE ATTENTION IN FEW-STEP DIFFUSION TRANSFORMER

**Anonymous authors**
Paper under double-blind review




(a) MiniMax H3 text-to-video generation

(b) Qwen-Image 4 step text-to-image generation

(c) Qwen-Image 4 step image editing

Figure 1: **Illustrative generation and editing examples.** Dense attention, Sol-Attn, and SPARE on (a) MiniMax H3 video generation, (b) four-step Qwen-Image generation, and (c) image editing. These examples illustrate output behavior; the matched quantitative evidence uses the Wan2.2 protocol in Section 5.1.

## ABSTRACT

Few-step diffusion compresses a multi-step teacher trajectory into a handful of student evaluations. Sparse attention further reduces cost by retaining only highly ranked interactions, but the student's current ranking need not represent the corresponding teacher-window support. On aligned probes of a four-step Wan2.2 student, the top-$b$ pool captures 51.91% of teacher attention mass and leaves a value-weighted residual. The useful core favors persistent interactions; its complement contains both missed salient blocks and a diffuse tail. This diagnosis motivates SPARE Attention: current-score candidates plus a small complement quota under a fixed sparse budget, with coarse routing and no teacher access at inference. A matched per-step-budget intervention reduces final Teacher40 RMSE from 0.959537 to 0.928340 (3.25%), improving all six reported records; larger quotas are not uniformly better. Separately, the original system configuration selects 14.84% of key blocks on the full 6,220-video VBench suite and denoises in 29.565 s, the shortest measured time among the timed sparse implementations, with Total comparable to the SLA family and PISA.

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.

# 1 INTRODUCTION

Few-step distillation makes video diffusion substantially cheaper by reducing the number of denoising evaluations. Each remaining evaluation, however, still processes a long video sequence. Sparse attention addresses this cost by restricting expensive interactions to a small set of key blocks. The usual rule is simple: rank keys by the student's current-step scores and compute only the highest-ranked blocks. We ask whether this local ranking is sufficient for a student that must approximate an entire multi-step teacher trajectory. A teacher-window probe provides a diagnostic of omitted interactions; the target of the sparse policy remains alignment with the final teacher endpoint.

A current-step score is a statistic of the student's compressed trajectory, whereas the corresponding teacher window contains several fine-grained state transitions. The teacher can place value-weighted attention contribution on blocks that the student does not rank highly, including a diffuse tail that need not enter the teacher's own top set. If the student ranking is used as a hard support boundary, this contribution becomes a support residual. The question is not whether the student ranking is useless, but whether its cutoff is a faithful support representation for the coarse state transition.

On aligned probes of a four-step Wan2.2-T2V-A14B student, the top-$b$ pool covers 51.91% of teacher-window attention mass. Persistent interactions are selected more often, including within overlapping teacher-importance strata, yet some persistent teacher-high blocks are still missed. The complement combines these omissions with a diffuse never-top tail. Value-sketch diagnostics motivate retaining the useful ranked core while allocating part of the budget beyond its cutoff.

We test this allocation principle through the final teacher endpoint, allowing later student steps to compensate for local changes. Holding the per-step keep schedule fixed, an outside quota of 16 reduces Teacher40 RMSE from 0.959537 to 0.928340, with improvement on all six reported records (Section 5.2). Smaller and larger tested quotas improve the mean less. Original multi-seed controls provide complementary evidence under their own configuration (Appendix F.1).

We instantiate the resulting support-completion policy as SPARE Attention (Figure 3). SPARE uses coarse query blocks for cheap routing, forms a current-score candidate pool, and spends a small complement quota outside that pool. The ranked pool preserves local relevance; the complement quota is intended to cover contributions outside the cutoff at low cost. Previous support is carried only as selection state for quota bookkeeping. The keep schedule is selected on development data. The policy requires neither teacher access nor an additional reference forward pass.

The original system configuration is evaluated separately on the full 6,220-video VBench suite. At 14.84% exact block density, SPARE has the shortest denoising time among the timed sparse implementations and a Total comparable to the SLA family and PISA. Teacher-endpoint agreement and VBench Total can favor different budgets; we therefore report them as distinct outcomes.

Our contributions are a block-level diagnosis of window-to-step support mismatch, a teacher-free support-completion policy, and matched-budget endpoint interventions paired with a system quality–cost evaluation. The intervention establishes an allocation-policy effect, without identifying a unique causal block class.

# 2 RELATED WORK

**Sparse structure and routing.** Sparse video attention exploits structure at several levels. Sliding Tile Attention uses spatial–temporal locality to construct hardware-efficient windows (Zhang et al., 2025d). Sparse VideoGen2 groups tokens through semantic-aware permutation (Yang et al., 2025), while XAttention uses antidiagonal samples to estimate block importance (Xu et al., 2025). These approaches improve the structure or estimation of sparse support. SPARE examines the boundary that selection draws: a few-step student's current ranking leaves out a diffuse share of the teacher's attention mass, motivating a quota outside its highest-ranked candidates.

**Approximating unselected interactions.** SLA combines sparse and linear attention and supports fine-tuning (Zhang et al., 2025a). PISA computes critical blocks exactly and approximates the remainder (Li et al., 2026b). Sol-Attn integrates threshold-based routing, sparse computation, and approximation correction within an online-softmax pass (Li et al., 2026a). Thus, sparse attention is not uniformly a keep-or-drop operation. Our study concerns how the selected-block budget is

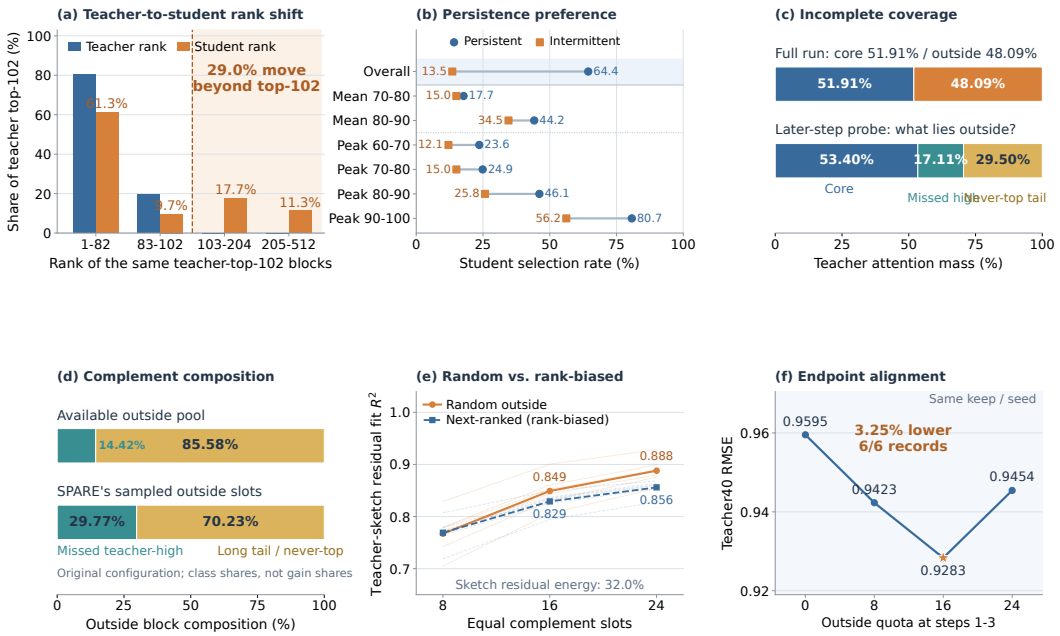



Figure 2: **A persistence-skewed core leaves a mixed complement.** (a) Rank transfer of the same teacher-top-102 blocks; the teacher reference follows bin sizes. Top-102 is the original rank-audit threshold, not deployed keep. (b) Student selection of persistent ($f_t \geq 0.8$) and intermittent ($0 < f_t \leq 0.2$) blocks, overall and within teacher mean/peak importance percentile strata. (c) Full-run paired teacher-mass coverage is 51.91%; 48.09% remains outside. The separate later-step probe splits mass into core 53.40%, missed teacher-high 17.11%, and never-top 29.50%. (d) Original outside-pool and sampled-slot compositions. Teacher-high ($f_t > 0$) includes persistent 17.38%, intermittent 2.95%, and intermediate 9.44% of sampled outside slots; never-top has $f_t = 0$. SPARE samples within state-constrained groups, so its composition differs from the whole pool. (e) At equal slots, random outside blocks fit the teacher-sketch residual better than next-ranked blocks at 16 and 24; faint lines are records. (f) A separate locked same-keep/seed intervention reduces final Teacher40 RMSE from 0.959537 to 0.928340 (3.25%). Panel (e) is a local fit comparison, not a direct endpoint contrast; (d) reports composition, not class-specific endpoint gains.

allocated; the experiments compare specified implementations and operating points under the same student inference setup.

Recent work also explores orthogonal ways to obtain or organize sparse support. SpargeAttention and VSA use training-free or trainable selection mechanisms, respectively (Zhang et al., 2025b;c); Radial Attention uses an energy-decay structure for long videos (Li et al., 2025); and AdaSPA adapts sparse decisions across the denoising process (Xia et al., 2025). These methods are structural points of comparison; our numerical comparison is restricted to released implementations that run under the same checkpoint, resolution, evaluator, and timing protocol.

**Cross-step support and few-step models.** LiteAttention propagates decisions about non-essential tiles using temporal coherence (Shmilovich et al., 2025), and HEART studies head-dependent sparse-attention behavior (Zheng et al., 2026). SPARE recomputes its candidate pool at every step and carries previous support only as selection state. We study this policy in the Wan family (Wan et al., 2025) and evaluate final videos with VBench (Huang et al., 2024). The teacher supplies diagnostic references; it is absent from SPARE inference.

# 3 WINDOW-TO-STEP SUPPORT MISMATCH

Figure 2 connects ranking mismatch, omitted teacher-window contribution, and a matched-budget endpoint intervention. We first describe the local support diagnosis, then test its allocation consequence after complete student generation (Section 5.2).

## 3.1 ALIGNED TEACHER–STUDENT PROBES

For each student step $t$, we associate a window $W_t$ of 10 teacher steps, partitioning the 40-step teacher trajectory into 4 disjoint windows. The probe uses 6 probe records, 6 layers, eight heads, and 16 sampled query rows per layer and step. Each row has 512 key blocks, and the diagnostic operates on these stored block-level probes.

Let $p_\tau(k)$ denote a teacher block's softmax probability and $T_\tau$ the teacher's top-102 block set at teacher step $\tau$. The within-window saliency frequency is

$$f_t(k) = \frac{1}{|W_t|} \sum_{\tau \in W_t} \mathbf{1}\{k \in T_\tau\}. \tag{1}$$

The student's current-step ranking induces a top-$b_t$ pool $G_t$ and its complement $O_t$, where $b_t$ is the per-step keep budget of the evaluated schedule (Section 4). We write $G$ and $O$ when the step is clear. The pool is a student-local relevance representation, while the teacher window supplies an aligned multi-step diagnostic reference; the latter is not treated as the student's stepwise ground truth. Teacher-high blocks are those with $f_t(k) > 0$; never-teacher-top blocks have $f_t(k) = 0$.

For each stored query row we form a teacher-window sketch $y$ by pooling the teacher's value-weighted attention output over $W_t$, and let $c(G)$ denote the contribution of the blocks in $G$ to that sketch. The residual $y - c(G)$ is what the student's pool leaves unexplained. To compare complement subsets we report the $R^2$ of an origin-constrained fit of this residual onto the subset's summed contribution, and collinearity as the mean absolute cosine between each subset block and the summed contribution of $G$. These are aggregate-direction diagnostics, not projections onto the full core span. Appendix A.6 gives the constructions. Quantities below are means over the 6 probe records unless noted.

The original sigma-aligned rank audit directly compares block identities (Figure 2a). Full-ranking correlation is 0.759, yet 29.0% of the teacher's top-102 blocks fall outside the student's top 102: 17.7% move to ranks 103–204 and 11.3% to ranks 205–512. With both sets fixed at 102 blocks, only 71.0% of either set is shared. Thus an informative student ranking still yields a different sparse support. This rank audit is distinct from the following ten-step-window probes; its top-102 cutoff is used to compare equal-size sets.

## 3.2 A USEFUL CORE WITH A PERSISTENCE PREFERENCE

We distinguish persistent blocks ($f_t \geq 0.8$), intermittent blocks ($0 < f_t \leq 0.2$), intermediate blocks, and never-top blocks ($f_t = 0$). In the additional persistence summaries, persistent blocks have median student rank 12.8% and selection rate 64.4%, versus 29.0% and 13.5% for intermittent blocks. This is not evidence that the student core is unimportant: the first 5% of student ranks captures 35.2% of teacher-window mass. Indeed, the earlier rank diagnosis finds that 98.5% of teacher-top-5% blocks missed by the student top 5% are persistent; only 1.9% of its intermittent group reaches the teacher's top 10% even at its peak. Mismatch thus includes missed persistent blocks, not just rare high-importance events.

Importance explains part of the selection gap. The supplied coarse importance strata nevertheless show higher persistent selection in both usable mean-importance strata and all four usable peak-importance strata (Figure 2b; Table 4). The result describes a persistence association within comparable importance ranges; it is not exact or joint matching on mean and peak, or an isolation of distillation as its cause.

### 3.3 VALUE-WEIGHTED CONTRIBUTION OUTSIDE THE SUPPORT

To summarize all four student steps, let $\bar{p}_t(k) = |W_t|^{-1} \sum_{\tau \in W_t} p_\tau(k)$ and define full-run average paired coverage as

$$C_{\text{pair}}(G) = \frac{1}{4} \sum_{t=0}^{3} \sum_{k} \bar{p}_t(k) \mathbf{1}\{k \in G_t\}. \tag{2}$$

The four coverages are 47.43, 49.05, 53.89, and 57.26%, averaging 51.91% and leaving 48.09% outside (Figure 2c). The windows partition all 40 teacher steps, but coverage retains each student's paired support mask; it is an aligned diagnostic, not a stepwise imitation objective.

The original last-three-step summary covers 53.4% of teacher mass, with 46.6% in the complement. The mean per-query relative residual energy is 32.0%; the additional probe's pooled-energy ratio is 25.1%, a different aggregation. These offline teacher-sketch residuals measure a contribution omitted by the student's support boundary, not the student's actual output error. Their endpoint consequence requires the separate intervention below.

### 3.4 WHAT THE COMPLEMENT CONTAINS

The complement contains both missed salient blocks and a diffuse tail. Never-teacher-top blocks hold 29.5% of total teacher mass, versus 17.1% on missed teacher-high blocks (Figure 2c). Admitted outside slots include 29.77% teacher-high and 70.23% never-top blocks (panel d). These composition measurements describe the original selection context, not the locked masks or either class's causal endpoint effect. Appendix A.7 retains the full pool/slot breakdown and distinguishes gross recovery from net selection.

### 3.5 EXTENDING THE RANKING VERSUS SAMPLING THE COMPLEMENT

A natural comparison is to take the next-ranked student blocks rather than sample the complement. This compares rank-biased recovery with random outside recovery at equal slot counts. In the teacher-sketch residual they explain, the two choices are tied at eight slots (0.769 versus 0.767) and separate as the quota grows: at sixteen and twenty-four slots the random complement reaches 0.849 and 0.888, against 0.829 and 0.856 for the next-ranked blocks (overall 0.842 versus 0.826, random higher on 4 of 6 records; Figure 2e). These fits concern a single summed predictor and do not identify residual rank or demonstrate new directions outside the core span. They also do not constitute a direct random-versus-next-ranked final-endpoint comparison. They motivate testing a complement allocation, rather than requiring any local residual to decrease at every step. Its final effect is measured by the completed student trajectory against Teacher40 in Section 5.2, with the original policy controls retained in Appendix F.1.

## 4 SPARE ATTENTION: SUPPORT COMPLETION

The diagnosis treats the student's current ranking as a local relevance representation of a coarse transition, not as a complete representation of the teacher-window support. SPARE implements partial support completion under a fixed budget: it keeps a current-score candidate pool and spends a small quota outside it. The design principle is this allocation between ranked relevance and support completion. Previous support is carried only as selection state that distinguishes retained from fresh candidates when the quotas are filled; it is an implementation choice whose endpoint advantage is not established by the original policy controls (Appendix F.1).

### 4.1 COARSE BLOCK ROUTING

For each self-attention head, we partition queries into blocks of 128 tokens and keys into blocks of 64 tokens. At the evaluated token length there are $M = 512$ key blocks. The routing score for query block $i$ and key block $j$ is

$$s_t(i, j) = \bar{q}_{t,i}^{\mathsf{T}} \bar{k}_{t,j}, \tag{3}$$

where the implementation mean-pools queries and mean-centered keys. Common positive scaling does not affect the ranking. For each query block we form the candidate pool $G_t = \text{Top}_{b_t}(s_t)$ and its

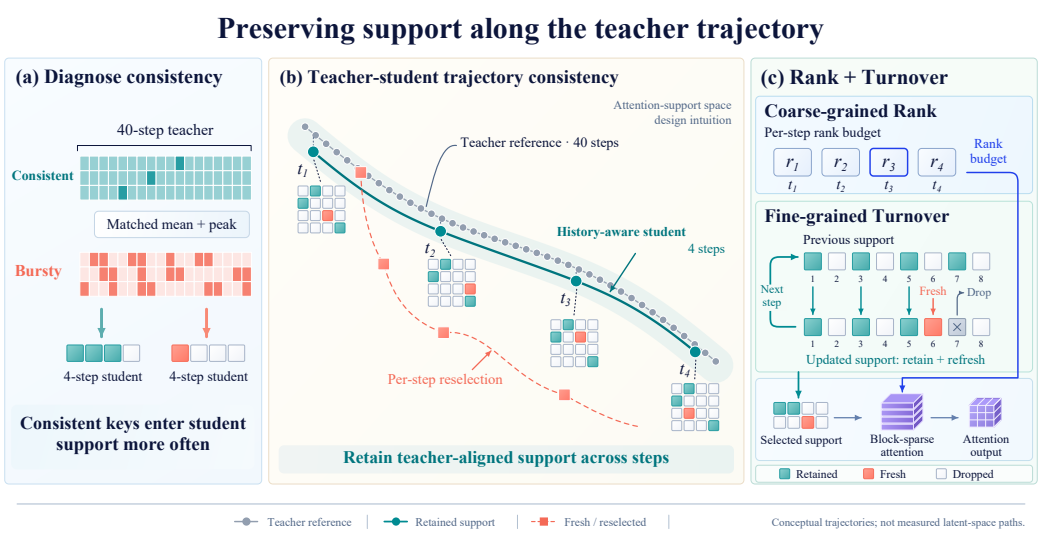



Figure 3: **SPARE method overview.** The original schematic connects ranking mismatch to rank-plus-turnover selection. Trajectories are conceptual, not measured paths. The mean/peak comparison denotes separate coarse importance strata, not exact joint matching; previous support is selection state rather than an independently established source of endpoint improvement.

complement $O_t$. Coarsening reduces the number of query blocks and amortizes execution. We claim the measured kernel saving in Appendix F.2, not equivalent support quality at the two granularities.

### 4.2 Candidate pool, complement quota, and support state

At each step the selected support contains $n_t$ blocks from $G_t$ and $o_t = b_t - n_t$ blocks from $O_t$. The previous support is $P = S_{t-1}$. For a category $X \in \{G_t, O_t\}$ with quota $q_X$, the feasible number of retained blocks is bounded by

$$\ell_X = \max(0, q_X - |X \setminus P|), \qquad h_X = \min(q_X, |X \cap P|). \tag{4}$$

The target $r_t$ is clipped to $[\ell_G + \ell_O, h_G + h_O]$. Retention above the feasible minimum is allocated first within $G_t$ and then within $O_t$. We sample uniformly without replacement inside each of the four resulting category-and-state groups. This is quota-based sampling from a ranked candidate pool; it does not guarantee that the globally highest $n_t$ blocks are always selected.

Algorithm 1 in Appendix A.5 gives the complete selection procedure.

Step zero is pure top-$b_0$; later steps select exactly $b_t$ blocks and retain the clipped target. Appendix A.3 records the schedule representation and execution details.

**Evaluated schedules.** The original system configuration uses

$$b^{\text{sys}} = [72, 68, 82, 82], \quad n^{\text{sys}} = [72, 54, 66, 66], \quad o^{\text{sys}} = [0, 14, 16, 16]. \tag{5}$$

The locked endpoint experiment instead uses

$$b^\star = [78, 66, 80, 80], \quad n^\star = [78, 50, 64, 64], \quad o^\star = [0, 16, 16, 16]. \tag{6}$$

Both schedules average 76 selected blocks, but their per-step allocations differ. The locked comparison changes only outside allocation within $b^\star$; the full-suite quality and latency results remain tied to the original system configuration. Neither schedule is claimed to be an analytical optimum. Teacher persistence labels are used offline for diagnosis and exploratory controls, not by the online selection rule.

**Execution.** Selected indices feed the block-sparse kernel in the same student forward pass, without a teacher or dense reference prepass. Layer/branch state resets for each record, and denoising latency includes all selection overhead.

# 5 EXPERIMENTS

## 5.1 EXPERIMENTAL SETUP

All video comparisons use the same Wan2.2-T2V-A14B four-step checkpoint, with dequantized BF16 weights, at 480×832 resolution and 81 frames. Self-attention is varied; cross-attention remains dense. Student generation uses RTX 4090 hardware. Dense denotes the four-step student with dense self-attention. SLA80 uses the evaluated LightX2V block-sparse backend at a nominal sparsity of $0.8$; this path has no linear-remainder branch and no additional SLA fine-tuning is performed for this comparison. The same backend is also reported at the additional nominal operating points SLA85, SLA90, and SLA95. Sol-Attn uses its official backend with threshold $\tau = 1$. The dense 40-step teacher is an additional quality reference, generated on A800 hardware in a different runtime, and is excluded from student latency comparisons.

The full VBench manifest has 6,220 video records and 943 unique prompts. Each dimension uses its official subset and aggregation. Generation uses the fixed seed 2026082613. Full-suite results are point estimates from one seed. Schedule selection and timing use a 20-record development cohort; the 6-record probe cohort of Section 3 is used for the endpoint interventions (Appendix A.8).

## 5.2 CLOSING THE LOOP: SAME-KEEP TEACHER-ENDPOINT ALIGNMENT

The diagnosis identifies a useful but persistence-skewed core and a heterogeneous complement. We test its allocation consequence by completing the four-step student generation and measuring the final latent against the matched 40-step teacher:

$$\mathcal{E}(\pi) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{RMSE}(x_{0,i}^{\pi}, x_{0,i}^{T}). \tag{7}$$

This teacher-alignment objective allows all downstream compensation; it is neither a sum of per-step attention residuals nor a distance to the dense student endpoint.

The new comparison holds $b^{\star}$ fixed and uses the same generation seed across its four policies. Step zero remains pure top-$b_0$; the outside quota at each later step is 0, 8, 16, or 24. All policies have mean keep 76. Figure 2f shows a reduction from 0.959537 to 0.928340 at quota 16, an absolute reduction of 0.031196 and relative reduction of 3.25%. The supplied locked report records improvement on all six records for the zero-versus-16 contrast. The quota response is non-monotone: 8 and 24 improve the mean less than 16. The corresponding means at quotas 8 and 24 are 0.942313 and 0.945419, respectively.

The result supports a controllable allocation effect on the completed trajectory. It does not identify a unique causal block class or require monotonic local residual reduction. The locked report supplies aggregate means and the all-six improvement count, but no per-record seed arrays; uncertainty from the original controls is not transferred to this setting.

## 5.3 SELECTION-POLICY CONTROLS

The original matched experiment fixes $b^{\mathrm{sys}}$, geometry, and generation noise on six records with three mask seeds. Pure top-$b_t$ is worst on all six records; SPARE minus pure top-$b_t$ is -0.0173 (95% record-bootstrap interval [-0.0225, -0.0105]). A zero-retention-target control does not establish an advantage for particular retained identities, supporting our treatment of history as selection state. The full policies, per-record plot, and quota sweep appear in Appendix F.1. These original-configuration intervals are not transferred to the locked endpoint comparison.

## 5.4 FULL-SUITE QUALITY AND DENOISING LATENCY

This system evaluation retains $b^{\mathrm{sys}}$, the original reported operating point. The locked endpoint result is additional mechanism evidence, not a replacement measurement of its video quality or speed.

Table 1 places SPARE among the sparse configurations. Among the timed implementations (SLA80, Sol-Attn, PISA-HYD), SPARE computes the fewest key blocks exactly (14.84%) and has the shortest

Table 1: **Video quality and denoising cost of one Wan2.2-T2V-A14B four-step student across three suites.** VBench columns are the official Total, Quality, and Semantic scores on the 0–100 scale. VBench-2.0 columns are the five leaderboard groups and their unweighted mean (Total), on the raw $[0, 1]$ scale: Human Fidelity (anatomy, identity, clothes), Creativity (diversity, composition), Controllability (dynamic spatial relationship, dynamic attribute, motion order, human interaction, complex landscape, complex plot, camera motion), Physics (mechanics, thermotics, material, multi-view consistency), and Commonsense (motion rationality, instance preservation). Each group score is the unweighted mean of its raw dimensions. CompBench Total is the unweighted mean of the seven scaled category scores, also on $[0, 1]$. Per-dimension values are in Tables 8 and 9 (Appendix D). These two suites were run on four arms; — marks a suite or a timing protocol that was not run. Density is the fraction of key blocks computed exactly per query block, averaged over steps (SLA and PISA rows: realized block count under the shared block rule, e.g. 76/512 at SLA85; Sol-Attn uses a data-dependent threshold). Time is the uncontended single-process denoising mean of Section 5.4; Speedup is relative to the dense student. SLA85/90/95 were not timed. Every entry is rounded to two decimal places. The shaded row is the proposed method.

| Method | VBench↑ | | | VBench-2.0↑ | | | | | | CompBench↑ | Efficiency | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Total | Quality | Semantic | Human | Creat. | Ctrl. | Phys. | Comm. | Total | Total | Dens. (%) | Time (s)↓ | Speedup↑ |
| Dense student | 81.29 | 83.39 | 72.88 | 0.88 | 0.48 | 0.37 | 0.58 | 0.65 | 0.59 | 0.59 | 100.00 | 46.78 | 1.00× |
| SLA80 | 82.13 | 83.77 | 75.53 | 0.85 | 0.47 | 0.38 | 0.58 | 0.66 | 0.59 | 0.58 | 19.92 | 32.58 | 1.44× |
| Sol-Attn | 80.37 | 83.10 | 69.42 | 0.80 | 0.51 | 0.31 | 0.56 | 0.64 | 0.56 | 0.53 | *dynamic* | 29.73 | 1.57× |
| PISA-HYD | 81.74 | 83.48 | 74.77 | — | — | — | — | — | — | — | 19.92 | 33.90 | 1.38× |
| **SPARE** | 82.34 | 83.93 | 76.01 | 0.83 | 0.44 | 0.37 | 0.50 | 0.65 | 0.56 | 0.58 | 14.84 | 29.57 | 1.58× |

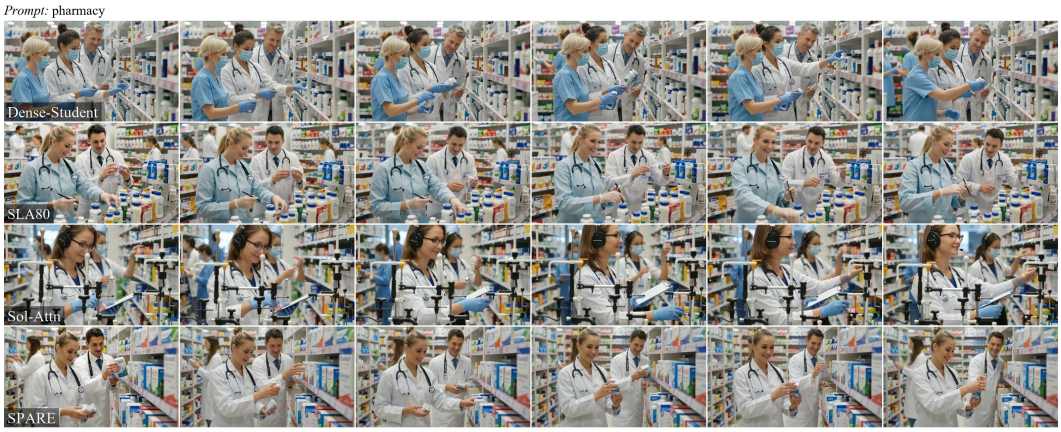



Figure 4: **Wan2.2 examples from VBench.** Unscored frames; see Table 1 for scores.

four-step denoising time (29.565 s). On quality, the sparse configurations occupy a narrow band: SPARE's Total of 82.34 sits with SLA80 (82.13), the untimed SLA85/90/95 points (82.35, 82.61, 82.38, at equal or lower density), and PISA (81.74), and above the dense student (81.29). The evidence for the complement quota is therefore the endpoint interventions and the cost side of this table, not a Total advantage over the SLA family. The dimension breakdown in Appendix C shows where the aggregate sits and which dimensions favor the baselines. A fixed-seed qualitative panel in Appendix A.13 provides a visual sanity check on matched prompts; it is illustrative and is not used to replace the aggregate evaluation.

**Qualitative comparison.** Figure 4 shows a Wan2.2 generation example from the VBench evaluation. The frame panel complements the aggregate scores; it is an unscored comparison rather than an additional quality metric.

**Denoising speed.** The timing protocol uses one GPU and one process without competing workload, and measures 20 development records, excluding the first warm-up record from the mean. The resulting 19-record mean is 29.565 seconds for SPARE, 32.579 for SLA80, and 46.784 for dense

attention. This corresponds to 9.25% less denoising time than SLA80 and a $1.582\times$ speedup over dense attention. The measurement includes support selection and the full denoising loop, but excludes decoding and video publication; it is not end-to-end generation latency.

Further kernel timing and interpretation of the teacher quality reference are given in Appendix F.2.

**Further controls.** The budget sweep and granularity checks are in Appendices B and F.2. Endpoint RMSE and VBench Total need not favor the same budget; the kernel-time saving does not establish equivalent support.

**Additional evaluations.** Table 1 also reports VBench-2.0 and T2V-CompBench summaries. Their per-dimension results and further qualitative examples are retained in Appendices D and G.

### 5.5 Image editing transfer

We also evaluate the four attention arms on Qwen-Image-Edit with four-step GEdit-Bench generation over 606 items. Table 2 reports the local Qwen2.5-VL judge mean: SPARE scores 8.139, the highest of the four arms. Its configuration uses keep $0.75$, outside $0.5$, and retain $1$. This is a transfer check under a local judge, not an official VIE-Score result or a measurement of the locked video configuration.

Table 2: **Qwen-Image-Edit on GEdit-Bench.** Four-step generation on 606 items, scored by a local Qwen2.5-VL judge. Higher is better; these are not official VIE-Score results.

| Method | Dense | SLA80 | Sol-Attn | **SPARE** |
|---|---|---|---|---|
| GEdit local mean↑ | 8.051 | 8.079 | 8.111 | **8.139** |

## 6 Limitations

The mechanism and endpoint studies use six development-cohort records; coarse importance-stratified persistence differences are associations, not an identified effect of distillation. The locked comparison provides aggregate means and reported improvement counts, without raw per-record seed arrays for uncertainty estimation. Original multi-seed controls remain tied to their own configuration. Neither evidence identifies a unique causal block class, an optimal schedule, or an advantage of particular retained identities.

The full-suite comparison uses one generation seed, one Wan2.2 checkpoint, and one resolution. Endpoint RMSE and VBench Total can rank budgets differently. Untimed SLA85/90/95 points reach equal or lower density with Total at or above SPARE; the measured latency claim applies only to the timed implementations. The small gap to Sol-Attn needs repeated interleaved timing for a robust ordering. Coarse query blocks save kernel time, but matched-mask checks do not establish equivalent support quality. Additional suite results are mixed (Appendix G); the locked variant's perceptual quality, latency, and cross-family video generalization remain unmeasured. The image-editing transfer check uses a local judge rather than the official VIE-Score pipeline.

## 7 Conclusion

A few-step student's current ranking supplies a useful core but leaves teacher-window support and value-weighted contribution outside its sparse cutoff. SPARE implements partial support completion by allocating a fixed budget between ranked candidates and their complement, without teacher access at inference. A matched per-step-budget intervention reduces final Teacher40 RMSE by 3.25%, connecting the diagnosis to a controllable endpoint effect after complete generation. Separately, the original system configuration combines competitive VBench Total with the shortest measured denoising time among the timed sparse implementations. The resulting design principle is to assess sparse support through the endpoint consequences of its omissions, while evaluating perceptual quality and execution cost independently.

## REFERENCES

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Haopeng Li, Yitong Li, Junsong Chen, Tian Ye, Haozhe Liu, Jincheng Yu, Duomin Wang, Ruihua Zhang, Zeke Xie, Enze Xie, and Song Han. Sol-Attn: Accelerating video generation inference via on-the-fly attention sparsification. *arXiv preprint arXiv:2607.24027*, 2026a.

Haopeng Li, Shitong Shao, Wenliang Zhong, Zikai Zhou, Lichen Bai, Hui Xiong, and Zeke Xie. PISA: Piecewise sparse attention is wiser for efficient diffusion transformers. *arXiv preprint arXiv:2602.01077*, 2026b.

Xingyang Li, Muyang Li, Tianle Cai, Haocheng Xi, Shuo Yang, Yujun Lin, Lvmin Zhang, Songlin Yang, Jinbo Hu, Kelly Peng, Maneesh Agrawala, Ion Stoica, Kurt Keutzer, and Song Han. Radial attention: $O(n \log n)$ sparse attention with energy decay for long video generation. *arXiv preprint arXiv:2506.19852*, 2025.

Dor Shmilovich, Tony Wu, Aviad Dahan, and Yuval Domb. LiteAttention: A temporal sparse attention for diffusion transformers. *arXiv preprint arXiv:2511.11062*, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Yifei Xia, Suhan Ling, Fangcheng Fu, Yujie Wang, Huixia Li, Xuefeng Xiao, and Bin Cui. Training-free and adaptive sparse attention for efficient long video generation. *arXiv preprint arXiv:2502.21079*, 2025.

Ruyi Xu, Guangxuan Xiao, Haofeng Huang, Junxian Guo, and Song Han. XAttention: Block sparse attention with antidiagonal scoring. *arXiv preprint arXiv:2503.16428*, 2025.

Shuo Yang, Haocheng Xi, Yilong Zhao, Muyang Li, Jintao Zhang, Han Cai, Yujun Lin, Xiuyu Li, Chenfeng Xu, Jianfei Chen, Song Han, Kurt Keutzer, and Ion Stoica. Sparse VideoGen2: Accelerate video generation with sparse attention via semantic-aware permutation. *arXiv preprint arXiv:2505.18875*, 2025.

Jintao Zhang, Haoxu Wang, Kai Jiang, Shuo Yang, Kaiwen Zheng, Haocheng Xi, Ziteng Wang, Hongzhou Zhu, Min Zhao, Ion Stoica, Joseph E. Gonzalez, Jun Zhu, and Jianfei Chen. SLA: Beyond sparsity in diffusion transformers via fine-tunable sparse-linear attention. *arXiv preprint arXiv:2509.24006*, 2025a.

Jintao Zhang, Chendong Xiang, Haofeng Huang, Jia Wei, Haocheng Xi, Jun Zhu, and Jianfei Chen. SpargeAttention: Accurate and training-free sparse attention accelerating any model inference. *arXiv preprint arXiv:2502.18137*, 2025b.

Peiyuan Zhang, Yongqi Chen, Haofeng Huang, Will Lin, Zhengzhong Liu, Ion Stoica, Eric Xing, and Hao Zhang. VSA: Faster video diffusion with trainable sparse attention. *arXiv preprint arXiv:2505.13389*, 2025c.

Peiyuan Zhang, Yongqi Chen, Runlong Su, Hangliang Ding, Ion Stoica, Zhengzhong Liu, and Hao Zhang. Fast video generation with sliding tile attention. *arXiv preprint arXiv:2502.04507*, 2025d.

Xuzhe Zheng, Yuexiao Ma, Jing Xu, Xiawu Zheng, Rongrong Ji, and Fei Chao. HEART: Exploiting head heterogeneity in sparse attention for video diffusion. *arXiv preprint arXiv:2605.14513*, 2026.

# A Reproducibility and measurement details

## A.1 Model and implementation

The evaluated student uses Wan2.2-T2V-A14B with a high-noise and a low-noise expert, with the expert boundary at student step index two. The recorded denoising step list is $[1000, 750, 500, 250]$. Weights are loaded in BF16 from a dequantized NVFP4-derived checkpoint. The configuration disables classifier-free guidance, uses sampling shift five, and retains dense cross-attention. The diagnosis in Section 3 describes the evaluated checkpoint and does not attribute the effect to a specific distillation objective.

The student runtime is PyTorch 2.10.0 with CUDA 12.8 and Triton 3.6.0 on RTX 4090 (SM89). The SPARE kernel uses eight warps and two stages. Generation workers use one visible GPU each. Timing comes from separate uncontended single-process measurements, not concurrent full-suite generation journals. The first development record used for warm-up is excluded from the timing mean; all development records contribute to endpoint RMSE. VAE decoding and video writing are excluded from denoising latency. Dense teacher full-suite videos come from an earlier A800 runtime; matched endpoint ablations instead use hash-bound teacher references at the corresponding endpoint runtime.

SLA80 denotes the existing LightX2V sparse backend at nominal sparsity $0.8$, not a new fine-tuned reproduction of every setting in the SLA paper; its evaluated branch is pooled-score top-$k$ block-sparse attention without a linear-remainder path. Sol-Attn uses the official Sol engine checkout, and PISA20 uses the official `piecewise_sparse_attention_hyd` backend with block size 64 and density $0.20$. Source commits and configuration hashes are retained in the machine-readable provenance. These are specific deployment comparisons; hardware-specific retuning could change their ordering.

## A.2 Exact policy semantics

The implementation forms $G_t$ from top-$b_t$, not top-$n_t$. When $n_t < b_t$, only a subset of $G_t$ is selected. For each category, Eq. (4) ensures that both its retained and fresh quotas are feasible. Because $r \geq \ell_G + \ell_O$ and $r \leq h_G + h_O$, assigning retained blocks within $G_t$ first leaves a feasible residual in $O_t$. Uniform random priorities implement the draws without replacement. Thus the algorithm enforces both the category quotas and the clipped total overlap with the previous support. The target is total retained support, not an outside-only retained count.

The variant called `turnover_b` in archived files sets $r_t = 0$ before feasibility clipping. It still needs the previous support to distinguish retained and fresh candidates. We label it "zero retention target" throughout the paper. It should not be interpreted as independent uniform sampling from the whole complement. The teacher-informed shuffled control is diagnostic only; it cannot serve as a deployable baseline or a latency comparison.

## A.3 Schedule representation

The evaluated schedules are represented by the step-indexed quantities

$$(b_t, n_t, o_t, r_t)_{t=0}^{T-1}, \qquad n_t + o_t = b_t, \qquad r_t \in [\ell_G + \ell_O, h_G + h_O]. \tag{8}$$

Numeric schedules are recorded with the configuration; figures use these symbols so that the mechanism is portable across few-step students. Step zero is pure top-$b_0$ because there is no previous support. The later steps retain exactly the clipped target and select exactly $b_t$ blocks. The mean exact-support density is 14.84%. The schedule is selected on development data. Section 5.4 reports its operating point together with a sweep of the total budget.

## A.4 Execution and state

The selected indices feed the existing block-sparse attention kernel. Routing, mask construction, retention bookkeeping, and sparse attention occur within the same student forward pass. The selection stages precede sparse execution without a dense reference prepass or teacher evaluation. Previous-support state is maintained separately by layer and branch and reset for each record. Random priorities

are derived from a fixed seed and call identifiers. All selection overhead is included in denoising latency. Coarse geometry and reduced density are both changed relative to SLA80, so the main comparison evaluates this configured sparse branch rather than a separately fine-tuned sparse-linear model.

### A.5 SELECTION PSEUDOCODE

**Algorithm 1** SPARE support selection for one head and query block

**Require:** scores $s_t$, previous support $P$, quotas $(b_t, n_t, o_t)$, target $r_t$
$G \leftarrow \mathrm{Top}_{b_t}(s_t)$; $O \leftarrow \{1, \dots, M\} \setminus G$
**if** $t = 0$ **then**
  **return** $G$
**end if**
Compute $(\ell_G, h_G)$ and $(\ell_O, h_O)$ using Eq. (4)
$r \leftarrow \mathrm{clip}(r_t, \ell_G + \ell_O, h_G + h_O)$
$r_G \leftarrow \ell_G + \min(r - \ell_G - \ell_O, h_G - \ell_G)$; $r_O \leftarrow r - r_G$
$S \leftarrow \emptyset$
**for** $(X, q, r_X) \in \{(G, n_t, r_G), (O, o_t, r_O)\}$ **do**
  Add $r_X$ uniform draws from $X \cap P$ to $S$
  Add $q - r_X$ uniform draws from $X \setminus P$ to $S$
**end for**
**return** $S$

### A.6 TEACHER-WINDOW SKETCHES AND RESIDUAL FITS

The coverage, composition, and residual-fit numbers in Section 3 are computed on the stored block-level probes. For a query row at student step $t$, let $y$ be a sketch of the teacher's value-weighted attention output pooled over the aligned window $W_t$, and let $c(G)$ be the contribution of the student's top-$b_t$ pool $G$ to that sketch. Coverage is the share of the teacher's window-pooled softmax mass on $G$ versus $O$; residual energy is the energy of $y - c(G)$ relative to that of $y$. Composition splits the teacher mass on $O$ into never-teacher-top blocks ($f_t = 0$) and missed teacher-high blocks ($f_t > 0$), both expressed as shares of the total teacher mass. When comparing a complement subset $S \subset O$ of a given size, we form $c(S) = \sum_{k \in S} c_k$ and fit $y - c(G)$ onto that single aggregate vector through the origin. The fit has one scalar coefficient, not one free coefficient per block. Collinearity is the mean absolute cosine between $c_k$ and $c(G)$, the summed core contribution, rather than the full span of core blocks. Neither statistic establishes residual rank or orthogonality to the core span. The next-ranked comparison takes the highest-scoring student blocks outside $G$; the random comparison samples blocks of $O$ uniformly; the deployed set is the complement set realized by the evaluated later-step quota. Fits are averaged over rows and then over the 6 probe records; the per-record counts quoted in Section 3 compare record means.

### A.7 COMPLEMENT COMPOSITION DETAILS

The omitted mass is not concentrated on a few missed high-attention keys. Blocks that never enter the teacher's top set ($f_t = 0$) hold 29.5% of the teacher's attention mass, compared with 17.1% on missed teacher-high blocks (Figure 2c); the ordering holds on all 6 of the 6 records. The complement therefore includes a diffuse tail as well as missed salient blocks. The original outside pool contains 14.42% teacher-high and 85.58% never-top blocks by count. In the outside-selection report, admitted outside slots comprise 29.77% missed teacher-high blocks and 70.23% long-tail (never-top) blocks. The teacher-high share includes 17.38% persistent, 2.95% intermittent, and 9.44% intermediate blocks, each expressed as a share of all admitted outside slots. SPARE's complement budget thus covers both kinds of omission. These are composition measurements for the original selection context, not measurements of the locked endpoint configuration (Figure 2d). SPARE samples uniformly within its category-and-state groups (Section 4), rather than independently from the entire outside pool; the pool and selected-slot shares therefore need not coincide. They establish a heterogeneous complement; these shares do not measure either class's contribution to the endpoint improvement in Figure 2f. Gross recovery and net selection must also be distinguished: intermittent selection remains

essentially unchanged in the new pooled summary, and the older report's smaller persistence gap is mainly associated with reduced persistent selection (Appendix A.11).

### A.8 Data usage and statistical units

The mechanism cohort contains 6 records; development selection and timing use 20 records; the full VBench suite contains 6,220 records from 943 unique prompts. The mechanism and development records are drawn from the same prompt manifest as the full suite, so the full-suite scores are point estimates rather than an independent confirmation set. T2V-CompBench uses 200 videos in each of seven categories for the same four arms. All seven categories are scored for every arm.

For the small-cohort mechanism and endpoint intervals, resampling occurs at the record level. In the three-seed endpoint experiment, seeds are averaged within each record before comparing policies; attention cells and mask seeds are not treated as additional independent prompts.

### A.9 Complement-quota experiment

The one-seed quota experiment keeps $b_t = [72, 68, 82, 82]$ fixed and uses the zero-retention-target rule; the pure-top-$b$ and standard-quota settings are shared with the A–D experiment. The setting labelled "16" uses the actual later-step schedule $(14, 16, 16)$ rather than an equal count at all three steps.

Table 3: Complement-quota interventions on the same 6 probe records and one mask seed. The keep budget is identical across rows: mean density 14.84%, or sparsity 85.16%.

| Setting | Outside schedule | Teacher-endpoint RMSE |
|---|---|---|
| Pure native | $[0, 0, 0, 0]$ | 0.9741 |
| Small outside | $[0, 8, 8, 8]$ | 0.9483 |
| Standard outside | $[0, 14, 16, 16]$ | 0.9518 |
| Larger outside | $[0, 24, 24, 24]$ | 0.9497 |

The paired small-minus-standard difference is -0.0035 with interval [-0.0128, 0.0063]; standard-minus-larger is 0.0021 [-0.0163, 0.0156]; small-minus-larger is -0.0014 [-0.0244, 0.0152]. The overlapping uncertainty does not establish equality among the nonzero quotas, and we do not retune the schedule to the smallest observed mean.

### A.10 Exact values behind the unified observation figure

Figure 2 combines the diagnostic and endpoint observations. Table 4 retains the detailed importance strata. The matched endpoint means are described in Section 5.2 and plotted in Figure 2f.

Table 4: **Persistence and student selection within importance strata.** Only strata with sufficient overlap between both classes are shown. Rates are pooled block-observation proportions, not independent-record confidence intervals.

| Importance | Percentile bin | Persistent (%) | Intermittent (%) | Gap (pp) |
|---|---|---|---|---|
| Mean | 70–80% | 17.73 | 15.02 | +2.71 |
| Mean | 80–90% | 44.18 | 34.55 | +9.64 |
| Peak | 60–70% | 23.61 | 12.07 | +11.54 |
| Peak | 70–80% | 24.89 | 15.04 | +9.85 |
| Peak | 80–90% | 46.12 | 25.76 | +20.37 |
| Peak | 90–100% | 80.72 | 56.19 | +24.53 |

### A.11 Additional persistence and locked-configuration evidence

The new evidence extends the original manuscript without pooling its experimental protocols. The full-run paired coverage in Eq. (2) uses the original four-step keep schedule; the geometry figure uses

its original last-three-step sketches. Persistence statistics summarize sampled block observations. The locked endpoint report uses $b^\star$ and a separately reported fixed generation seed. No new record-level bootstrap is computed from those aggregate means, and the six records are not a held-out multi-seed confirmation.

**Keep timing and protocol identity.** The newly supplied original-keep summary reports RMSE 0.974141, 0.973098, 0.949364, and 0.951519 for later-step quotas 0, 8, 16, and 24. The locked zero-outside mean is instead 0.959537. Thus the larger cross-schedule improvement from 0.974141 to 0.928340 combines keep timing and complement allocation. Only 0.959537 versus 0.928340 isolates allocation within $b^\star$. The original zero-retention-target quota study in Appendix A.9 has its own state rule and uses $[0, 14, 16, 16]$ for its standard quota. Its record-level results are preserved; similarly named arms from the new summaries are not substituted for or pooled with that experiment.

**Composition is distinct from net recovery.** In the older outside-set report, the intermittent selection rate changes from 16.76% to 16.42%, while the persistent rate changes from 51.09% to 43.42%. The narrower gap does not demonstrate increased net intermittent selection. In the new unconditional persistence CSV, intermittent selection is 13.5446% before and 13.5457% after the reported policy change, effectively unchanged. Outside slots can recover omitted intermittent blocks while others leave the inside pool. Gross recovery, net selection, and gap reduction are different quantities. These statistics motivate heterogeneous support completion without establishing a unique causal mediator.

**Geometry does not imply new span.** In the additional 32-dimensional value-sketch probe, the core matrix rank averages about 31.55. A lower cosine to its sum therefore does not imply new directions outside its span. The active-step temporal cosine is 0.875 for persistent blocks and 0.923 for intermittent blocks; persistence alone does not establish greater temporal directional repetition. The arm labeled *mass matched* has mean mass about 0.01823 versus 0.03616 for next-ranked blocks, so it is not used as a successful equal-mass control. The value-sketch evidence remains descriptive; final teacher alignment supplies the task-level intervention outcome.

### A.12 Granularity and empirical allocation

The fine-versus-coarse check uses 24 captured attention calls with matched lookup masks. The coarse kernel repeats exactly in this check, whereas splitting the same mask across two fine blocks produces attention-output RMSE in $[4.95, 22.12] \times 10^{-6}$. All captures exceed the predeclared threshold $\max(2\epsilon_{\text{repeat}}, 10^{-6})$. This prevents attributing the endpoint difference solely to routing geometry. Separately, the stored query rows are nonadjacent, so they cannot directly measure the support disagreement between the adjacent halves of a deployed coarse query block.

We do not claim that the keep schedule is an analytical optimum; it is selected on development data. The later budget sweep below is reported in full for the same reason.

### A.13 Qualitative matched frames

Figure 5 shows midpoint frames from three fixed prompts at the same manifest seed index across the completed Dense, SLA80, SPARE, and Sol-Attn video runs. The panel is a visual sanity check rather than a selection claim: it is not scored, and SPARE is not expected to dominate every row. The aggregate VBench table remains the quantitative comparison.

## B Additional quality–cost curves

**Budget curves and metric disagreement.** We evaluate all 16 VBench dimensions at seven additional budgets. Figure 6 plots all settings, including the reused main point. The smallest tested budget, mean keep 48, gives the highest observed Total, 82.60, at 27.832 seconds, while its teacher-endpoint error is higher than at the main setting. Thus, endpoint agreement and video-quality aggregation need not rank budgets the same way. The main setting remains the fixed reported configuration; the complete exploratory sweep exposes the sensitivity of this choice. Table 5 gives all values, followed by the quality–latency plot.

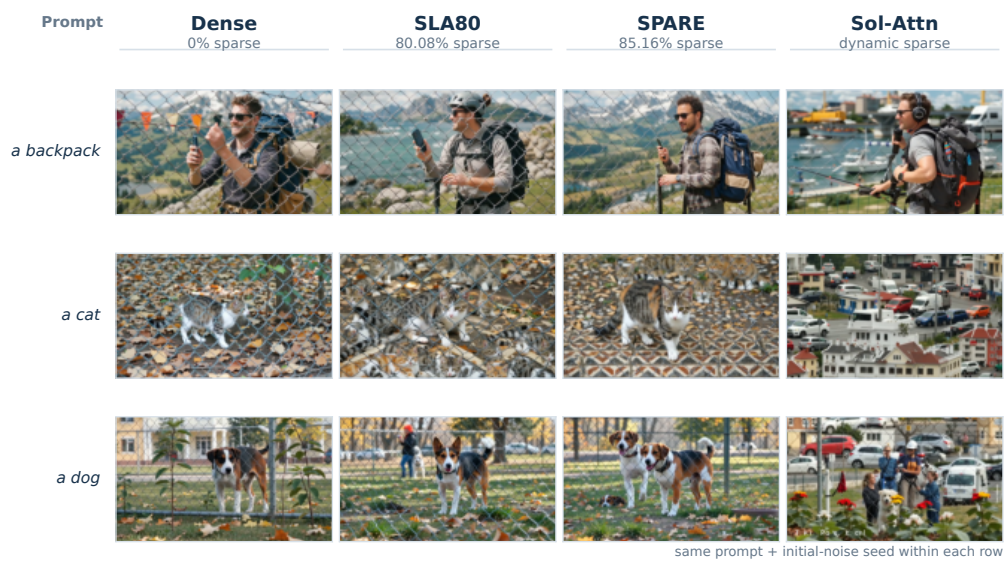



Figure 5: **Matched qualitative frames.** Midpoint frames for three fixed prompts at the same manifest seed index across Dense, SLA80, SPARE, and Sol-Attn. This unscored panel is a visual sanity check; aggregate quality is reported in Table 1.

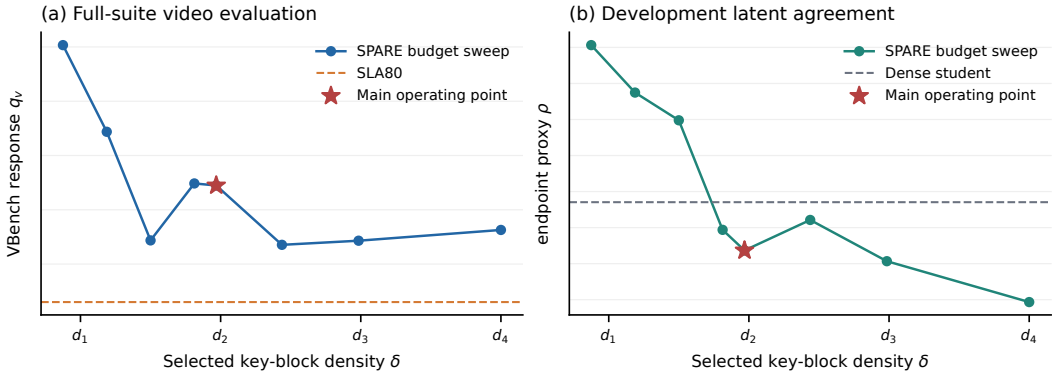



Figure 6: **Quality changes with sparse budget.** (a) Official full-suite VBench Total. (b) Teacher-endpoint RMSE on development records. Every marker is a completed setting; lines only connect measured points. The main setting is marked with a star. The different trends show that improved teacher-endpoint agreement is not a guarantee of improved VBench Total.

Table 5 reports every completed budget setting. The schedules scale the original shape and its within-pool/outside quotas using the same generation seed. The main point is reused, not regenerated as an independent replicate. Full-suite VBench, development teacher-endpoint RMSE, and uncontended denoising time are separate measurements with their own cohorts. The full-suite scores are single-seed point estimates.

## C Full VBench dimension results

Table 6 gives raw dimension aggregates, before official normalization. Quality and Semantic are computed with the official dimension normalization and weights; their aggregate Total uses the official group weights. We do not replace these with an unweighted average of raw scores. SPARE is above SLA80 on 11 of 16 dimensions; the remaining dimensions favor a baseline. Dimension-specific

Table 5: Complete exploratory budget sweep. Mean keep is the number of key blocks selected per query averaged over the four steps. The original method uses mean keep 76; all settings have complete 16-dimension VBench results.

| Mean keep | Density (%) | Sparsity (%) | Total | Quality | Semantic | Time (s) | RMSE |
|---|---|---|---|---|---|---|---|
| 48 | 9.38 | 90.62 | 82.60 | 84.04 | 76.87 | 27.832 | 0.968128 |
| 56 | 10.94 | 89.06 | 82.44 | 83.94 | 76.46 | 28.180 | 0.965495 |
| 64 | 12.50 | 87.50 | 82.24 | 83.74 | 76.27 | 28.586 | 0.963956 |
| 72 | 14.06 | 85.94 | 82.35 | 83.82 | 76.45 | 28.962 | 0.957876 |
| 76 | 14.84 | 85.16 | 82.34 | 83.93 | 76.01 | 29.565 | 0.956733 |
| 88 | 17.19 | 82.81 | 82.24 | 83.81 | 75.94 | 29.700 | 0.958425 |
| 102 | 19.92 | 80.08 | 82.24 | 83.92 | 75.54 | 30.370 | 0.956135 |
| 128 | 25.00 | 75.00 | 82.26 | 83.92 | 75.65 | 31.563 | 0.953867 |

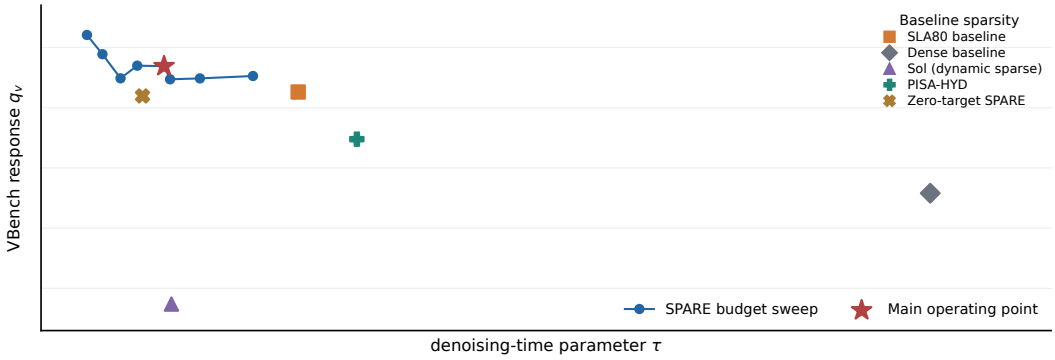



Figure 7: **Measured quality–latency points.** The curve connects all completed SPARE budget settings. Baseline points use the configurations in Table 1. The main setting is marked separately from the later exploratory sweep. These points do not establish a global Pareto frontier or statistical dominance.

sample counts are reported in Table 7. The color metric has different returned entry counts across methods, so its aggregate is not a paired mean over an identical set of videos. Completeness of generation and availability of all dimensions should not be confused with identical evaluator coverage.

Table 6: Raw VBench dimension aggregates for the original comparison and two completed companion arms. "Zero target" is the SPARE variant with preferred retention set to zero, subject to feasibility. Method-specific densities are listed in Table 1.

| Dimension | SPARE | SLA80 | Dense | Sol | Teacher | PISA20 | Zero target |
|---|---|---|---|---|---|---|---|
| Subject consistency | 0.9571 | 0.9539 | 0.9571 | 0.9590 | 0.9464 | 0.9615 | 0.9564 |
| Background consistency | 0.9539 | 0.9530 | 0.9576 | 0.9449 | 0.9456 | 0.9599 | 0.9550 |
| Temporal flickering | 0.9883 | 0.9882 | 0.9873 | 0.9853 | 0.9850 | 0.9876 | 0.9886 |
| Motion smoothness | 0.9835 | 0.9836 | 0.9850 | 0.9861 | 0.9788 | 0.9862 | 0.9843 |
| Dynamic degree | 0.5694 | 0.5500 | 0.4917 | 0.5361 | 0.7722 | 0.4611 | 0.5472 |
| Aesthetic quality | 0.6601 | 0.6633 | 0.6613 | 0.6485 | 0.6298 | 0.6656 | 0.6582 |
| Imaging quality | 0.7029 | 0.7043 | 0.6982 | 0.6866 | 0.6999 | 0.7015 | 0.7030 |
| Object class | 0.8856 | 0.8831 | 0.8608 | 0.7918 | 0.8367 | 0.8878 | 0.8797 |
| Multiple objects | 0.8646 | 0.8555 | 0.7918 | 0.6834 | 0.6838 | 0.8370 | 0.8570 |
| Human action | 0.8360 | 0.8380 | 0.8320 | 0.8000 | 0.8400 | 0.8260 | 0.8400 |
| Color | 0.8552 | 0.8516 | 0.8546 | 0.8745 | 0.8634 | 0.8720 | 0.8462 |
| Spatial relationship | 0.8851 | 0.8818 | 0.8070 | 0.8162 | 0.7879 | 0.8633 | 0.8755 |
| Scene | 0.3869 | 0.3667 | 0.3244 | 0.2797 | 0.3525 | 0.3517 | 0.3641 |
| Appearance style | 0.1936 | 0.1957 | 0.1961 | 0.1956 | 0.2032 | 0.1974 | 0.1925 |
| Temporal style | 0.2416 | 0.2388 | 0.2350 | 0.2205 | 0.2342 | 0.2313 | 0.2405 |
| Overall consistency | 0.2560 | 0.2553 | 0.2502 | 0.2373 | 0.2469 | 0.2511 | 0.2551 |

Table 7: Number of per-video entries returned by each official dimension evaluator. These counts describe evaluator outputs rather than independent prompts.

| Dimension | SPARE | SLA80 | Dense | Sol | Teacher | PISA20 | Zero target |
|---|---|---|---|---|---|---|---|
| Subject consistency | 360 | 360 | 360 | 360 | 360 | 360 | 360 |
| Background consistency | 430 | 430 | 430 | 430 | 430 | 430 | 430 |
| Temporal flickering | 375 | 375 | 375 | 375 | 375 | 375 | 375 |
| Motion smoothness | 360 | 360 | 360 | 360 | 360 | 360 | 360 |
| Dynamic degree | 360 | 360 | 360 | 360 | 360 | 360 | 360 |
| Aesthetic quality | 465 | 465 | 465 | 465 | 465 | 465 | 465 |
| Imaging quality | 465 | 465 | 465 | 465 | 465 | 465 | 465 |
| Object class | 395 | 395 | 395 | 395 | 395 | 395 | 395 |
| Multiple objects | 410 | 410 | 410 | 410 | 410 | 410 | 410 |
| Human action | 500 | 500 | 500 | 500 | 500 | 500 | 500 |
| Color | 397 | 389 | 354 | 340 | 352 | 376 | 393 |
| Spatial relationship | 420 | 420 | 420 | 420 | 420 | 420 | 420 |
| Scene | 430 | 430 | 430 | 430 | 430 | 430 | 430 |
| Appearance style | 450 | 450 | 450 | 450 | 450 | 450 | 450 |
| Temporal style | 500 | 500 | 500 | 500 | 500 | 500 | 500 |
| Overall consistency | 465 | 465 | 465 | 465 | 465 | 465 | 465 |

## D VBENCH-2.0 AND T2V-COMPBENCH DIMENSION RESULTS

Table 8 lists the 18 raw VBench-2.0 dimension scores behind the five group scores in Table 1, and Table 9 lists the seven T2V-CompBench category scores behind Table 1's CompBench total.

Table 8: **VBench-2.0 raw dimension scores.** Four student arms on 18 dimensions. The unweighted mean of these raw scores is not the official VBench-2.0 aggregate. Table 1 instead reports the five leaderboard group scores and their unweighted mean.

| Dimension | Dense | SLA80 | Sol-Attn | SPARE |
|---|---|---|---|---|
| Camera Motion | 0.290 | 0.290 | 0.231 | **0.296** |
| Complex Landscape | 0.173 | 0.171 | 0.151 | **0.187** |
| Complex Plot | 0.116 | 0.131 | 0.105 | **0.134** |
| Composition | **0.497** | 0.485 | 0.443 | 0.480 |
| Diversity | 0.465 | 0.446 | **0.571** | 0.392 |
| Dynamic Attribute | **0.381** | 0.311 | 0.242 | 0.293 |
| Dynamic Spatial Relationship | 0.493 | **0.570** | 0.440 | 0.536 |
| Human Anatomy | 0.958 | **0.961** | 0.953 | 0.953 |
| Human Clothes | **0.932** | 0.867 | 0.780 | 0.856 |
| Human Identity | **0.735** | 0.721 | 0.675 | 0.688 |
| Human Interaction | 0.810 | **0.837** | 0.697 | 0.797 |
| Instance Preservation | **0.918** | 0.901 | 0.895 | 0.906 |
| Material | **0.617** | 0.583 | 0.589 | 0.553 |
| Mechanics | 0.578 | 0.613 | **0.615** | 0.514 |
| Motion Order Understanding | 0.303 | 0.320 | 0.293 | **0.323** |
| Motion Rationality | 0.391 | **0.420** | 0.385 | 0.397 |
| Multi-View Consistency | **0.471** | 0.456 | 0.398 | 0.349 |
| Thermotics | 0.659 | **0.672** | 0.655 | 0.597 |
| *Mean of 18 dimensions* | **0.544** | 0.542 | 0.507 | 0.514 |

Table 9: **T2V-CompBench category scores.** Each entry is the mean over 200 videos, divided by the category's official scale. The last row is the unweighted mean of the seven scaled category scores.

| Category | Dense | SLA80 | Sol-Attn | SPARE |
|---|---|---|---|---|
| Consistent attribute | 0.850 | 0.855 | 0.814 | **0.858** |
| Dynamic attribute | **0.128** | 0.101 | 0.041 | 0.082 |
| Action binding | 0.819 | 0.813 | **0.824** | 0.818 |
| Object interactions | 0.822 | 0.828 | 0.775 | **0.838** |
| Spatial relationships | 0.661 | 0.683 | 0.660 | **0.690** |
| Generative numeracy | **0.529** | 0.443 | 0.348 | 0.460 |
| Motion binding | 0.301 | 0.335 | 0.272 | **0.345** |
| *Mean of 7 categories* | **0.587** | 0.580 | 0.533 | 0.584 |

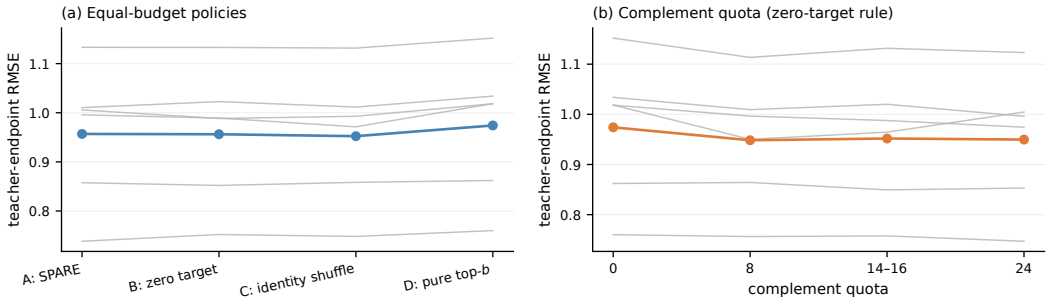



Figure 8: **Equal-budget interventions on the probe records.** The original system keep schedule is fixed throughout this figure. (a) Teacher-endpoint RMSE of the four policies; thin lines are individual records, the bold line is the cohort mean. (b) Teacher-endpoint RMSE as the later-step complement quota varies under the zero-target rule. Lower is closer to the teacher endpoint; the metric measures trajectory proximity, not perceptual quality.

# E ARTIFACT PROVENANCE

All numerical macros, tables, and result figures in this manuscript are generated by scripts from the completed experiment artifacts. The source bundle records source paths, transformations, and SHA256 hashes for every reported number and figure; regeneration reads existing artifacts and launches no model inference. Full-suite scores come from the official merged 16-dimension result files; endpoint and mechanism numbers come from the stored probe analyses; the budget curves come from the full-suite sweep together with its development and timing analyses.

# F EXTENDED POLICY AND SYSTEM CONTROLS

## F.1 ORIGINAL EQUAL-BUDGET POLICY CONTROLS

The original study compares four selection policies on the 6 probe records of Section 3 with three mask seeds, holding $b^{\mathrm{sys}} = [72, 68, 82, 82]$, block geometry, and generation noise fixed. This is a separate matched experiment from Figure 2f; its policy means and uncertainty estimates remain attached to the original configuration. The outcome is the RMSE between the student's final latent and the matched teacher endpoint. Policy A is SPARE. Policy B sets the retention target to zero, so the complement quota is refilled from fresh blocks. Policy C shuffles which previous-support blocks are retained while preserving the quotas; it is a diagnostic, non-deployable control. Policy D spends the entire budget on the current top-$b_t$ blocks.

Pure top-$b_t$ selection (D) is worst on all 6 records; the paired A-minus-D contrast is -0.0173 (95% record-bootstrap interval [-0.0225, -0.0105]). The tested nonzero-quota policies improve on the all-inside mean in this original protocol. Holding the keep schedule fixed and varying the later-step complement quota under the zero-target rule gives mean RMSE 0.9741 at quota zero versus 0.9483, 0.9518, and 0.9497 at the three nonzero settings (Figure 8b; Appendix A.9). The zero quota is worst on 5 of 6 records, and the pairwise intervals among the nonzero quotas contain zero.

The original controls do not establish an advantage for retaining the particular previous-support identities: A minus B is 0.0006 (95% interval [-0.0080, 0.0089]), with policy means 0.9569, 0.9562, and 0.9523 for A, B, and C. This does not establish equivalence of all complement identities, or apply those means to the locked configuration. SPARE keeps previous support as selection state that separates retained from fresh candidates. The exact policy semantics are given in Appendix A.

## F.2 ADDITIONAL EFFICIENCY AND TEACHER-REFERENCE INTERPRETATION

SPARE selects 25.5% fewer key blocks per query than SLA80. A separate operator measurement gives 28.296 ms with query blocks of 128 tokens and 47.083 ms with blocks of 64 tokens. This

supports an execution benefit from coarse blocks, but does not isolate their contribution to the whole-system speedup. Sol-Attn takes 29.730 seconds under the same protocol; the 0.56% difference is too small to support a robust speed advantage without repeated interleaved measurements. PISA-HYD takes 33.897 seconds on the same official backend.

**Retention-target comparison.** The zero-retention-target variant scores 82.10 at 29.080 seconds with unchanged geometry and quotas; together with the A–B endpoint contrast, this is why retention is treated as state rather than as a quality mechanism.

**Interpreting the teacher reference.** The teacher has the highest Quality-group score among the comparison arms, although SPARE has a higher aggregate Total. Dynamic degree is 0.7722 for the teacher, compared with 0.5694 for SPARE and 0.4917 for the dense student. Thus, a higher Total does not imply better motion. Because the lower dynamic-degree score also appears in the dense student, it cannot be attributed solely to sparse selection. These comparisons do not isolate the effects of distillation, checkpoint differences, or runtime.

**Query-block granularity.** A comparison of fine and coarse policies gives a mean fine-minus-coarse endpoint difference of -0.0009 (95% interval [-0.0136, 0.0143]). However, a matched-mask kernel check fails the predeclared numerical-identifiability threshold on all 24 of 24 captures. The comparison cannot separate routing changes from kernel numerics, and the interval containing zero does not prove support equivalence. We therefore claim measured execution savings, not equal support quality between the two block sizes.

## G Additional evaluations and qualitative examples

**Qualitative panels.** The supplied generation panels and the image examples in Figure 1 are illustrative, unscored comparisons; they do not establish cross-family quantitative generalization.

*Prompt:* A whale and a seagull flying side by side, the whale soaring through the clouds while the seagull glides effortlessly.




Figure 9: **Wan2.2 examples from VBench-2.0.** Unscored frames; see Table 8 for scores.

*Prompt:* Four horses gallop across the beach.




Figure 10: **Compositional generation examples.** Unscored frames; see Table 9 for scores.

*Prompt:* A British Shorthair cat wearing a yellow helmet rides through a night market; the helmet reads SPARE, with a small wind-driven three-blade rotor; smooth side tracking shot, natural motion and night lights.

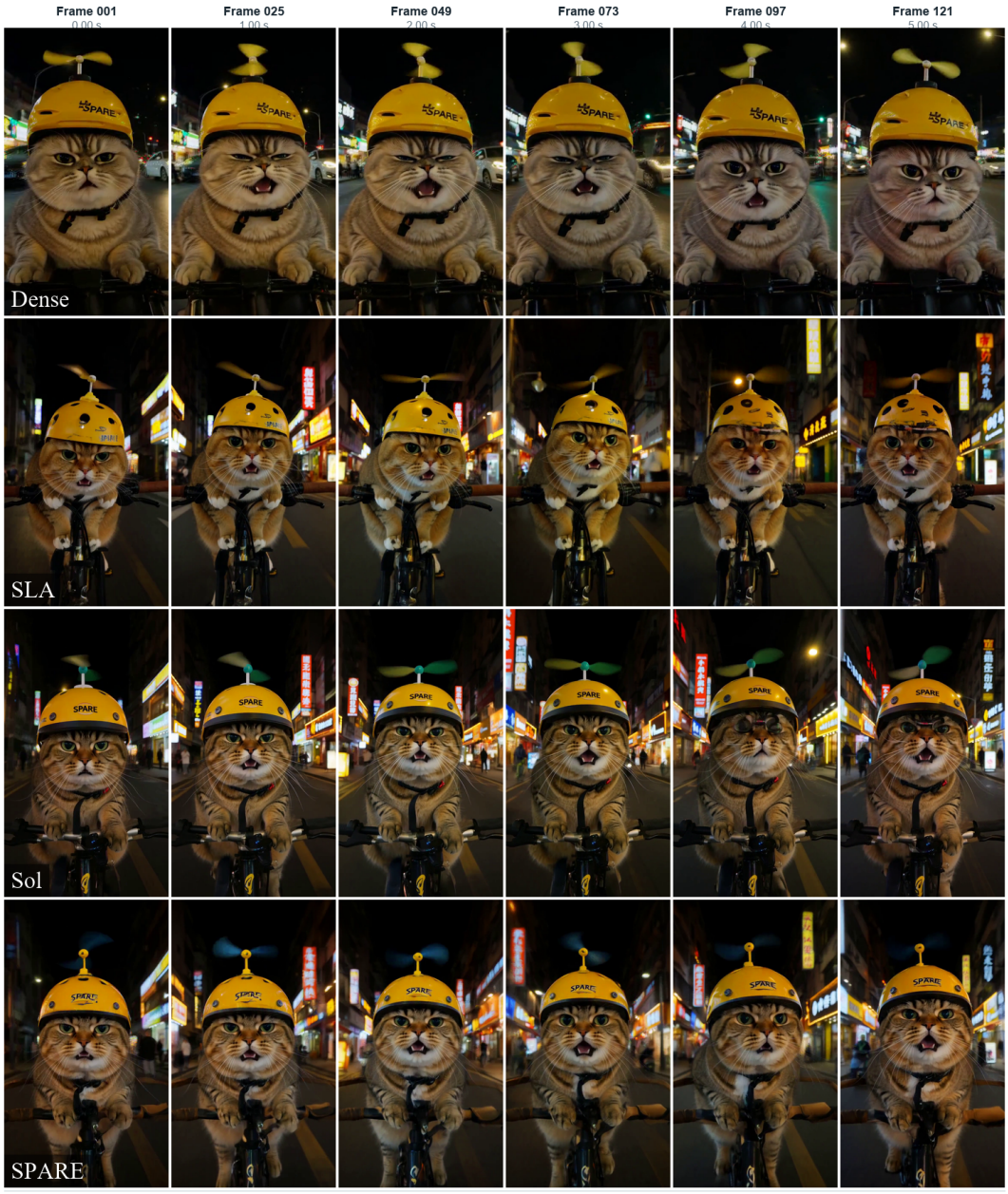



Figure 11: **MiniMax-H3 qualitative sequence.** The supplied generation frames are illustrative and are not an additional controlled endpoint experiment.