# ASYMREFRESH: MODALITY-ASYMMETRIC REFRESH FOR EFFICIENT JOINT AUDIO-VIDEO GENERATION

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

## ABSTRACT

Feature caching accelerates joint audio–video generation, but joint refreshes tie additional audio computation to video recomputation. We introduce AsymRefresh, a training-free method that makes modality choice part of the refresh decision. Between full joint refreshes, selected steps recompute the native audio pathway against reconstructed video contexts while predicting the video output. A shallow joint prefix and periodically renewed contexts preserve audio access to the evolving video trajectory, with unchanged model weights and samplers. We evaluate JavisDiT++ on all 10,140 JavisBench prompts and both JavisDiT++ and JavisDiT on all 500 T2AV-Compass prompts, covering joint-attention and dual-stream backbones. On JavisBench, AsymRefresh achieves a $1.93\times$ speedup in a separate timing study with competitive full-benchmark quality against four caching baselines. On JavisDiT, the selected T2AV operating point leads the compared accelerators on four of six primary quality means. Controlled JavisDiT++ studies on 512 prompts and three seeds show improved audio–text alignment over additional joint refreshes at comparable latency under two budgets, with positive multiplicity-adjusted intervals. The six-update gain is also observed at a second prefix depth. Stage and context controls characterize the alignment benefit and accompanying quality and memory trade-offs. These results support modality selection as a useful dimension of inference-budget allocation in joint generation.

## 1 INTRODUCTION

Joint audio–video generators synthesize visual events and their sounds through a coupled sampling process. Models such as JavisDiT, JavisDiT++, and Ovi exchange information between the two modalities within transformer blocks (Liu et al., 2026a;b; Low et al., 2025). Repeatedly evaluating these blocks is expensive. Feature caching addresses this cost by reusing or predicting intermediate features while retaining the pretrained model (Ma et al., 2024b; Liu et al., 2025a;b; Bu et al., 2026).

For a coupled block stack, a joint refresh treats video and audio as one unit of computation: allocating an additional refresh to audio also pays for a video update. Cross-modal attention offers a finer choice. Audio queries can read video keys and values without recomputing video outputs. This raises a budget-allocation question: can selective audio updates improve audio–text alignment more than additional joint refreshes at similar latency, while preserving visual conditioning throughout generation?

Development diagnostics provide an empirical starting point. Replacing the audio output with a fully computed output in late sampling steps improved audio–text alignment; the middle-stage intervention did not show the same benefit (Appendix H). Obtaining these diagnostic outputs required extra full computation. An efficient method must instead supply each audio block with useful video context while avoiding video-output updates, even as both modalities continue to evolve.

We address this challenge with **AsymRefresh**, a training-free method for *modality-asymmetric refresh*. A shallow joint prefix observes the current coupled state at every step. On selected steps, the remaining audio blocks execute against video contexts reconstructed from the two latest full computations, while the video output is predicted. Periodic joint refreshes update both streams and renew the stored contexts. AsymRefresh makes the modality of recomputation an explicit budget choice while preserving audio's dependence on the evolving video trajectory. We use the DiCache residual



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.

predictor (Bu et al., 2026) for output and context reconstruction, and keep it fixed in the joint-only controls to evaluate the effect of computation allocation.

The evaluation provides evidence across benchmarks and backbone structures. JavisDiT++ covers the full JavisBench (Liu et al., 2026a) and T2AV-Compass (Cao et al., 2026) benchmarks; the original JavisDiT is also evaluated on T2AV. On JavisBench, AsymRefresh reduces generation time from 34.70 to 17.95 seconds in the separate timing study, a $1.93\times$ speedup. T2AV tests both a 50-step joint-attention backbone and a 30-step backbone with separate modality streams and bidirectional cross-attention. At the selected JavisDiT operating point, AsymRefresh has higher mean video technical quality and all three cross-modal alignment scores than each of the four caching baselines.

Controlled JavisDiT++ studies then examine the allocation itself. On 512 prompts and three seeds, selective audio updates improve audio–text alignment over extra joint refreshes at comparable latency under two budgets. Stage, context, and prefix controls characterize where this benefit arises and its quality and memory costs.

Our contributions are:

- We introduce AsymRefresh, a training-free policy that decouples audio recomputation from video-output recomputation within joint sampling. Reconstructed intermediate video context supports native audio computation while video outputs are predicted and both modalities evolve.
- We evaluate JavisDiT++ on JavisBench and T2AV-Compass, and JavisDiT on T2AV-Compass, against four public caching baselines. The evaluations cover joint-attention and separate-stream architectures under their native generation settings.
- Paired JavisDiT++ allocation studies hold the predictor fixed and show audio–text alignment gains over extra joint refreshes at the tested latency budgets. Stage, context, and prefix controls characterize the gain and associated quality and memory trade-offs.

## 2 Related Work

**Joint and conditioned audiovisual generation.** MM-Diffusion, JavisDiT, JavisDiT++, UniVerse-1, and Ovi couple audio and video generation (Ruan et al., 2023; Liu et al., 2026a;b; Wang et al., 2025; Low et al., 2025). Latent aligners also couple separately pretrained generators (Xing et al., 2024). Video-to-audio systems such as FoleyCrafter, MMAudio, and ReWaS condition on an available video (Zhang et al., 2026; Cheng et al., 2025; Jeong et al., 2025); audio-driven video methods instead take sound as input (Yariv et al., 2024; Jeong et al., 2023). AsymRefresh operates while both modalities are being generated, so reused video context must remain useful along an evolving joint trajectory.

**Caching, prediction, and scheduling.** Fast solvers reduce sampling evaluations (Lu et al., 2022); caching reduces the computation within a retained sampling trajectory. Feature reuse has been studied at the level of U-Net features, transformer layers, and individual tokens (Ma et al., 2024b; Selvaraju et al., 2024; Ma et al., 2024a; Zou et al., 2025). Jenga instead reduces token interactions through dynamic attention carving and progressive resolution (Zhang et al., 2025). PAB broadcasts attention outputs, FasterCache also caches classifier-free guidance outputs, and AdaCache, TeaCache, MagCache, and EasyCache adapt reuse decisions (Zhao et al., 2025; Lyu et al., 2025; Kahatapitiya et al., 2025; Liu et al., 2025a; Ma et al., 2025; Zhou et al., 2025). TaylorSeer and FoCa forecast features (Liu et al., 2025b; Zheng et al., 2026). DiCache uses shallow-layer trajectories to predict deep-layer features, while RFC estimates output changes from module inputs (Bu et al., 2026; Son et al., 2026). AsymRefresh adds modality selection to the execution of coupled blocks. It builds on DiCache's predictor, which is shared with our joint-only allocation controls.

**Modality-asymmetric audiovisual computation.** SyncCache reuses visual computation while recomputing audio blocks for portrait animation conditioned on supplied audio (Ma et al., 2026). The H3 AudioRefine implementation computes audio queries against cached video context during an audio refinement pass with the video latent held fixed (Adudeguyman, 2026). AsymRefresh places selective updates inside joint sampling: video latents keep evolving, joint refreshes renew the context, and the audio budget competes with additional joint computation. Synchrony-aware

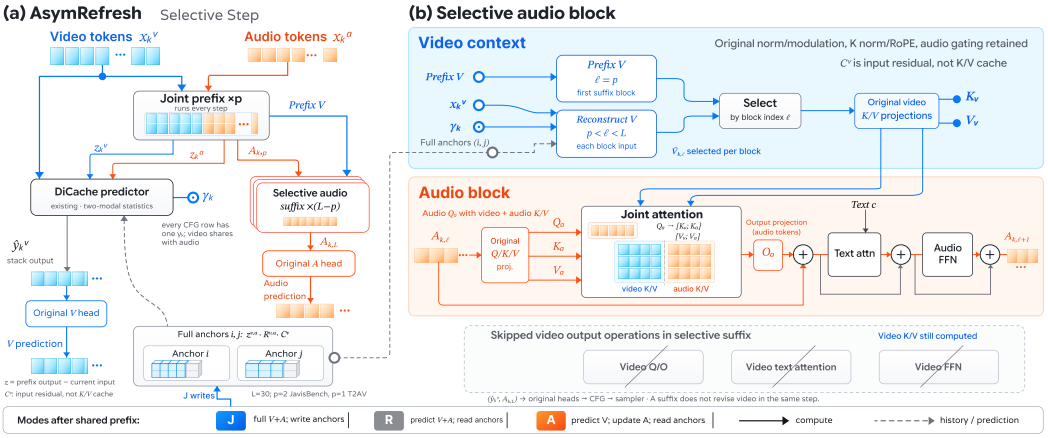



Figure 1: AsymRefresh's selective audio path on JavisDiT++. (a) A joint prefix runs at every sampling step. Selective steps predict video stack outputs with DiCache and recompute the original audio suffix; only joint refreshes update full anchors. (b) Audio queries attend to joint video–audio keys and values. The first suffix block uses current prefix video states; later blocks reconstruct video inputs from full anchors. Video K/V are recomputed; video Q/O projections, video text attention, and video FFNs are skipped. The stack has $L = 30$ blocks; its main-table operating points use prefix depth $p = 2$ for JavisBench and $p = 1$ for T2AV.

acceleration uses cross-modal saliency to guide sparse attention and cache reuse (Gao et al., 2026). AsymRefresh retains the backbone's native audio attention to video context and changes which modality outputs are recomputed.

## 3 AsymRefresh

### 3.1 From joint refresh to selective computation

Joint refresh couples the cost of updating audio to that of updating video. AsymRefresh introduces a finer choice: recompute audio outputs between full joint refreshes, and compare this allocation with extra joint computation at similar measured latency.

The key distinction is between *updating video outputs* and *providing video context to audio*. Audio queries need video keys and values but can produce outputs without computing video attention outputs. The late-stage diagnostics in Appendix H motivate using this option for audio.

Predicting the final video output alone cannot supply the intermediate video states required by deeper audio blocks. A shallow joint prefix gives the first selective block current context; full-computation histories reconstruct deeper video inputs. Joint refreshes renew these histories as both modalities evolve (Figure 1).

### 3.2 Execution policy

Let $k \in \{0, \ldots, N-1\}$ index sampling steps and $x_k = (x_k^v, x_k^a)$ enter the $L$-block controlled stack $F_k$. For JavisDiT++, $x_k$ follows input projection; the JavisDiT boundary is specified in Appendix A. The prefix $P_k$ comprises the first $p$ joint blocks and runs every step. Disjoint schedules $\mathcal{J}$ and $\mathcal{A}$ select joint and audio refreshes:

$$y_k = \begin{cases} F_k(x_k), & k \in \mathcal{J} \quad \text{(joint refresh)}, \\ (\widehat{y}_k^v, A_{k,L}), & k \in \mathcal{A} \quad \text{(selective audio)}, \\ (\widehat{y}_k^v, \widehat{y}_k^a), & \text{otherwise} \quad \text{(reuse)}. \end{cases} \tag{1}$$

Here $\widehat{y}_k^m$ predicts modality $m$'s stack output; $A_{k,L}$ is recomputed by the native audio suffix. Original output heads, classifier-free guidance (CFG), and the sampler consume $y_k$. Only joint refreshes

create *full anchors*, storing fully computed states; invalid prediction history also triggers a joint refresh.

### 3.3 AUDIO COMPUTATION WITH REUSED VIDEO CONTEXT

In a JavisDiT++ joint block, the audio attention output is

$$H_a = \mathrm{softmax}\left(\frac{Q_a[K_v; K_a]^\top}{\sqrt{d}}\right)[V_v; V_a], \tag{2}$$

where $[\,;\,]$ concatenates tokens and $d$ is the per-head dimension. The native audio path retains query/key normalization, rotary encoding, output projection, text attention, modulation, gating, residuals, and FFN. Video K/V projections remain necessary; video Q/O projections, video text attention, and video FFNs are skipped.

At full anchor $j$, we store $C^v_{j,\ell} = V_{j,\ell} - x^v_j$, the residual of the video state entering suffix block $\ell$. The two latest full anchors $i < j < k$ reconstruct its current input:

$$\widetilde{V}_{k,\ell} = \begin{cases} P_k(x_k)^v, & \ell = p, \\ x^v_k + C^v_{j,\ell} + \gamma_k(C^v_{j,\ell} - C^v_{i,\ell}), & p < \ell < L. \end{cases} \tag{3}$$

The first suffix block reads the current prefix; deeper blocks use the inherited coefficient $\gamma_k$. From $A_{k,p} = P_k(x_k)^a$, audio follows

$$A_{k,\ell+1} = \mathcal{A}_\ell(\widetilde{V}_{k,\ell}, A_{k,\ell}; k, c), \qquad \ell = p, \ldots, L-1, \tag{4}$$

where $\mathcal{A}_\ell$ is the original block's audio-output path and $c$ is the text condition.

Selective steps return predicted video and recomputed audio. Their suffix passes video information to audio; bidirectional interaction remains in the joint prefix and full refreshes. Both outputs update sampler states. Reconstructed contexts and predicted outputs introduce trajectory approximation, distinct from the same-input audio-slicing check (Appendix A).

For JavisDiT's separate modality streams, selective steps compute native audio-directed cross-attention to current or reconstructed video states, then the original audio blocks. They skip the video-output direction and subsequent video blocks, retaining native conditioning and sampling. Appendix A details the grouping at interaction boundaries.

### 3.4 SHARED PREDICTOR AND FULL-ANCHOR UPDATES

Define shallow and full-stack residuals for $m \in \{v, a\}$ as

$$z^m_k = P_k(x_k)^m - x^m_k, \qquad R^m_k = F_k(x_k)^m - x^m_k. \tag{5}$$

We use the DiCache predictor (Bu et al., 2026), with one scalar shared by both modalities for each CFG row:

$$\gamma_k = \mathrm{clip}_{[1,2]}\left(\frac{\sum_{m\in\{v,a\}} \|z^m_k - z^m_i\|_1}{\max\left(\sum_{m\in\{v,a\}} \|z^m_j - z^m_i\|_1, \epsilon\right)}\right) - 1, \quad \epsilon = 10^{-8}. \tag{6}$$

The predicted stack output is

$$\widehat{y}^m_k = x^m_k + R^m_j + \gamma_k(R^m_j - R^m_i). \tag{7}$$

L1 reductions use FP32; residuals retain the model dtype. Only full computations update the two anchors, keeping predicted states out of reference histories. Missing or invalid history and nonfinite statistics trigger joint refreshes. JavisDiT++ joint-only controls share this predictor and prefix.

### 3.5 OPERATING POINTS AND COMPUTATIONAL COST

JavisDiT++ uses $N = 50$ steps and $L = 30$ joint blocks. Its JavisBench point uses $p = 2$ and 20 joint/six audio refreshes; T2AV uses $p = 1$ and 23 joint/three audio refreshes. JavisDiT uses

$N = 30$, $p = 1$, and ten joint/three audio refreshes. Selection is specified in Section 4.1, with exact schedules in Appendix A.

For $K_J$ joint refreshes and $K_A$ selective steps, the controlled stack uses $Np+(L-p)K_J$ joint block calls and $(L-p)K_A$ audio block calls. The JavisDiT++ configurations use 660 joint and 168 audio calls on JavisBench, and 717 and 87 on T2AV; their Native reference uses 1500 joint calls. JavisDiT additionally runs its initial unimodal groups every step. Joint/audio calls have different costs, so latency is measured directly. JavisDiT++ prefix-two contexts occupy 2.206 GiB, distinct from peak memory (Appendix C).

# 4 EXPERIMENTS

Tables 1 and 2 compare benchmarks and backbones; JavisDiT++ controls then examine allocation, timing, context, and prefix depth.

## 4.1 EVALUATION PROTOCOL

**Backbones and benchmarks.** Both backbones retain released weights, native geometries, and samplers in BF16. JavisDiT++ (Liu et al., 2026b) uses LoRA, FlashAttention2 (Dao, 2024), CFG scale 5 (Ho & Salimans, 2022), and 50 rectified-flow steps (Liu et al., 2023). The original JavisDiT v0.1 (Liu et al., 2026a) uses CFG scale 7 and 30 steps, without LoRA. JavisDiT++ covers all 10,140 JavisBench prompts (Liu et al., 2026a); both backbones cover all 500 T2AV-Compass prompts (Cao et al., 2026). Each main-table row uses one generation per prompt (seed 42), with stochastic inputs paired across methods within each backbone. JavisDiT++ controls use 128 prompts/two seeds for independent confirmation and 512 prompts/three seeds for ablations. Fixed JavisDiT++ T2AV configurations also undergo 256-prompt/two-seed new-input validation (Appendix G); output settings are in Appendix A.

**Baselines and operating points.** Native is the unchanged backbone. We adapt the authors' public implementations of TeaCache, PAB, TaylorSeer, and DiCache (Liu et al., 2025a; Zhao et al., 2025; Liu et al., 2025b; Bu et al., 2026) to joint generation. JavisBench baseline parameters were fixed after bounded development, using latency proximity for selection. On T2AV, DiCache uses thresholds 0.10/0.16 for JavisDiT++/JavisDiT. The T2AV AsymRefresh configurations and JavisDiT baseline settings were chosen after inspecting benchmark results and remain exploratory. Appendix A records parameters and selection details. Joint-only allocation controls share AsymRefresh's predictor and prefix.

**Metrics.** JavisBench reports FVD/FAD, ImageBind similarities (TV-IB, TA-IB, AV-IB), CLIP, CLAP, AVHScore, and JavisScore (Unterthiner et al., 2018; Kilgour et al., 2019; Guzhov et al., 2022; Girdhar et al., 2023; Radford et al., 2021; Wu et al., 2023b; Mao et al., 2024; Liu et al., 2026a). T2AV has six primary metrics: video technical quality (VT), video aesthetics (VA), audio aesthetics $\mathrm{AA} = (\mathrm{CU} + \mathrm{PQ})/2$, and audio–video, text–audio, and text–video similarities (Cao et al., 2026; Wu et al., 2023a; Tjandra et al., 2025; Girdhar et al., 2023; Wu et al., 2023b; Wang et al., 2024). The similarities use ImageBind, CLAP, and VideoCLIP-XL-v2, respectively, on all 500 prompts; T-V is reported on the cosine scale. Speech quality (SQ, NISQA (Mittag et al., 2021)) is a conditional diagnostic on 150 speech-requested prompts. Development and allocation studies additionally use VideoReward VQ/MQ/TA, CLIP, CLAP, and basic audio PQ (Liu et al., 2025c; Radford et al., 2021; Wu et al., 2023b; Tjandra et al., 2025). Common-scorer and official scores are kept separate (Appendix B).

**Study roles and uncertainty.** The 20-joint/six-audio schedule was fixed before independent confirmation; full benchmarks and expanded ablations include previously observed inputs. Paired intervals resample prompts after averaging seed differences within prompt. Confirmation and exploratory T2AV comparisons use marginal 95% intervals; budget, prefix, and new-input studies adjust their primary comparison families. Details are in Section 4.4 and Appendix B. Main-table ranks are descriptive; zero-crossing intervals establish neither a directional difference nor equivalence.

Table 1: Quality and generation time on JavisBench with JavisDiT++ (10,140 prompts). Ours uses 20 joint refreshes, six audio updates, and prefix depth two. Bold/underlined values mark the best/second-best accelerated quality scores, including ties at displayed precision; Native is the unaccelerated reference. ‡: separate four-prompt timing study, three repetitions per prompt.

| | **AV-Quality** | | **Text-Consistency** | | | | **AV-Consistency** | | **AV-Synchrony** | **Runtime** |
|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | FVD↓ | FAD↓ | TV-IB↑ | TA-IB↑ | CLIP↑ | CLAP↑ | AV-IB↑ | AVHScore↑ | JavisScore↑ | Gen.‡ (s) |
| Native | 146.9 | 5.8 | 0.287 | 0.156 | 0.315 | 0.446 | 0.183 | 0.168 | 0.146 | 34.70 |
| TeaCache | 144.7 | 6.0 | **0.287** | 0.155 | **0.315** | 0.444 | 0.180 | 0.166 | 0.144 | 19.06 |
| PAB | 144.6 | 6.1 | 0.286 | **0.156** | **0.315** | 0.445 | 0.180 | 0.167 | 0.145 | 23.57 |
| TaylorSeer | **141.9** | 5.9 | 0.286 | 0.155 | 0.313 | 0.442 | 0.181 | 0.166 | 0.144 | 19.77 |
| DiCache | 146.1 | **5.8** | **0.287** | **0.156** | **0.315** | 0.445 | **0.182** | **0.168** | **0.146** | 19.55 |
| **Ours** | 145.9 | **5.8** | **0.287** | **0.156** | **0.315** | **0.446** | **0.182** | **0.168** | **0.146** | 17.95 |

Table 2: T2AV-Compass results on 500 prompts for JavisDiT++ (50 steps) and JavisDiT v0.1 (30 steps). Ours uses 23/10 joint refreshes, respectively, with three audio updates and prefix depth one for both backbones. Bold/underlined values mark the best/second-best accelerated quality means within each backbone, including ties at displayed precision; Native is the unaccelerated reference. †: SQ covers 150 speech-requested prompts. ‡: timing cohorts are specified in Appendix C. Operating-point selection follows Section 4.1.

| | **Video Quality** | | **Audio Quality** | | **Cross-modal Alignment** | | | **Runtime** |
|---|---|---|---|---|---|---|---|---|
| **Method** | VT↑ | VA↑ | AA↑ | SQ† ↑ | A-V↑ | T-A↑ | T-V↑ | Gen.‡ (s) |
| **JavisDiT++** | | | | | | | | |
| Native | 6.4948 | 3.8822 | 5.5821 | 2.0223 | 0.166365 | 0.217110 | 0.2364 | 35.09 |
| TeaCache | 6.3376 | **3.8945** | 5.5631 | 1.9068 | 0.163266 | 0.217006 | 0.2360 | 19.29 |
| PAB | 6.3963 | 3.8760 | 5.5668 | 1.8757 | 0.163368 | 0.217185 | **0.2362** | 21.91 |
| TaylorSeer | 6.1628 | 3.8455 | **5.5992** | **2.0098** | 0.164397 | 0.215673 | 0.2347 | 19.95 |
| DiCache | 6.4474 | 3.8766 | 5.5799 | 1.9919 | **0.165650** | 0.217481 | 0.2360 | 18.00 |
| **Ours** | **6.4529** | 3.8795 | 5.5711 | 1.9811 | 0.165384 | **0.217579** | **0.2362** | 17.68 |
| **JavisDiT** | | | | | | | | |
| Native | 4.9699 | 3.1771 | 4.9004 | 1.3374 | 0.143336 | 0.160734 | 0.2206 | 88.82 |
| TeaCache | 4.4640 | 3.2446 | 4.7533 | 1.1158 | 0.133700 | 0.140959 | 0.2154 | 31.07 |
| PAB | 4.1331 | **3.3428** | 4.6874 | 0.9579 | 0.132301 | 0.134399 | 0.2124 | 40.37 |
| TaylorSeer | 4.2963 | 3.1309 | **4.8826** | **1.2935** | 0.132899 | 0.144417 | 0.2077 | 35.44 |
| DiCache | 4.6325 | 3.0327 | 4.8757 | 1.2388 | 0.135295 | 0.143620 | 0.2097 | 33.17 |
| **Ours** | **4.6896** | 3.0618 | 4.8804 | 1.2805 | **0.142141** | **0.162803** | **0.2169** | 33.10 |

**Latency.** Generation time includes encoding, denoising, cache/context operations, and decoding, excluding loading, warmup, scoring, and media export. JavisBench uses four development prompts and three interleaved repetitions on one MetaX C500. Budget and prefix studies are timed separately. T2AV retains distinct timing cohorts; JavisDiT Native uses full-benchmark generation logs (Appendix C).

### 4.2 Full-benchmark quality and efficiency

**JavisBench.** AsymRefresh has the lowest reported generation time in Table 1: 17.95 seconds versus 34.70 for Native, a $1.93\times$ speedup. At the displayed precision, it matches or exceeds the best accelerated-baseline value on eight of nine quality metrics, covering FAD, text similarities, and audiovisual consistency scores. FVD favors TaylorSeer, PAB, and TeaCache. Seven displayed quality scores match DiCache's, at 17.95 versus 19.55 seconds, combining shorter measured latency with competitive quality.

**T2AV-Compass: JavisDiT++.** In Table 2, AsymRefresh's VT, AA, A-V, and T-A means exceed those of TeaCache and PAB; TaylorSeer has the highest AA. Compared with DiCache, AsymRefresh

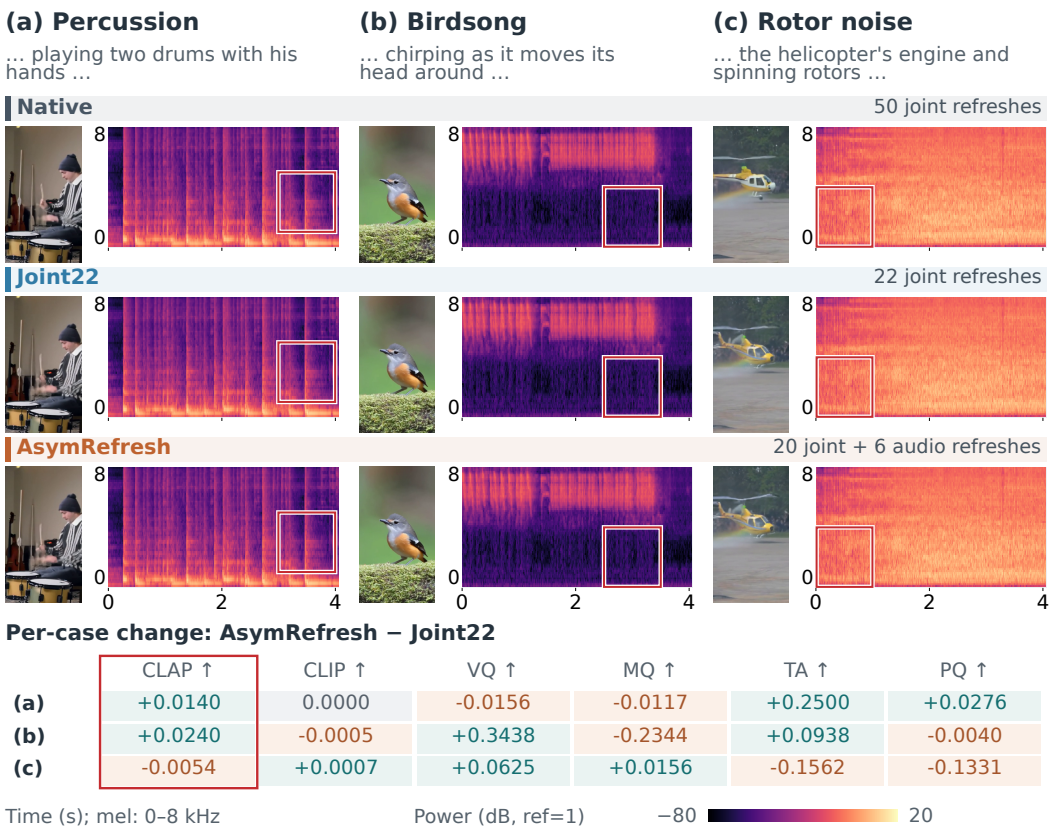



Figure 2: Paired JavisDiT++ examples: (a) percussion, (b) birdsong, and (c) rotor noise. Methods share each case's prompt and seed. Each cell pairs a frame at 2 s with the full 4.0625 s log-mel spectrum, using shared axes and power scales. The matrix reports all six AsymRefresh-minus-Joint22 score differences. Red boxes mark matched spectral windows and CLAP. Cases were selected for sound-source diversity after evaluation; selection details and full scores are in Appendix K.

has higher VT, VA, and T-A means and lower AA and A-V. The paired 95% intervals for these five differences and conditional SQ include zero (Appendix F).

A separate validation of these fixed configurations on 256 new prompts and two seeds also yields six Bonferroni-adjusted intervals containing zero (Appendix G).

**T2AV-Compass: JavisDiT.** With the original 30-step backbone, AsymRefresh attains the highest accelerated-method means on VT, A-V, T-A, and T-V, and the second-highest on AA and SQ. AsymRefresh improves all seven quality means over the selected DiCache configuration. PAB and TeaCache retain higher VA, while TaylorSeer leads AA and SQ.

Together, Tables 1 and 2 examine quality across two benchmark protocols and selective execution across joint-attention and separate-stream backbones. In both implementations, audio keeps access to video context within the native attention pathway while the selective suffix omits video-output computation. The following JavisDiT++ allocation controls then test how to spend additional computation: with the predictor and prefix held fixed within each budget comparison, selective audio updates are compared with extra joint refreshes at similar measured latency.

### 4.3 Selective audio versus additional joint refreshes

On a shared 20-joint-refresh schedule, six audio updates use 17.947 seconds versus 18.060 seconds for two additional joint refreshes (Joint22). Across all 10,140 JavisBench prompts, this allocation improves eight of nine quality scores, including FAD by 0.1050, CLAP by 0.00409, AV-IB by

Table 3: JavisDiT++ budget, stage, and context ablations at prefix depth two on 512 prompts and three seeds. Joint/Audio count scheduled updates; Gen. averages 12 separate development timings per configuration. Subscripts $c$ and $o$ denote common and official scorers. Late J22 adds joint refreshes at steps 40/45; the original full-benchmark Joint22 uses 30/40. Complete metrics, paired intervals, per-seed results, and memory appear in Appendix I.

| Configuration | Joint | Audio | Gen. (s)↓ | $\text{CLAP}_c$ ↑ | $\text{AV-IB}_o$ ↑ | $\text{JavisScore}_o$ ↑ |
|---|---|---|---|---|---|---|
| Native | 50 | 0 | 35.095 | 0.3130 | 0.1908 | 0.1544 |
| J20 | 20 | 0 | 17.039 | 0.3086 | 0.1868 | 0.1515 |
| Late J21 | 21 | 0 | 17.652 | 0.3074 | 0.1865 | 0.1511 |
| Late A3 | 20 | 3 | 17.654 | 0.3109 | 0.1887 | 0.1526 |
| Late J22 | 22 | 0 | 18.265 | 0.3115 | 0.1893 | 0.1530 |
| **AsymRefresh (late A6)** | 20 | 6 | 18.162 | 0.3122 | 0.1899 | 0.1534 |
| Middle A6 | 20 | 6 | 18.152 | 0.3085 | 0.1876 | 0.1517 |
| Held-context A6 | 20 | 6 | 18.110 | 0.3123 | 0.1894 | 0.1533 |

0.00244, and JavisScore by 0.00182; CLIP decreases by 0.000055. Joint22 adds steps 30/40, so modality and update location both differ; the late-stage controls below examine two further schedules. Appendix D retains all outcomes.

Independent confirmation (128 prompts, two seeds) gives common CLAP $+0.003512$ for AsymRefresh minus Joint22, with 95% interval $[+0.001338, +0.005817]$ and positive means for both seeds. The remaining five intervals cross zero. Appendix E reports all six metrics against every reference.

**Paired examples.** Figure 2 compares paired frames, spectra, and scores across percussion, birdsong, and rotor noise. Selection and complete per-case outcomes appear in Appendix K.

### 4.4 Budget allocation and ablations

JavisDiT++ ablations test allocation, update timing, context reconstruction, and prefix depth. Eight configurations use 512 prompts and seeds 42/43/44. Starting from Joint20 (J20), A3/A6 add three/six audio updates, while late J21/J22 add joint refreshes at $\{40\}/\{40, 45\}$. Middle A6 changes the six audio-update locations; held-context A6 removes only context extrapolation, retaining the output predictor and context banks. All eight use prefix depth two; three additional configurations test depth one. Ablation A3 uses J20, distinct from both T2AV main-table schedules.

Table 3 summarizes quality and separate four-prompt, three-repeat timing; Figure 3 shows paired effects. $\text{CLAP}_c$ denotes the common scorer. Four designated $\text{CLAP}_c$ contrasts use 10,000 prompt-cluster bootstrap draws with 98.75% individual Bonferroni intervals. The cohort includes 128 previously observed prompts, as detailed in Appendix I.

**Allocation at two budgets.** Three audio updates increase $\text{CLAP}_c$ from late J21's 0.307419 to 0.310937, at 17.654 versus 17.652 seconds. The paired gain is $+0.003518$, with adjusted interval $[+0.002603, +0.004421]$, and is positive for all three seeds. Six audio updates yield a smaller gain of $+0.000722$ over late J22, with adjusted interval $[+0.000133, +0.001316]$, at 18.162 versus 18.265 seconds. Figure 3(a) shows these paired alignment gains and their adjusted intervals. Official CLAP and AV-IB also favor the audio allocation at both budgets, with positive exploratory 95% intervals. Complete results, including the seed-dependent sign of the smaller A6 effect, appear in Appendix I.

**Update timing.** At a fixed 20-joint/six-audio budget, late A6 improves $\text{CLAP}_c$ over middle A6 by $+0.003701$, with adjusted interval $[+0.002474, +0.004938]$ (Figure 3(a), bottom row). This fixed-count comparison supports the tested late-stage choice.

**Context reconstruction.** The predicted-minus-held context difference in $\text{CLAP}_c$ is $-0.000076$, with interval $[-0.000637, +0.000461]$. Both variants retain the output predictor and context banks; the additional alignment benefit of intermediate context extrapolation remains unresolved.

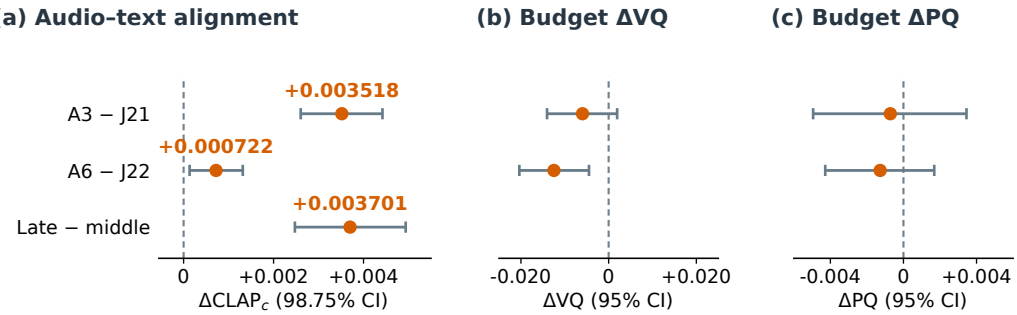



Figure 3: Paired JavisDiT++ effects at prefix depth two (512 prompts, three seeds). (a) CLAP$_c$ differences at two budgets (A3–J21, A6–J22) and two update stages (late minus middle A6), retaining the four-contrast Bonferroni intervals (98.75%). (b,c) VQ/basic-PQ differences for the same budget pairs (audio minus joint), with exploratory 95% intervals. All intervals use prompt-cluster bootstrap. Absolute scores and timing appear in Table 3; complete results in Appendix I.

Table 4: JavisDiT++ prefix-depth controls on 512 prompts and three seeds. The three primary CLAP$_c$ contrasts use 98.33% Bonferroni intervals. Interaction is $(\text{A6} - \text{late J22})_{p=1} - (\text{A6} - \text{late J22})_{p=2}$.

| Contrast | ΔCLAP$_c$ | Adjusted interval |
|---|---|---|
| A6 − late J22 at $p = 1$ | +0.000898 | $[+0.000301, +0.001496]$ |
| A6: $p = 1$ minus $p = 2$ | -0.000101 | $[-0.000672, +0.000460]$ |
| Allocation × prefix interaction | +0.000176 | $[-0.000596, +0.000914]$ |

**Prefix depth and allocation.** We repeat J20, late A6, and late J22 at depth one on the same prompts and seeds, reusing their depth-two controls. A6 improves CLAP$_c$ over late J22 (Table 4), with positive mean differences for all three seeds. The allocation gain is observed at both tested depths; the direct prefix comparison and interaction remain inconclusive. Depth-one A6 and late J22 take 17.287 and 17.427 seconds in separate timing (Appendix J).

**Quality and resource trade-offs.** For both original depth-two budget comparisons, the exploratory 95% intervals for basic audio PQ include zero (Figure 3(c)). A3 has lower MQ than late J21, and A6 has lower VQ/MQ than late J22. Late A6 also has lower basic audio PQ than middle A6; the intervals for these differences exclude zero. Depth-two audio configurations use 19.964 GiB peak allocated memory versus 17.400 GiB for joint-only controls. At depth one, A6 also has lower VQ and basic PQ than late J22, with negative exploratory 95% intervals, and adds 2.680 GiB of peak allocated memory. For complete metrics, per-seed differences, and resources, see Appendices I and J.

## 5 Discussion and Conclusion

AsymRefresh allocates refresh computation by modality while preserving video context within joint sampling. JavisBench and T2AV-Compass provide evidence across joint-attention JavisDiT++ and dual-stream JavisDiT. Controlled JavisDiT++ studies show audio–text alignment gains over extra joint refreshes at similar measured latency, supporting modality selection for efficient joint generation.

Alignment gains involve quality trade-offs and additional context memory. An independent extrapolation benefit and a preferred prefix depth remain unresolved. Evidence covers two JavisDiT backbones at native settings; other families, hardware, video-selective policies, and optimal schedules remain open. T2AV points were selected on observed results, and new-input validation is inconclusive. Timing scope is in Appendix C. Perceptual assessment and temporal-synchronization gains remain untested. Future policies could adapt allocation to inputs, compute, and memory.

### AI USE STATEMENT

We used GPT-5.6-sol for language polishing and analysis of experimental results. OpenAI Codex assisted with manuscript structure and editing, literature and citation checks, method–implementation consistency, interpretation and presentation of recorded results, selection and layout of qualitative examples, and preparation of experiment, scoring, aggregation, media-extraction, and plotting scripts. AI also authored and reviewed the supplementary validation prompts. GPT generated an initial architecture-figure reference and edited its labels and schematic connections. Quantitative results come from recorded experiment reports. The human authors retain responsibility for the claims, citations, measurements, and accompanying artifacts.

### ETHICS STATEMENT

AsymRefresh changes the inference schedule of pretrained audiovisual generators. Lower generation costs can support creative applications while also making misleading audiovisual content easier to produce. The method does not address biases or harmful capabilities of the underlying models, and the reported quality benchmarks are not safety evaluations. Use and redistribution of models, benchmark data, and generated media should respect their licenses, consent conditions, privacy interests, and usage restrictions.

### REPRODUCIBILITY STATEMENT

Section 3 specifies the predictor, context reconstruction, and anchor update rules. Appendix A records the exact schedules and baseline parameters, Appendix B specifies metric scope, and Appendix C gives the measurement boundaries. Appendices D–J retain complete quality outcomes and distinguish development, independent confirmation, and expanded ablations. Experiment records retain configuration identities, complete metric outcomes, and source manifests for tracing the reported studies.

## REFERENCES

Adudeguyman. ComfyUI-H3-AudioRefine. Software repository, 2026. URL https://github.com/Adudeguyman/ComfyUI-H3-AudioRefine. Reference implementation at commit d78d34f2f1100b0422047e9d784a3bdb4e1061d3.

Jiazi Bu, Pengyang Ling, Yujie Zhou, Yibin Wang, Yuhang Zang, Dahua Lin, and Jiaqi Wang. DiCache: Let Diffusion Model Determine Its Own Cache. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/78288ef33b18a351c3cd679dc9a15c8d-Abstract-Conference.html.

Zhe Cao, Tao Wang, Jiaming Wang, Yanghai Wang, Yuanxing Zhang, Jiahao Wang, Jialu Chen, Miao Deng, Chenxi Liao, Yize Zhang, Yubin Guo, Zhaoxiang Zhang, and Jiaheng Liu. T2AV-Compass: Towards Unified Evaluation for Text-to-Audio-Video Generation. In *International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://arxiv.org/abs/2512.21094.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. MMAudio: Taming Multimodal Joint Training for High-Quality Video-to-Audio Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28901–28911, 2025. doi: 10.1109/CVPR52734.2025.02691. URL https://openaccess.thecvf.com/content/CVPR2025/html/Cheng_MMAudio_Taming_Multimodal_Joint_Training_for_High-Quality_Video-to-Audio_Synthesis_CVPR_2025_paper.html.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/98ed250b203d1ac6b24bbcf263e3d4a7-Abstract-Conference.html.

Shengchuan Gao, Teng Hu, Bohao Feng, Luchen Li, Wenqiang Wang, Hongqian Deng, and Ran Yi. Efficient Audio-Visual Generation via Synchrony-Aware Cross-Modal Sparse Attention, 2026. URL https://arxiv.org/abs/2608.15522.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One Embedding Space To Bind Them All. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023. doi: 10.1109/CVPR52729.2023.01457. URL https://openaccess.thecvf.com/content/CVPR2023/html/Girdhar_ImageBind_One_Embedding_Space_To_Bind_Them_All_CVPR_2023_paper.html.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. AudioCLIP: Extending CLIP to Image, Text and Audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 976–980. IEEE, 2022. doi: 10.1109/ICASSP43922.2022.9747631. URL https://www.dfki.de/en/web/research/projects-and-publications/publication/12157.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient Synchronization from Sparse Cues. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5325–5329, 2024. doi: 10.1109/ICASSP48485.2024.10448489. URL https://doi.org/10.1109/ICASSP48485.2024.10448489.

Yujin Jeong, Wonjeong Ryoo, Seunghyun Lee, Dabin Seo, Wonmin Byeon, Sangpil Kim, and Jinkyu Kim. The Power of Sound (TPoS): Audio Reactive Video Generation with Stable Diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7822–7832, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Jeong_The_Power_of_Sound_TPoS_Audio_Reactive_Video_Generation_with_ICCV_2023_paper.html.

Yujin Jeong, Yunji Kim, Sanghyuk Chun, and Jiyoung Lee. Read, Watch and Scream! Sound Generation from Text and Video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17590–17598, 2025. doi: 10.1609/aaai.v39i17.33934. URL https://ojs.aaai.org/index.php/AAAI/article/view/33934.

Kumara Kahatapitiya, Haozhe Liu, Sen He, Ding Liu, Menglin Jia, Chenyang Zhang, Michael S. Ryoo, and Tian Xie. Adaptive Caching for Faster Video Generation with Diffusion Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15240–15252, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Kahatapitiya_Adaptive_Caching_for_Faster_Video_Generation_with_Diffusion_Transformers_ICCV_2025_paper.html.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms. In *Interspeech*, pp. 2350–2354, 2019. doi: 10.21437/Interspeech.2019-2219. URL https://www.isca-archive.org/interspeech_2019/kilgour19_interspeech.html.

Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep Embedding Tells: It's Time to Cache for Video Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7353–7363, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liu_Timestep_Embedding_Tells_Its_Time_to_Cache_for_Video_Diffusion_CVPR_2025_paper.html.

Jiacheng Liu, Chang Zou, Yuanhuiyi Lyu, Junjie Chen, and Linfeng Zhang. From Reusing to Forecasting: Accelerating Diffusion Models with TaylorSeers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15853–15863, 2025b. URL https://openaccess.thecvf.com/content/ICCV2025/html/Liu_From_Reusing_to_Forecasting_Accelerating_Diffusion_Models_with_TaylorSeers_ICCV_2025_paper.html.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving Video Generation with Human Feedback. In *Advances in Neural Information Processing Systems*, volume 38, 2025c. doi: 10.52202/085713-2750. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/76227feb18ea0ee40bd15cf02c33e18e-Abstract-Conference.html.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Fei, and Tat-Seng Chua. JavisDiT: Joint Audio-Video Diffusion Transformer with Hierarchical Spatio-Temporal Prior Synchronization. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e11afe29671baa67c9f15fa77fc97357-Abstract-Conference.html.

Kai Liu, Yanhao Zheng, Kai Wang, Shengqiong Wu, Rongjunchen Zhang, Jiebo Luo, Dimitrios Hatzinakos, Ziwei Liu, Hao Fei, and Tat-Seng Chua. JavisDiT++: Unified Modeling and Optimization for Joint Audio-Video Generation. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/f3c27bd0616d08c5168e18e8deb159c6-Abstract-Conference.html.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=xVjTT1nw5z.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin Backbone Cross-Modal Fusion for Audio-Video Generation, 2025. URL https://arxiv.org/abs/2510.01284.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. In *Advances in Neural Information Processing Systems*, volume 35, pp. 5775–5787, 2022. doi: 10.52202/068431-0418. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/260a14acce2a89dad36adc8eefe7c59e-Abstract-Conference.html.

Zhengyao Lyu, Chenyang Si, Junhao Song, Zhenyu Yang, Yu Qiao, Ziwei Liu, and Kwan-Yee K. Wong. FasterCache: Training-Free Video Diffusion Model Acceleration with High Quality. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/518046d86bbc41a0707727c38301ad8e-Abstract-Conference.html.

Juncheng Ma, Yuxuan Du, Yanan Sun, Zhening Xing, Changlin Li, Zhenyu Tang, Bo Li, Peng-Tao Jiang, Li Yuan, Daquan Zhou, and Yonghong Tian. SyncCache: Exploiting Asymmetric Dynamics for Fast Audio-Driven Portrait Animation. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2606.30849.

Xinyin Ma, Gongfan Fang, Michael Bi Mi, and Xinchao Wang. Learning-to-Cache: Accelerating Diffusion Transformer via Layer Caching. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-4236. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f0b1515be276f6ba82b4f2b25e50bef0-Abstract-Conference.html.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. DeepCache: Accelerating Diffusion Models for Free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15762–15772, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ma_DeepCache_Accelerating_Diffusion_Models_for_Free_CVPR_2024_paper.html.

Zehong Ma, Longhui Wei, Feng Wang, Shiliang Zhang, and Qi Tian. MagCache: Fast Video Generation with Magnitude-Aware Cache. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1153. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/311207bb626e36a8f1d3eb92aa67af22-Abstract-Conference.html.

Yuxin Mao, Xuyang Shen, Jing Zhang, Zhen Qin, Jinxing Zhou, Mochu Xiang, Yiran Zhong, and Yuchao Dai. TAVGBench: Benchmarking Text to Audible-Video Generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6607–6616, 2024. doi: 10.1145/3664647.3680612.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets. In *Interspeech*, pp. 2127–2131, 2021. doi: 10.21437/Interspeech.2021-299. URL https://www.isca-archive.org/interspeech_2021/mittag21_interspeech.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. MM-Diffusion: Learning Multi-Modal Diffusion Models for Joint Audio and Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10219–10228, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Ruan_MM-Diffusion_Learning_Multi-Modal_Diffusion_Models_for_Joint_Audio_and_Video_CVPR_2023_paper.html.

Pratheba Selvaraju, Tianyu Ding, Tianyi Chen, Ilya Zharkov, and Luming Liang. FORA: Fast-Forward Caching in Diffusion Transformer Acceleration, 2024. URL https://arxiv.org/abs/2407.01425.

Byunggwan Son, Jeimin Jeon, Jeongwoo Choi, and Bumsub Ham. Relational Feature Caching for Accelerating Diffusion Transformers. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4019681ef05e8e347c3067d7f973a364-Abstract-Conference.html.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta Audiobox Aesthetics: Unified Automatic Quality Assessment for Speech, Music, and Sound, 2025. URL https://arxiv.org/abs/2502.05139.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards Accurate Generative Models of Video: A New Metric & Challenges, 2018. URL https://arxiv.org/abs/1812.01717.

Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang Yu. UniVerse-1: Unified Audio-Video Generation via Stitching of Experts, 2025. URL https://arxiv.org/abs/2509.06155.

Jiapeng Wang, Chengyu Wang, Kunzhe Huang, Jun Huang, and Lianwen Jin. VideoCLIP-XL: Advancing Long Description Understanding for Video CLIP Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16061–16075. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.898. URL https://aclanthology.org/2024.emnlp-main.898/.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring Video Quality Assessment on User Generated Contents from Aesthetic and Technical Perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20144–20154, 2023a. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Wu_Exploring_Video_Quality_Assessment_on_User_Generated_Contents_from_Aesthetic_ICCV_2023_paper.pdf.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5. IEEE, 2023b. doi: 10.1109/ICASSP49357.2023.10095969.

Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and Hearing: Open-domain Visual-Audio Generation with Diffusion Latent Aligners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7151–7161, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Xing_Seeing_and_Hearing_Open-domain_Visual-Audio_Generation_with_Diffusion_Latent_Aligners_CVPR_2024_paper.html.

Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. Diverse and Aligned Audio-to-Video Generation via Text-to-Video Model Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6639–6647, 2024. doi: 10.1609/aaai.v38i7.28486. URL https://ojs.aaai.org/index.php/AAAI/article/view/28486.

Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, Bin Liu, and Kai Chen. FoleyCrafter: Bring Silent Videos to Life with Lifelike and Synchronized Sounds. *International Journal of Computer Vision*, 134(1):46, 2026. doi: 10.1007/s11263-025-02649-3. URL https://link.springer.com/article/10.1007/s11263-025-02649-3.

Yuechen Zhang, Jinbo Xing, Bin Xia, Shaoteng Liu, Bohao Peng, Xin Tao, Pengfei Wan, Eric Lo, and Jiaya Jia. Training-Free Efficient Video Generation via Dynamic Token Carving. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2738. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/75b2a6632ab7b9196e9cb93750986e45-Abstract-Conference.html.

Xuanlei Zhao, Xiaolong Jin, Kai Wang, and Yang You. Real-Time Video Generation with Pyramid Attention Broadcast. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/092c2d45005ea2db40fc24c470663416-Abstract-Conference.html.

Shikang Zheng, Liang Feng, Xinyu Wang, Qinming Zhou, Peiliang Cai, Chang Zou, Jiacheng Liu, Yuqi Lin, Junjie Chen, Yue Ma, and Linfeng Zhang. Forecast Then Calibrate: Feature Caching as ODE for Efficient Diffusion Transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13449–13457, 2026. doi: 10.1609/aaai.v40i16.38349. URL https://ojs.aaai.org/index.php/AAAI/article/view/38349.

Xin Zhou, Dingkang Liang, Kaijin Chen, Tianrui Feng, Xiwu Chen, Hongkai Lin, Yikang Ding, Feiyang Tan, Hengshuang Zhao, and Xiang Bai. Less is Enough: Training-Free Video Diffusion Acceleration via Runtime-Adaptive Caching, 2025. URL https://arxiv.org/abs/2507.02860.

Chang Zou, Xuyang Liu, Ting Liu, Siteng Huang, and Linfeng Zhang. Accelerating Diffusion Transformers with Token-wise Feature Caching. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bbe024e0517fe12ac3a8d388b19ff9fe-Abstract-Conference.html.

# A IMPLEMENTATION DETAILS

**Backbone output settings.** JavisDiT++ generates $432 \times 240$ video with 65 frames at 16 FPS and 65,000 audio samples at 16 kHz (4.0625 seconds). JavisDiT v0.1 generates $424 \times 240$ video with 102 frames at 24 FPS and 68,000 audio samples at 16 kHz (4.25 seconds). Each backbone retains its native geometry across all compared methods. Weights, precision, CFG, and sampling steps are specified in Section 4.1.

**JavisDiT++ execution schedules.** All indices are zero-based sampling steps in execution order.

$$\mathcal{J}_{20} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 14, 18, 23, 28, 33, 38, 42, 47, 48, 49\},$$
$$\mathcal{A}_{\text{late},6} = \{40, 41, 43, 44, 45, 46\},$$
$$\mathcal{J}_{22} = \mathcal{J}_{20} \cup \{30, 40\}, \qquad \mathcal{J}_{24} = \mathcal{J}_{20} \cup \{30, 35, 40, 45\}.$$

A configured audio window of steps 40–49 contains only six selective updates because joint refresh takes precedence at 42 and 47–49. Reused or selectively updated outputs do not become full anchors. In the held-context variant, Equation 3 drops only the extrapolation term for $\ell > p$; the first suffix block still uses current shallow video, and the main video-output residual predictor remains active. Holding these contexts therefore does not disable the predictor.

The separately selected JavisDiT++ T2AV configuration instead uses $\mathcal{J}_{23} = \mathcal{J}_{20} \cup \{41, 43, 45\}$, $\mathcal{A}_3^{\text{T2AV}} = \{40, 44, 46\}$, and prefix depth 1. It executes 717 joint block calls and 87 audio-only block calls; these counts are not a latency estimate.

**JavisDiT v0.1 adaptation and schedule.** The original backbone interleaves separate video and audio spatial, temporal, and prior blocks with bidirectional audiovisual cross-attention. The first unimodal groups execute at every sampling step. A controlled block then consists of native cross-attention at depth $\ell$, followed by the unimodal groups at depth $\ell + 1$, except at the final depth. Thus $x_k$ in Section 3 is the pair of states after the first unimodal groups, and stored video contexts are inputs to native cross-attention. A selective block retains the audio-directed attention to video keys and values, including its normalization, projection, gate, and residual, then executes the next audio group. The video-output direction and next video group are omitted. Input embeddings, text and prior conditioning, masks, output heads, and the original sampler are retained. The same residual predictor and full-anchor update rule operate at this boundary; Equation 2 specifically describes JavisDiT++'s joint attention, rather than replacing JavisDiT's native cross-attention.

The 30-step T2AV operating point uses prefix depth one and

$$\mathcal{J}_{10}^{\text{JavisDiT}} = \{0, 1, 3, 6, 10, 15, 21, 25, 27, 29\},$$
$$\mathcal{A}_3^{\text{JavisDiT}} = \{24, 26, 28\}.$$

This is a separate schedule for the original backbone, not a rescaling of the 50-step JavisDiT++ schedule.

Block-call counts refer to batched evaluations including the original CFG rows. They do not equate joint and audio calls or represent measured FLOPs. The nominal counts apply to the stated schedule; fallback refreshes, when needed, execute and record full computation. The JavisDiT++ audio-slice warmup compares one original block and its sliced path on identical inputs in finite precision. This checks the slicing implementation, separately from the approximation introduced by historical context.

**Generation randomness.** For JavisDiT++, initial latent-noise seeds are deterministic functions of the prompt identifier, generation seed, and modality. The same tuple is shared across methods, while different prompt identifiers receive their own video and audio noise. Thus seed 42 specifies one paired generation per prompt, not one common noise tensor for the entire benchmark. The JavisDiT adapter instead resets Python, NumPy, and Torch CPU/GPU random states to seed 42 before each sample, pairing latent noise and sampled prior randomness across methods. These are sampling seeds; pretrained weights remain fixed across all studies.

**Public method parameters and identity.** The JavisDiT++ DiCache adaptation computes joint audio–video statistics and refreshes the batched computation when either CFG row requests it. Public method names identify the adapted mechanism, not an author-released AV implementation.

Table 5: Original public baseline operating points used on JavisBench. The JavisDiT++ T2AV block uses DiCache threshold 0.10; the other listed baseline parameter choices are retained. These local AV adaptations are distinct from the fixed-schedule allocation controls.

| Method | Fixed parameters | Computation reused |
|---|---|---|
| TeaCache (Liu et al., 2025a) | threshold 0.08; warmup 1, final 1 | joint stack residual |
| PAB (Zhao et al., 2025) | period 4; $100 < \text{timestep} < 850$ | attention outputs |
| TaylorSeer (Liu et al., 2025b) | interval 2; order 1 | module features |
| DiCache (Bu et al., 2026) | threshold 0.08; prefix 1; retention 0.2 | joint stack residual |

For the JavisDiT++ TeaCache, TaylorSeer, and PAB adaptations, the retained candidate pairs were thresholds $\{0.07, 0.08\}$, intervals $\{2, 3\}$ (order 1), and periods $\{3, 4\}$, respectively. From the candidates retained in development32, selection minimized the absolute generation-time difference from AsymRefresh using the earlier development16 study's four-prompt isolated timing. The choices agreed on both datasets. These opened development measurements determine the fixed parameters; the full-benchmark scores do not enter the selection rule.

For JavisDiT v0.1, Table 2 uses TeaCache threshold 0.30; PAB spatial/temporal/cross periods 4/8/12 over timesteps 300–970; TaylorSeer refresh interval four and order one; and DiCache threshold 0.16, prefix depth one, and retention ratio 0.2. These implementations retain the public algorithms with the required audiovisual interfaces.

**T2AV operating-point selection.** The JavisDiT++ AsymRefresh point was selected from 24 configurations on the observed 500-prompt set by equal-weight average ranks over the six primary metrics. The JavisDiT rows use separately chosen configurations from the completed parameter study, after inspection of full-benchmark results. Both T2AV blocks therefore report exploratory operating points; the JavisDiT++ development rule above does not describe selection of the JavisDiT rows. SQ is a conditional diagnostic and is not part of the six-primary-metric selection criterion.

## B Metric definitions and evaluation scope

**JavisBench.** The original JavisBench protocol and JavisScore follow JavisDiT (Liu et al., 2026a). FVD uses I3D features (Unterthiner et al., 2018); the released FAD implementation uses AudioCLIP features (Kilgour et al., 2019; Guzhov et al., 2022), so its values must be compared within that feature and preprocessing configuration. ImageBind provides text–video, text–audio, and audio–video similarities (Girdhar et al., 2023); CLIP and CLAP provide their respective text-alignment scores (Radford et al., 2021; Wu et al., 2023b). AVHScore follows the referenced audiovisual evaluation protocol (Mao et al., 2024). FVD and FAD are lower-is-better; the remaining listed metrics are higher-is-better. These aggregate distances and semantic similarities do not directly measure event timing.

We preserve all 10,140 benchmark rows, including repeated prompt text. The main-benchmark distribution features aggregate all seed-42 outputs with one copy of the reference features. The official reference protocol uses 16 video frames and a maximum audio length of 4 seconds for distribution features. Feature-extraction sampling does not change the generated media geometry.

**T2AV-Compass.** Both backbone blocks in Table 2 use $\text{AA} = (\text{CU} + \text{PQ})/2$, CLAP audio–text similarity, and VideoCLIP-XL-v2 video–text cosine similarity (T-V) (Wang et al., 2024), each on all 500 inputs. We divide the released script's $100\times$ cosine scores by 100 for the T-V column, matching the benchmark's published scale. Text inputs are the released modality descriptions; the empty audio description at index 103 is retained without a prompt fallback. NISQA SQ (Mittag et al., 2021) uses the same 150 speech-requested prompts for every displayed method; the remaining 350 prompts are marked not applicable. SQ measures predicted speech quality on speech-requested inputs, without a separate verification that speech was generated.

VT uses the benchmark's transformed DOVER technical score (Wu et al., 2023a), displayed on its specified $\times 100$ scale. VA uses the released SigLIP-based aesthetic predictor. The legacy AA4 summary averages the four AudioBox Aesthetics axes—content enjoyment (CE), content usefulness

(CU), production complexity (PC), and production quality (PQ)—under the benchmark configuration (Tjandra et al., 2025; Cao et al., 2026). The original allocation tables retain AA4 and the legacy ImageBind text similarities. A–V and these two text similarities use ImageBind and the benchmark's preprocessing (Girdhar et al., 2023). TA-IB uses *audio descriptions* and TV-IB uses *video descriptions*, rather than substituting the integrated prompt.

Under that legacy text protocol, the eight empty audio descriptions are indices $\{3, 151, 157, 175, 408, 424, 438, 468\}$: TA-IB has 492 valid observations and eight specified N/A records. Its remaining primary groups use all 500 observations. Failed evaluations are not silently removed from the denominator. The native JavisDiT++ and JavisDiT outputs last 4.0625 and 4.25 seconds, respectively. Neither meets the adopted 4.8-second DeSync window (Iashin et al., 2024), so DeSync is omitted from the tables.

**Development and confirmation.** VideoReward provides visual quality (VQ), motion quality (MQ), and text alignment (TA) (Liu et al., 2025c); CLIP, CLAP, and basic audio PQ complete the six-metric suite. The basic PQ scorer uses the saved waveform, whereas the official T2AV AA4 protocol scores audio extracted from the MP4. Protocol-specific readers, sampling, and preprocessing remain fixed. These scores are not pooled with the official benchmark columns. All six development metrics are higher-is-better.

**Paired uncertainty estimates.** The independent 128-prompt confirmation averages the two seed differences within each prompt before 10,000 prompt-cluster bootstrap draws. The exploratory 500-prompt JavisDiT++ T2AV comparisons use 5,000 paired prompt-bootstrap draws. Both report marginal, unadjusted 95% intervals; selection-set intervals do not account for operating-point selection. The expanded ablation, prefix extension, and fixed-configuration new-input study specify their separate primary comparison families and resampling counts in Appendices I, J, and G, respectively.

## C TIMING AND MEMORY

Table 6: Original JavisBench operating-point timing on one C500: 12 measurements per configuration (four prompts, seed 42, three repetitions). SD is the sample standard deviation across measurements. Peak allocated memory is distinct from reserved memory, which is approximately 28.52 GiB for the listed configurations. These are the timings used in Table 1, separate from T2AV and the later ablation timing study.

| Configuration | Generation (s) | SD (s) | With export (s) | Peak alloc. (GiB) |
|---|---|---|---|---|
| Native | 34.701 | 0.134 | 36.053 | 17.40 |
| TeaCache | 19.058 | 0.025 | 20.428 | 17.40 |
| PAB | 23.574 | 0.037 | 24.982 | 18.82 |
| TaylorSeer | 19.774 | 0.144 | 21.165 | 23.82 |
| DiCache | 19.553 | 0.291 | 20.904 | 17.40 |
| Joint-only (22) | 18.060 | 0.025 | 19.449 | 17.40 |
| **AsymRefresh (ours)** | 17.947 | 0.052 | 19.285 | 19.96 |

The generation boundary includes text encoding, sampling, online cache statistics, video-context reconstruction, video K/V and audio suffix computation, and decoding. Media export is measured separately. Loading, warmup, and scoring are outside this per-sample boundary. For input-adaptive methods, a four-prompt cohort provides a limited estimate of latency under another prompt distribution. The 2.206 GiB video-context figure measures persistent banks, not total allocated or reserved device memory.

**JavisDiT v0.1 runtime scope.** For the original backbone, generation time includes text/prior encoding, rectified-flow sampling, and decoding, with device synchronization; model loading, latent hashing, scoring, and media export are excluded. The four public-baseline entries in Table 2 use separate four-prompt, three-repetition operating-point measurements on equivalently configured C500 hosts. The Native entry instead averages the 500 original per-sample runtime records collected dur-

ing concurrent full-benchmark generation. These cohorts are distinct, and the Native runtime is not an exclusive-device measurement paired with the accelerated rows.

## D COMPLETE ALLOCATION COMPARISONS

Table 7: All nine JavisBench outcomes for AsymRefresh and Joint22, seed 42. FVD/FAD are lower-is-better; other quality metrics are higher-is-better. The eight improved scores are descriptive, without nine individual significance claims. ‡: common development timing.

| Configuration | FVD↓ | FAD↓ | TV-IB | TA-IB | AV-IB | CLIP | CLAP | AVH | Javis | Gen.‡ (s) |
|---|---|---|---|---|---|---|---|---|---|---|
| Joint-only (22 refreshes) | 146.136 | 5.955 | 0.28679 | 0.15570 | 0.17950 | 0.31498 | 0.44154 | 0.16548 | 0.14381 | 18.060 |
| **AsymRefresh (20 joint + 6 audio)** | 145.873 | 5.850 | 0.28701 | 0.15599 | 0.18194 | 0.31492 | 0.44563 | 0.16760 | 0.14563 | 17.947 |

Table 8: Legacy JavisDiT++ T2AV quality groups for the original 20-joint/six-audio allocation comparison: 500 inputs, with 492 valid TA-IB descriptions, seed 42. All columns are higher-is-better. AA4 and ImageBind text similarities follow the historical protocol; these are distinct from the AA, T-A, and T-V columns and the 23-joint/three-audio point in Table 2.

| Configuration | VT↑ | VA↑ | AA4↑ | A–V↑ | TA-IB↑ | TV-IB↑ |
|---|---|---|---|---|---|---|
| Joint-only (22 refreshes) | 6.423 | 3.882 | 4.555 | 0.16198 | 0.13927 | 0.06700 |
| **AsymRefresh (20 joint + 6 audio)** | 6.440 | 3.875 | 4.559 | 0.16525 | 0.13774 | 0.06733 |

## E COMPLETE INDEPENDENT CONFIRMATION

Table 9: AsymRefresh minus each reference on 128 previously unused prompts, seeds 42 and 43. All six outcomes are higher-is-better. Intervals use 10,000 prompt-cluster resamples with both seeds averaged within prompt; they are marginal 95% intervals. Per-seed mean differences are descriptive. This original confirmation remains separate from the later ablations, including the expanded 512-prompt, three-seed study in Appendix I.

| Reference | Metric | Difference | CI lower | CI upper | Seed 42 | Seed 43 |
|---|---|---|---|---|---|---|
| Native | VQ | -0.029042 | -0.058289 | 0.000012 | -0.050470 | -0.007614 |
| Native | MQ | -0.019765 | -0.039019 | -0.000347 | -0.022427 | -0.017104 |
| Native | TA | -0.001035 | -0.056899 | 0.056683 | 0.080473 | -0.082542 |
| Native | CLIP | 0.000138 | -0.000613 | 0.000885 | 0.000375 | -0.000099 |
| Native | CLAP | -0.001338 | -0.003000 | 0.000367 | -0.001076 | -0.001599 |
| Native | PQ | -0.017406 | -0.032542 | -0.001938 | -0.022188 | -0.012625 |
| Joint-only (20) | VQ | 0.000606 | -0.018632 | 0.020172 | -0.005930 | 0.007141 |
| Joint-only (20) | MQ | -0.011519 | -0.025250 | 0.002430 | -0.020442 | -0.002595 |
| Joint-only (20) | TA | -0.008698 | -0.040962 | 0.023277 | 0.031067 | -0.048462 |
| Joint-only (20) | CLIP | -0.000005 | -0.000237 | 0.000224 | 0.000173 | -0.000183 |
| Joint-only (20) | CLAP | 0.002805 | 0.000655 | 0.005032 | 0.002026 | 0.003583 |
| Joint-only (20) | PQ | -0.008881 | -0.020500 | 0.002456 | -0.010469 | -0.007294 |
| Joint-only (22) | VQ | 0.012022 | -0.007723 | 0.032280 | 0.009090 | 0.014954 |
| Joint-only (22) | MQ | -0.012092 | -0.026257 | 0.001493 | -0.014065 | -0.010119 |
| Joint-only (22) | TA | 0.019394 | -0.016846 | 0.055680 | 0.086403 | -0.047615 |
| Joint-only (22) | CLIP | -0.000224 | -0.000540 | 0.000097 | 0.000107 | -0.000555 |
| Joint-only (22) | CLAP | 0.003512 | 0.001338 | 0.005817 | 0.001982 | 0.005043 |
| Joint-only (22) | PQ | -0.002380 | -0.013926 | 0.008943 | -0.002344 | -0.002416 |
| Joint-only (24) | VQ | -0.004846 | -0.024476 | 0.014976 | -0.013445 | 0.003754 |
| Joint-only (24) | MQ | -0.014864 | -0.028434 | -0.001644 | -0.027444 | -0.002283 |
| Joint-only (24) | TA | -0.026201 | -0.064126 | 0.011526 | 0.026154 | -0.078556 |
| Joint-only (24) | CLIP | 0.000041 | -0.000245 | 0.000328 | 0.000067 | 0.000015 |
| Joint-only (24) | CLAP | -0.000807 | -0.002042 | 0.000454 | -0.001428 | -0.000185 |
| Joint-only (24) | PQ | 0.000682 | -0.008435 | 0.009476 | 0.000537 | 0.000827 |
| DiCache | VQ | -0.013251 | -0.039870 | 0.014620 | -0.033842 | 0.007339 |
| DiCache | MQ | -0.004807 | -0.022385 | 0.012525 | -0.021135 | 0.011522 |
| DiCache | TA | -0.000514 | -0.052168 | 0.053249 | 0.027865 | -0.028893 |
| DiCache | CLIP | -0.000283 | -0.000847 | 0.000281 | -0.000193 | -0.000374 |
| DiCache | CLAP | 0.000001 | -0.001538 | 0.001657 | -0.000005 | 0.000008 |
| DiCache | PQ | -0.014402 | -0.025897 | -0.003527 | -0.025256 | -0.003548 |

## F ADDITIONAL JAVISDIT++ T2AV COMPARISONS

For the selected 23-joint/three-audio AsymRefresh and threshold-0.10 DiCache configurations, the paired differences on 500 prompts are VT $+0.005496$ ($[-0.014979, +0.026964]$), VA $+0.002916$ ($[-0.004020, +0.009955]$), AA $-0.008746$ ($[-0.018259, +0.000524]$), A-V $-0.000266$ ($[-0.001647, +0.001115]$), and T-A $+0.000098$ ($[-0.000938, +0.001183]$). Differences are AsymRefresh minus DiCache; parentheses give the marginal 95% intervals from 5,000 paired prompt-bootstrap draws under the exploratory protocol in Appendix B.

On the 150 speech-requested prompts, the AsymRefresh–DiCache paired SQ difference is $-0.010775$, with unadjusted 95% interval $[-0.032769, +0.010446]$. This conditional diagnostic uses the shared prompt subset specified in Appendix B; the interval does not establish equivalence.

Table 10: JavisDiT++ audio components on all 500 T2AV prompts, seed 42, matching the configurations in Table 2. AA4 averages CE/CU/PC/PQ; both main-table backbone blocks use AA=(CU+PQ)/2. All columns are higher-is-better. These production-quality values use the official audio protocol, separately from basic development PQ.

| Method | CE | CU | PC | PQ | AA4 |
|---|---|---|---|---|---|
| Native | 3.7638 | 5.2018 | 3.3552 | 5.9623 | 4.5708 |
| TeaCache | 3.7136 | 5.1827 | 3.3155 | 5.9434 | 4.5388 |
| PAB | 3.7062 | 5.1878 | 3.2995 | 5.9458 | 4.5348 |
| TaylorSeer | 3.7333 | 5.2174 | 3.2951 | 5.9811 | 4.5567 |
| DiCache | 3.7479 | 5.2001 | 3.3477 | 5.9597 | 4.5639 |
| **Ours** | 3.7417 | 5.1951 | 3.3479 | 5.9472 | 4.5579 |

## G FIXED-CONFIGURATION VALIDATION ON NEW PROMPTS

The exploratory JavisDiT++ T2AV operating points were frozen before this supplementary study: AsymRefresh uses 23 joint refreshes, three audio updates, and prefix one; DiCache uses threshold 0.10, prefix one, and retention 0.2. We evaluate both on 256 new prompts with seeds 42 and 43, using the generation geometry and six primary metric definitions in Section 4.1. All 1,024 media outputs and 6,144 scalar scores are complete, with no missing, failed, or unscorable items. The prompts were authored and reviewed by AI; they were not human-reviewed. This separately assembled input set is not an official T2AV-Compass split, and no exclusion from the pretrained model's training data is asserted.

For each metric, we average the two paired method differences within each prompt and perform 50,000 prompt-bootstrap resamples. All six comparisons belong to the prespecified primary family, using individual Bonferroni intervals at $1 - 0.05/6 = 99.1667\%$. The intervals condition on the two measured seeds. Table 11 reports the means, differences, and adjusted intervals; Table 12 retains marginal 95% intervals and both seed differences. All adjusted intervals include zero, so this study establishes neither a directional quality difference nor equivalence. Parallel generation supplies no latency comparison. These fixed-configuration results are distinct from the selected 500-prompt benchmark point and from the component ablations.

## H DEVELOPMENT STUDIES AND OPERATING-POINT SELECTION

The diagnostics in this section use JavisDiT++.

**Modality correction as a diagnostic.** On separate 32-prompt Javis and T2AV development cohorts with two seeds, we completed the full suffix at selected reuse steps, replaced only the audio output, and continued sampling. Relative to the shared 20-refresh trajectory, late correction (steps 30–49) increased common CLAP by $+0.007486$ on Javis and $+0.006802$ on T2AV, with exploratory 95% intervals $[+0.002899, +0.012570]$ and $[+0.001719, +0.012504]$. Middle correction (steps 10–29) gave negative mean differences; only the Javis interval excluded zero. The windows used 14 and 16 interventions, respectively, so they motivated a stage hypothesis rather than isolating a count-matched effect. Because correction required an extra full suffix, this was a diagnostic of allocation

Table 11: Fixed-configuration validation on 256 new AI-authored and AI-reviewed prompts, seeds 42/43. Difference is AsymRefresh minus DiCache. All six prespecified 99.1667% Bonferroni intervals cross zero. T–V is VideoCLIP-XL-v2 cosine similarity. This is separate from the official 500-prompt benchmark and from component ablation.

| Metric | AsymRefresh | DiCache | Difference | 99.1667% interval |
|---|---|---|---|---|
| VT | 5.901257 | 5.902795 | -0.001538 | $[-0.026429, +0.023503]$ |
| VA | 3.706107 | 3.709396 | -0.003290 | $[-0.014140, +0.007444]$ |
| AA | 5.247248 | 5.245362 | +0.001886 | $[-0.010385, +0.014143]$ |
| A–V | 0.134421 | 0.134077 | +0.000344 | $[-0.001383, +0.002047]$ |
| T–A | 0.199123 | 0.200139 | -0.001016 | $[-0.002545, +0.000516]$ |
| T–V | 0.222987 | 0.223102 | -0.000114 | $[-0.000435, +0.000189]$ |

Table 12: Marginal 95% intervals and descriptive seed-wise differences for all six new-prompt outcomes. Inference uses the corrected intervals in Table 11; there is no per-seed significance claim.

| Metric | 95% interval | Seed 42 | Seed 43 |
|---|---|---|---|
| VT | $[-0.020031, +0.016855]$ | -0.006551 | +0.003476 |
| VA | $[-0.011325, +0.004743]$ | -0.000732 | -0.005847 |
| AA | $[-0.007197, +0.011000]$ | -0.000044 | +0.003816 |
| A–V | $[-0.000947, +0.001599]$ | -0.000005 | +0.000693 |
| T–A | $[-0.002157, +0.000122]$ | -0.001802 | -0.000230 |
| T–V | $[-0.000348, +0.000112]$ | -0.000261 | +0.000032 |

potential, not a deployable speedup. The equal-count controls below and in Appendix I evaluate the efficient selective path.

**Stage at the earlier 14-update operating point.** On 32-prompt development sets with seeds 42 and 43, late minus middle audio refresh gave CLAP differences of $+0.01044$ on Javis and $+0.00710$ on T2AV, with exploratory prompt-cluster 95% intervals $[+0.00611, +0.01494]$ and $[+0.00121, +0.01360]$. Both used 14 audio updates; measured generation times were 19.197 and 19.204 seconds for late and middle, respectively. These times come from a separate earlier isolated C500 study using four fixed development prompts, seed 42, and three repetitions. Both quality cohorts share this timing reference; it is neither their 32-prompt mean latency nor a remeasurement in the current main-table timing study. This comparison identifies a difference between those fixed windows. It does not establish a universal order in which modalities converge. Later windows starting at steps 30 and 40 changed both timing and update count and are not a single-factor stage test.

**Context and budget trade-offs.** Predicted minus held contexts gave CLAP differences of $-0.000324$ on Javis and $+0.001215$ on T2AV, with both intervals crossing zero. At the 14-update point, a near-equal-time 24-joint control had higher T2AV VQ by 0.05248, with interval $[0.01471, 0.09228]$. With 16 joint refreshes, selective updates increased CLAP but reduced Javis VQ relative to a near-equal-time 20-joint control by 0.10931, with interval $[-0.17603, -0.05649]$. These results motivated a more conservative 20-joint/six-audio operating point, fixed before independent confirmation.

**Exploratory coefficient separation.** An earlier 32-prompt, two-seed study replaced the shared predictor with a bounded modality-wise least-squares fit. Separating the coefficients reduced the local fitting objective, but the matched final VQ difference was $+0.0160$ with interval $[-0.0471, +0.0828]$; seed-specific differences had opposite signs, and all six final-metric intervals crossed zero. This comparison motivates evaluating final samples rather than treating a smaller shallow residual error as sufficient evidence of improvement. The present method uses the DiCache coefficient in Equation 6; coefficient separation is not part of AsymRefresh. The detailed historical derivation and results are preserved in the project archive, separately from the submitted method description.

# I LATE-STAGE BUDGET ALLOCATION AND CONTROLS

All configurations in this study use JavisDiT++.

## I.1 EXPANDED COHORT AND FIXED SCHEDULES

The expanded ablation uses 512 fixed JavisBench prompts: the 128 prompts from the earlier ablation and 384 additional prompts selected by a fixed hash rule from the remaining benchmark inputs. Original prompt text and official metadata are retained. Each of eight configurations uses all three seeds (42, 43, 44), producing 12,288 clips with complete coverage for all 13 per-sample metrics. No failed-input intersection or score-dependent subset is used. Because the original cohort and benchmark were already observed, this expansion is not an untouched confirmation set. The independent 128-prompt, two-seed study remains separately reported in Appendix E.

Weights, LoRA, geometry, precision, CFG, and sampling follow Section 4.1; the primary 20-joint/six-audio configuration is unchanged. With $\mathcal{J}_{20}$ from Appendix A, the zero-based schedules are

$$\mathcal{A}_3 = \{40, 43, 46\}, \qquad \mathcal{A}_6 = \{40, 41, 43, 44, 45, 46\},$$
$$\mathcal{J}_{21}^{\text{late}} = \mathcal{J}_{20} \cup \{40\}, \qquad \mathcal{J}_{22}^{\text{late}} = \mathcal{J}_{20} \cup \{40, 45\},$$
$$\mathcal{A}_{\text{mid},6} = \{20, 21, 24, 25, 26, 27\}.$$

A3 and A6 use predicted contexts; middle A6 changes only the audio-update locations, and held-context A6 changes the context reconstruction rule. The held control retains the same video-context banks. Native computes all 30 joint blocks at every step; J20/J21/J22 have no selective audio updates. The full-benchmark Joint22 control adds $\{30, 40\}$, so its results must not be substituted for late J22 in this study. The compared allocation paths use distinct update locations within the late stage; they are fixed operating points, not identical per-step interventions.

## I.2 COMPLETE QUALITY OUTCOMES AND UNCERTAINTY

The common suite reports VideoReward VQ/MQ/TA, CLIP, CLAP, and basic audio PQ. The official per-sample suite reports TV-IB, TA-IB, AV-IB, CLIP, CLAP, AVHScore, and JavisScore. Their text fields and scoring protocols differ; common CLAP and official CLAP are not interchangeable. Tables below retain the two suites separately. No subset FVD/FAD is substituted for the full benchmark, and DeSync remains inapplicable at the native Javis duration.

Table 13: All six common-scorer means for all eight configurations, each using 512 prompts and three seeds. All metrics are higher-is-better; CLAP here is $\text{CLAP}_c$ in the main ablation table.

| Configuration | VQ | MQ | TA | CLIP | CLAP | PQ |
|---|---|---|---|---|---|---|
| Native | 1.946566 | 0.432074 | 2.777054 | 0.330666 | 0.313003 | 6.176623 |
| J20 | 1.921078 | 0.422883 | 2.758467 | 0.330516 | 0.308573 | 6.173301 |
| Late J21 | 1.925704 | 0.430147 | 2.751368 | 0.330549 | 0.307419 | 6.169477 |
| Late A3 | 1.919743 | 0.419281 | 2.747204 | 0.330526 | 0.310937 | 6.168760 |
| Late J22 | 1.929483 | 0.426401 | 2.758518 | 0.330524 | 0.311477 | 6.163842 |
| **AsymRefresh (late A6)** | 1.917004 | 0.418038 | 2.752008 | 0.330551 | 0.312199 | 6.162566 |
| Middle A6 | 1.918163 | 0.422214 | 2.748977 | 0.330463 | 0.308498 | 6.178810 |
| Held-context A6 | 1.922250 | 0.424148 | 2.752794 | 0.330474 | 0.312275 | 6.159939 |

**Paired estimation and comparison family.** The four designated contrasts are A3 minus late J21, A6 minus late J22, late minus middle A6, and predicted minus held contexts. For each metric, we first average the three paired seed differences within each prompt, then resample the 512 prompts 10,000 times. Thus the resampling unit is the prompt, not one of the 1,536 prompt–seed pairs. The four common-CLAP contrasts additionally use 98.75% individual Bonferroni intervals ($1 - 0.05/4$). This adjustment applies only to that four-contrast family. The other metrics' 95% intervals are exploratory and unadjusted. These intervals quantify prompt variation conditional on the three evaluated seeds; per-seed differences are reported separately.

Table 14: All seven official per-sample metric means on the same 512-prompt, three-seed ablation cohort. All metrics are higher-is-better. These subset means do not replace the full 10,140-prompt benchmark table.

| Configuration | TV-IB | TA-IB | AV-IB | CLIP | CLAP | AVHScore | JavisScore |
|---|---|---|---|---|---|---|---|
| Native | 0.279230 | 0.159367 | 0.190784 | 0.314819 | 0.440905 | 0.176390 | 0.154416 |
| J20 | 0.278871 | 0.158473 | 0.186765 | 0.314740 | 0.436834 | 0.173136 | 0.151516 |
| Late J21 | 0.278763 | 0.158341 | 0.186479 | 0.314782 | 0.435970 | 0.172915 | 0.151058 |
| Late A3 | 0.278897 | 0.158891 | 0.188701 | 0.314695 | 0.438947 | 0.174816 | 0.152563 |
| Late J22 | 0.278779 | 0.159110 | 0.189253 | 0.314752 | 0.439477 | 0.175155 | 0.153004 |
| **AsymRefresh (late A6)** | 0.278845 | 0.159524 | 0.189892 | 0.314704 | 0.440334 | 0.175750 | 0.153417 |
| Middle A6 | 0.278806 | 0.158354 | 0.187627 | 0.314747 | 0.436787 | 0.173827 | 0.151734 |
| Held-context A6 | 0.278921 | 0.159128 | 0.189402 | 0.314728 | 0.440210 | 0.175384 | 0.153273 |

Table 15: Four designated common-CLAP contrasts. Marginal 95% and individual 98.75% Bonferroni intervals use prompt-cluster bootstrap resampling after averaging seeds within prompt. A zero-crossing interval does not establish equivalence.

| Contrast | $\Delta\text{CLAP}_c$ | 95% interval | 98.75% interval |
|---|---|---|---|
| A3 − late J21 | +0.003518 | $[+0.002796, +0.004235]$ | $[+0.002603, +0.004421]$ |
| A6 − late J22 | +0.000722 | $[+0.000264, +0.001194]$ | $[+0.000133, +0.001316]$ |
| Late − middle A6 | +0.003701 | $[+0.002764, +0.004673]$ | $[+0.002474, +0.004938]$ |
| Predicted − held context | -0.000076 | $[-0.000502, +0.000346]$ | $[-0.000637, +0.000461]$ |

The tighter-budget A3 comparison improves both common and official audio–text alignment, as well as AV-IB, AVHScore, and JavisScore. The smaller A6 comparison has a positive corrected common-CLAP interval, but its seed-42 difference is negative and its official JavisScore interval crosses zero. The gains are specific to the alignment measures and evaluated schedules. In the common suite, A3 minus late J21 changes MQ by $-0.010866$; A6 minus late J22 changes VQ by $-0.012479$ and MQ by $-0.008363$. Late minus middle A6 changes PQ by $-0.016243$. All four exploratory 95% intervals are negative. The predicted-versus-held comparison does not establish a common-CLAP gain. Together, these contrasts locate the supported effects in allocation and timing; an additional benefit from context extrapolation remains inconclusive.

### I.3 Separate common timing and resource measurements

All eight configurations were interleaved on one monitored MetaX C500 using four fixed development prompts, seed 42, and three repetitions, for 96 measured generations after warmup. These measurements form a single separate study; main-benchmark times retain their original source. Generation includes text encoding, denoising, residual prediction, context reconstruction, selective computation, and decoding, excluding model loading, warmup, scoring, and media export. Export-inclusive time is reported separately. The timing cohort does not estimate average latency over the 512-prompt cohort or the full benchmark.

A3 and late J21 differ by 0.0021 seconds in mean generation time (0.012%). A6 is 0.1031 seconds faster than late J22 (0.564%). These are descriptive runtime comparisons, not formal equivalence tests. All four audio configurations retain 2.206 GiB of persistent video contexts and use 19.964 GiB peak allocated memory; joint-only controls use 17.400 GiB, a difference of 2.564 GiB. Held contexts do not remove the memory banks. All rows have 28.523 GiB peak reserved memory. Figure 3 shows paired quality effects from this expanded study. Horizontal bars are prompt-cluster bootstrap intervals: 98.75% for the primary CLAP contrasts and exploratory 95% for VQ/basic PQ. Timing and absolute scores remain in Table 3 and the complete tables above.

### I.4 Earlier exploratory cohort

For provenance, Table 21 retains the earlier 128-prompt, seed-42 budget comparison. Its two-comparison family used 97.5% individual intervals and found a positive A3 effect but an incon-

Table 16: Descriptive common-CLAP differences for each of the three seeds. A3–late J21 and late–middle A6 are positive for every evaluated seed; the smaller A6–late J22 effect varies in sign. No per-seed significance claim is made.

| Contrast | Seed 42 | Seed 43 | Seed 44 |
|---|---|---|---|
| A3 − late J21 | +0.004039 | +0.003273 | +0.003241 |
| A6 − late J22 | -0.000401 | +0.001251 | +0.001315 |
| Late − middle A6 | +0.004527 | +0.003228 | +0.003349 |
| Predicted − held context | -0.000066 | +0.000084 | -0.000247 |

Table 17: All 24 common-scorer paired differences with marginal 95% intervals. These include the four designated CLAP contrasts, whose adjusted intervals appear separately in Table 15.

| Contrast | Metric | Difference | 95% lower | 95% upper |
|---|---|---|---|---|
| A3 − late J21 | VQ | -0.005961 | -0.014048 | +0.001942 |
| | MQ | -0.010866 | -0.016789 | -0.004844 |
| | TA | -0.004164 | -0.019576 | +0.010808 |
| | CLIP | -0.000023 | -0.000131 | +0.000085 |
| | CLAP | +0.003518 | +0.002796 | +0.004235 |
| | PQ | -0.000717 | -0.004944 | +0.003451 |
| A6 − late J22 | VQ | -0.012479 | -0.020370 | -0.004482 |
| | MQ | -0.008363 | -0.014398 | -0.002568 |
| | TA | -0.006510 | -0.021076 | +0.008224 |
| | CLIP | +0.000027 | -0.000085 | +0.000142 |
| | CLAP | +0.000722 | +0.000264 | +0.001194 |
| | PQ | -0.001275 | -0.004281 | +0.001686 |
| Late − middle A6 | VQ | -0.001159 | -0.009300 | +0.007169 |
| | MQ | -0.004176 | -0.010301 | +0.001799 |
| | TA | +0.003031 | -0.011610 | +0.017665 |
| | CLIP | +0.000088 | -0.000035 | +0.000215 |
| | CLAP | +0.003701 | +0.002764 | +0.004673 |
| | PQ | -0.016243 | -0.022345 | -0.010391 |
| Predicted − held context | VQ | -0.005246 | -0.012958 | +0.002090 |
| | MQ | -0.006110 | -0.011636 | -0.000672 |
| | TA | -0.000786 | -0.013809 | +0.012400 |
| | CLIP | +0.000077 | -0.000034 | +0.000190 |
| | CLAP | -0.000076 | -0.000502 | +0.000346 |
| | PQ | +0.002627 | -0.000203 | +0.005387 |

clusive A6 effect. The expanded study includes these prompts with 384 additional prompts and two additional seeds; it does not retroactively change the earlier result or turn that exploratory cohort into independent confirmation. Earlier quality and timing records remain archived separately.

## J Prefix depth and refresh allocation

**Controlled extension.** We extend the JavisDiT++ 512-prompt, three-seed ablation with J20, late A6, and late J22 at prefix depth one. For each refresh schedule, prefix depth is the only changed algorithm setting relative to its depth-two counterpart; weights, predictor, output geometry, prompts, seeds, and evaluation protocols are retained. The three depth-two cells reuse complete frozen results from Appendix I. The extension adds 4,608 outputs and 59,904 scalar scores. Together, the six cells contain 9,216 configuration samples and 119,808 finite scores, with complete coverage for all 13 metrics. Reuse does not create an independent replication, and the shared development cohort retains its previously observed inputs.

Table 18: All 28 official per-sample metric contrasts, completing the 52 reported paired outcomes. Intervals are exploratory marginal 95% prompt-cluster bootstrap intervals, without multiplicity adjustment.

| Contrast | Metric | Difference | 95% lower | 95% upper |
|---|---|---|---|---|
| A3 − late J21 | TV-IB | +0.000135 | -0.000024 | +0.000289 |
| | TA-IB | +0.000550 | -0.000080 | +0.001196 |
| | AV-IB | +0.002222 | +0.001407 | +0.003051 |
| | CLIP | -0.000087 | -0.000212 | +0.000033 |
| | CLAP | +0.002978 | +0.002217 | +0.003746 |
| | AVHScore | +0.001902 | +0.001130 | +0.002694 |
| | JavisScore | +0.001505 | +0.000772 | +0.002259 |
| A6 − late J22 | TV-IB | +0.000066 | -0.000079 | +0.000212 |
| | TA-IB | +0.000414 | +0.000048 | +0.000794 |
| | AV-IB | +0.000639 | +0.000074 | +0.001217 |
| | CLIP | -0.000048 | -0.000170 | +0.000078 |
| | CLAP | +0.000856 | +0.000397 | +0.001330 |
| | AVHScore | +0.000595 | +0.000051 | +0.001160 |
| | JavisScore | +0.000413 | -0.000032 | +0.000867 |
| Late − middle A6 | TV-IB | +0.000039 | -0.000120 | +0.000201 |
| | TA-IB | +0.001170 | +0.000337 | +0.002036 |
| | AV-IB | +0.002265 | +0.001163 | +0.003389 |
| | CLIP | -0.000043 | -0.000176 | +0.000093 |
| | CLAP | +0.003546 | +0.002571 | +0.004557 |
| | AVHScore | +0.001923 | +0.000879 | +0.002996 |
| | JavisScore | +0.001682 | +0.000799 | +0.002611 |
| Predicted − held context | TV-IB | -0.000077 | -0.000227 | +0.000072 |
| | TA-IB | +0.000396 | +0.000028 | +0.000761 |
| | AV-IB | +0.000489 | -0.000034 | +0.001015 |
| | CLIP | -0.000024 | -0.000132 | +0.000083 |
| | CLAP | +0.000123 | -0.000280 | +0.000526 |
| | AVHScore | +0.000366 | -0.000105 | +0.000838 |
| | JavisScore | +0.000144 | -0.000243 | +0.000538 |

Table 19: Common isolated timing for all eight configurations. Each row summarizes 12 generations; SD is the sample standard deviation of generation time. Denoising is included in generation; the last column additionally includes media export.

| Configuration | Generation (s) | SD (s) | Denoising (s) | With export (s) |
|---|---|---|---|---|
| Native | 35.0951 | 0.0444 | 33.3578 | 36.5324 |
| J20 | 17.0388 | 0.1398 | 15.3059 | 18.3807 |
| Late J21 | 17.6524 | 0.0421 | 15.9195 | 19.0376 |
| Late A3 | 17.6545 | 0.0250 | 15.9219 | 18.9717 |
| Late J22 | 18.2650 | 0.0234 | 16.5320 | 19.6235 |
| **AsymRefresh (late A6)** | 18.1619 | 0.0329 | 16.4284 | 19.5987 |
| Middle A6 | 18.1519 | 0.0391 | 16.4189 | 19.5231 |
| Held-context A6 | 18.1098 | 0.0326 | 16.3785 | 19.4915 |

Let $B_p$, $A_p$, and $J_p$ denote J20, late A6, and late J22 at prefix depth $p$. Their schedules are the ones in Appendix I: A6 adds audio steps $\{40, 41, 43, 44, 45, 46\}$ to J20, whereas late J22 adds joint steps $\{40, 45\}$. Changing prefix depth changes computation even with the refresh schedule held fixed. At depth one, J20 and A6 use 630 joint block calls, A6 additionally uses 174 audio block calls, and late J22 uses 688 joint calls. These are different call types, not equal-cost FLOP counts.

**Estimation and primary comparisons.** Table 24 defines all nine contrasts. For each, we average three seed differences within each prompt, then use 10,000 prompt-bootstrap draws. The three pre-

Table 20: Peak GPU memory and actual block calls from the same 96-measurement study. Persistent video-context storage is included in allocated memory, not an additional additive cost. Joint and selective audio block calls have different costs and must not be summed as equivalent FLOPs.

| Configuration | Alloc. (GiB) | Reserv. (GiB) | Context (GiB) | Joint calls | Audio calls |
|---|---|---|---|---|---|
| Native | 17.400 | 28.523 | 0.000 | 1500 | 0 |
| J20 | 17.400 | 28.523 | 0.000 | 660 | 0 |
| Late J21 | 17.400 | 28.523 | 0.000 | 688 | 0 |
| Late A3 | 19.964 | 28.523 | 2.206 | 660 | 84 |
| Late J22 | 17.400 | 28.523 | 0.000 | 716 | 0 |
| **AsymRefresh (late A6)** | 19.964 | 28.523 | 2.206 | 660 | 168 |
| Middle A6 | 19.964 | 28.523 | 2.206 | 660 | 168 |
| Held-context A6 | 19.964 | 28.523 | 2.206 | 660 | 168 |

Table 21: Historical 128-prompt, seed-42 budget results. These are the earlier two-contrast family, with 97.5% individual Bonferroni intervals; they are distinct from the current 512-prompt, three-seed analysis.

| Contrast | ΔCLAP | 95% interval | 97.5% interval |
|---|---|---|---|
| A3 − J21 | +0.002745 | $[+0.000956, +0.004568]$ | $[+0.000665, +0.004815]$ |
| A6 − J22 | -0.000962 | $[-0.002711, +0.000737]$ | $[-0.002932, +0.000985]$ |

declared primary common-CLAP comparisons are $A_1 - J_1$, $A_1 - A_2$, and $(A_1 - J_1) - (A_2 - J_2)$, with individual Bonferroni intervals at $1 - 0.05/3 = 98.3333\%$. All other 114 metric/contrast combinations are exploratory; marginal 95% intervals and seed-wise differences are retained for every contrast. The original four-comparison family in Appendix I keeps its original adjusted intervals. C8/C9 reuse earlier outcomes and are not additional independent evidence. For C9, separate bootstrap draws in the two reports yield slightly different marginal endpoints; Table 15 retains the original primary inference.

The depth-one allocation difference is $+0.000898$, with adjusted interval $[+0.000301, +0.001496]$; all three measured seed differences are positive. The direct A6 prefix difference is $-0.000101$, with interval $[-0.000672, +0.000460]$, and the allocation–prefix interaction is $+0.000176$, with interval $[-0.000596, +0.000914]$. The allocation gain is therefore observed at both tested depths, while these controls do not establish a preferred depth or equivalence of the allocation effects. Table 25 preserves the complete primary family. The common and official means and all 117 paired contrasts follow, without filtering metrics or seeds by effect direction. For depth-one A6 versus late J22, VQ changes by $-0.008659$ and basic PQ by $-0.004481$, with negative exploratory 95% intervals. MQ's interval crosses zero. These results retain the quality trade-offs accompanying the alignment gain.

Table 22: Complete common-suite means for the six prefix/allocation cells. Each uses 512 prompts and seeds 42/43/44. Subscripts denote prefix depth; $B_p$, $A_p$, and $J_p$ are J20, late A6, and late J22. Depth-two cells reuse the original expanded ablation. All metrics are higher-is-better.

| Cell | VQ | MQ | TA | CLIP | CLAP | PQ |
|---|---|---|---|---|---|---|
| $B_1$ | 1.919228 | 0.423106 | 2.747890 | 0.330436 | 0.308426 | 6.175719 |
| $A_1$ | 1.923664 | 0.425485 | 2.748298 | 0.330465 | 0.312098 | 6.161369 |
| $J_1$ | 1.932323 | 0.429429 | 2.740866 | 0.330470 | 0.311200 | 6.165850 |
| $B_2$ | 1.921078 | 0.422883 | 2.758467 | 0.330516 | 0.308573 | 6.173301 |
| $A_2$ | 1.917004 | 0.418038 | 2.752008 | 0.330551 | 0.312199 | 6.162566 |
| $J_2$ | 1.929483 | 0.426401 | 2.758518 | 0.330524 | 0.311477 | 6.163842 |

**Separate timing and memory.** The three depth-one configurations and Native were interleaved on one monitored C500 using four fixed development prompts, seed 42, and three repetitions after warmup. Table 32 contains all 48 observations summarized by configuration. Generation follows

Table 23: Complete official-suite means for the six prefix/allocation cells. Each uses 512 prompts and seeds 42/43/44. Subscripts denote prefix depth; $B_p$, $A_p$, and $J_p$ are J20, late A6, and late J22. Depth-two cells reuse the original expanded ablation. All metrics are higher-is-better.

| Cell | TV-IB | TA-IB | AV-IB | CLIP | CLAP | AVHScore | JavisScore |
|---|---|---|---|---|---|---|---|
| $B_1$ | 0.278976 | 0.158542 | 0.187124 | 0.314654 | 0.436808 | 0.173429 | 0.151672 |
| $A_1$ | 0.278938 | 0.159393 | 0.189690 | 0.314696 | 0.440180 | 0.175719 | 0.153450 |
| $J_1$ | 0.278884 | 0.158891 | 0.189093 | 0.314647 | 0.439425 | 0.175314 | 0.153305 |
| $B_2$ | 0.278871 | 0.158473 | 0.186765 | 0.314740 | 0.436834 | 0.173136 | 0.151516 |
| $A_2$ | 0.278845 | 0.159524 | 0.189892 | 0.314704 | 0.440334 | 0.175750 | 0.153417 |
| $J_2$ | 0.278779 | 0.159110 | 0.189253 | 0.314752 | 0.439477 | 0.175155 | 0.153004 |

Table 24: All nine prefix/allocation contrasts. Only C1–C3 on common CLAP form the new three-comparison primary family; other metric/contrast combinations are exploratory.

| ID | Difference | Question |
|---|---|---|
| C1 | $A_1 - J_1$ | Allocation at prefix one |
| C2 | $A_1 - A_2$ | Prefix change for A6 |
| C3 | $(A_1 - J_1) - (A_2 - J_2)$ | Allocation–prefix interaction |
| C4 | $B_1 - B_2$ | Prefix change for J20 |
| C5 | $J_1 - J_2$ | Prefix change for late J22 |
| C6 | $(A_1 - B_1) - (A_2 - B_2)$ | Audio-increment–prefix interaction |
| C7 | $A_1 - B_1$ | Audio increment at prefix one |
| C8 | $A_2 - B_2$ | Audio increment at prefix two |
| C9 | $A_2 - J_2$ | Reused allocation at prefix two |

the same measurement boundary as Appendix C; export-inclusive time is separate. A6 and late J22 average 17.287251 and 17.427168 seconds, respectively: a descriptive difference of $-0.139917$ seconds and a ratio of 0.991971. No timing equivalence test is claimed. A6 adds 2.680 GiB of peak allocated memory, while peak reserved memory is the same. These measurements do not support a direct causal p=1 versus p=2 speed comparison across timing batches, and do not replace either main table.

## K Selection and provenance of paired examples

The examples in Figure 2 use seed 42 from the already observed 128-prompt confirmation cohort. The existing score ledger requires successful, identity-matched VQ, MQ, TA, CLIP, CLAP, and basic PQ records for all three methods and every prompt. We inspected the recorded scores and three extracted AsymRefresh frames for 16 candidate prompts, then chose three visible sound-source categories for illustration: percussion, birdsong, and rotor noise. Panels (a–c) are respectively `Dk-020qo4YQ`, `nrkLWsK6tzw_scene-0`, and `K18oFrxsrvg`. This editorial selection is post-hoc, not a random sample or an estimate of typical perceptual quality.

Table 33 reports all six scores for these cases and two additional examples. Case C, `0kU56fLQY8E_scene-0`, minimizes the mean absolute distance of AsymRefresh's six empirical midrank percentiles from 0.5; each percentile is the average ascending rank minus one half, divided by 128. Case R, `6zbgi4vJ7pQ_0_10000`, minimizes CLAP(AsymRefresh)−CLAP(Joint22) among negative differences. Both rules use the original fixed hash ordering for ties. R's CLAP decreases by 0.0651 while its other five scores increase. These scores complement the paired media; aggregate allocation results are reported in Section 4.3.

The main figure shows frame 32 (2 s) from each method's own generated MP4 without cropping or enhancement, beside its full-duration spectrum. The extraction archive retains frames 0, 16, 32, 48, and 64. Spectra use the corresponding original 65,000-sample, 16-kHz WAV, without per-clip amplitude normalization or temporal shifts, on a shared $-80$ to 20 dB scale with unit reference power. Per-case scores are reported together in Table 33; the main figure also shows all six AsymRefresh-

Table 25: All three predeclared common-CLAP comparisons, with 98.3333% individual Bonferroni intervals and descriptive seed-wise differences. Prompt bootstrap averages seeds before resampling. Zero-crossing intervals do not establish equivalence.

| Contrast | $\Delta\text{CLAP}_c$ | 98.3333% interval | Seed 42 | Seed 43 | Seed 44 |
|---|---|---|---|---|---|
| C1: $A_1 - J_1$ | +0.000898 | $[+0.000301, +0.001496]$ | +0.001004 | +0.000657 | +0.001032 |
| C2: $A_1 - A_2$ | -0.000101 | $[-0.000672, +0.000460]$ | +0.000561 | -0.000454 | -0.000409 |
| C3: $(A_1 - J_1) - (A_2 - J_2)$ | +0.000176 | $[-0.000596, +0.000914]$ | +0.001406 | -0.000594 | -0.000283 |

Table 26: Complete common-suite results for C1–C3 (Table 24), including all three seed differences. The 95% intervals are marginal and unadjusted; corrected primary intervals appear in Table 25.

| ID | Metric | Difference | 95% interval | Seed 42 | Seed 43 | Seed 44 |
|---|---|---|---|---|---|---|
| C1 | VQ | -0.008659 | $[-0.016481, -0.000820]$ | -0.008236 | -0.011683 | -0.006058 |
| | MQ | -0.003944 | $[-0.009364, +0.001504]$ | -0.000155 | -0.001400 | -0.010278 |
| | TA | +0.007433 | $[-0.006256, +0.021487]$ | +0.011370 | -0.001455 | +0.012382 |
| | CLIP | -0.000004 | $[-0.000121, +0.000110]$ | -0.000042 | +0.000067 | -0.000038 |
| | CLAP | +0.000898 | $[+0.000411, +0.001383]$ | +0.001004 | +0.000657 | +0.001032 |
| | PQ | -0.004481 | $[-0.007630, -0.001406]$ | +0.001139 | -0.006414 | -0.008168 |
| C2 | VQ | +0.006659 | $[-0.002171, +0.015355]$ | +0.008625 | -0.004135 | +0.015488 |
| | MQ | +0.007447 | $[+0.001410, +0.013860]$ | +0.010699 | +0.005515 | +0.006127 |
| | TA | -0.003710 | $[-0.020818, +0.013046]$ | +0.009781 | -0.003653 | -0.017258 |
| | CLIP | -0.000086 | $[-0.000230, +0.000060]$ | +0.000016 | -0.000094 | -0.000179 |
| | CLAP | -0.000101 | $[-0.000567, +0.000346]$ | +0.000561 | -0.000454 | -0.000409 |
| | PQ | -0.001198 | $[-0.004666, +0.002243]$ | -0.000571 | -0.003282 | +0.000260 |
| C3 | VQ | +0.003820 | $[-0.007319, +0.015017]$ | +0.008101 | -0.009013 | +0.012372 |
| | MQ | +0.004419 | $[-0.003451, +0.012345]$ | +0.009240 | +0.004694 | -0.000676 |
| | TA | +0.013943 | $[-0.006513, +0.034040]$ | +0.016472 | +0.007231 | +0.018124 |
| | CLIP | -0.000031 | $[-0.000189, +0.000124]$ | -0.000115 | +0.000156 | -0.000134 |
| | CLAP | +0.000176 | $[-0.000447, +0.000785]$ | +0.001406 | -0.000594 | -0.000283 |
| | PQ | -0.003206 | $[-0.007327, +0.000884]$ | +0.000109 | -0.005339 | -0.004387 |

minus-Joint22 differences. Cell colors encode sign, with each metric retaining its own units. Red boxes mark the CLAP column and one matched spectral window per case across all three methods. We select the window with the largest mean absolute Joint22–AsymRefresh difference in display-clipped dB, scanning one-second windows at 0.5-second offsets from 0 to 3 seconds and 32-bin mel bands at eight-bin offsets from 0 to 32. These boxes locate spectral differences without assigning a quality or synchronization judgment. Extraction and selection manifests retain input hashes, score identities, complete prompts, and display parameters. Prompt excerpts mark omitted spans with ellipses. The accompanying playback collection retains the original MP4 and WAV exports for all five cases. The figure illustrates generated outputs and automated scores; no listening study or temporal-synchronization annotation was performed for this figure.

Table 27: Complete official-suite results for C1–C3 (Table 24), including all three seed differences. The 95% intervals are marginal and unadjusted; corrected primary intervals appear in Table 25.

| ID | Metric | Difference | 95% interval | Seed 42 | Seed 43 | Seed 44 |
|---|---|---|---|---|---|---|
| C1 | TV-IB | +0.000054 | $[-0.000102, +0.000211]$ | -0.000034 | +0.000118 | +0.000076 |
| | TA-IB | +0.000502 | $[+0.000148, +0.000862]$ | -0.000128 | +0.001062 | +0.000572 |
| | AV-IB | +0.000597 | $[+0.000084, +0.001102]$ | +0.000574 | +0.000674 | +0.000542 |
| | CLIP | +0.000049 | $[-0.000062, +0.000162]$ | +0.000016 | +0.000126 | +0.000006 |
| | CLAP | +0.000755 | $[+0.000308, +0.001198]$ | +0.000977 | +0.000452 | +0.000836 |
| | AVHScore | +0.000405 | $[-0.000065, +0.000881]$ | +0.000613 | +0.000462 | +0.000139 |
| | JavisScore | +0.000145 | $[-0.000284, +0.000580]$ | +0.000594 | +0.000036 | -0.000194 |
| C2 | TV-IB | +0.000093 | $[-0.000094, +0.000282]$ | -0.000180 | +0.000079 | +0.000379 |
| | TA-IB | -0.000131 | $[-0.000493, +0.000220]$ | -0.000375 | -0.000297 | +0.000278 |
| | AV-IB | -0.000202 | $[-0.000729, +0.000333]$ | +0.000392 | -0.000810 | -0.000189 |
| | CLIP | -0.000008 | $[-0.000163, +0.000147]$ | +0.000177 | -0.000068 | -0.000134 |
| | CLAP | -0.000153 | $[-0.000617, +0.000294]$ | +0.000357 | -0.000460 | -0.000356 |
| | AVHScore | -0.000031 | $[-0.000531, +0.000470]$ | +0.000844 | -0.000669 | -0.000269 |
| | JavisScore | +0.000034 | $[-0.000389, +0.000467]$ | +0.000411 | -0.000484 | +0.000175 |
| C3 | TV-IB | -0.000012 | $[-0.000214, +0.000193]$ | -0.000173 | +0.000068 | +0.000068 |
| | TA-IB | +0.000088 | $[-0.000368, +0.000547]$ | -0.000358 | +0.000264 | +0.000357 |
| | AV-IB | -0.000042 | $[-0.000722, +0.000629]$ | +0.000520 | -0.000922 | +0.000275 |
| | CLIP | +0.000097 | $[-0.000057, +0.000253]$ | +0.000044 | +0.000233 | +0.000013 |
| | CLAP | -0.000101 | $[-0.000686, +0.000476]$ | +0.000712 | -0.000578 | -0.000438 |
| | AVHScore | -0.000190 | $[-0.000817, +0.000435]$ | +0.000907 | -0.001267 | -0.000210 |
| | JavisScore | -0.000267 | $[-0.000844, +0.000307]$ | +0.000869 | -0.001461 | -0.000211 |

Table 28: Complete common-suite results for C4–C6 (Table 24), including all three seed differences. The 95% intervals are marginal and unadjusted; corrected primary intervals appear in Table 25.

| ID | Metric | Difference | 95% interval | Seed 42 | Seed 43 | Seed 44 |
|---|---|---|---|---|---|---|
| C4 | VQ | -0.001850 | $[-0.010496, +0.006635]$ | -0.001834 | +0.001892 | -0.005608 |
| | MQ | +0.000223 | $[-0.006125, +0.006492]$ | +0.005476 | -0.000401 | -0.004406 |
| | TA | -0.010577 | $[-0.027279, +0.006733]$ | -0.004487 | -0.016384 | -0.010860 |
| | CLIP | -0.000080 | $[-0.000213, +0.000054]$ | -0.000034 | -0.000123 | -0.000083 |
| | CLAP | -0.000146 | $[-0.000672, +0.000379]$ | +0.000209 | -0.000221 | -0.000427 |
| | PQ | +0.002418 | $[-0.001189, +0.006140]$ | -0.000045 | +0.003379 | +0.003920 |
| C5 | VQ | +0.002840 | $[-0.005323, +0.011161]$ | +0.000525 | +0.004878 | +0.003116 |
| | MQ | +0.003028 | $[-0.003085, +0.009246]$ | +0.001460 | +0.000821 | +0.006803 |
| | TA | -0.017652 | $[-0.034966, -0.000352]$ | -0.006691 | -0.010884 | -0.035382 |
| | CLIP | -0.000055 | $[-0.000186, +0.000078]$ | +0.000132 | -0.000251 | -0.000045 |
| | CLAP | -0.000277 | $[-0.000761, +0.000205]$ | -0.000845 | +0.000140 | -0.000126 |
| | PQ | +0.002008 | $[-0.001441, +0.005535]$ | -0.000680 | +0.002056 | +0.004647 |
| C6 | VQ | +0.008510 | $[-0.002144, +0.019289]$ | +0.010460 | -0.006027 | +0.021096 |
| | MQ | +0.007224 | $[-0.000527, +0.015306]$ | +0.005223 | +0.005916 | +0.010533 |
| | TA | +0.006867 | $[-0.012572, +0.026410]$ | +0.014268 | +0.012731 | -0.006397 |
| | CLIP | -0.000006 | $[-0.000164, +0.000149]$ | +0.000050 | +0.000029 | -0.000096 |
| | CLAP | +0.000046 | $[-0.000657, +0.000746]$ | +0.000352 | -0.000232 | +0.000017 |
| | PQ | -0.003616 | $[-0.008285, +0.000962]$ | -0.000526 | -0.006662 | -0.003659 |

Table 29: Complete official-suite results for C4–C6 (Table 24), including all three seed differences. The 95% intervals are marginal and unadjusted; corrected primary intervals appear in Table 25.

| ID | Metric | Difference | 95% interval | Seed 42 | Seed 43 | Seed 44 |
|---|---|---|---|---|---|---|
| C4 | TV-IB | +0.000105 | $[-0.000089, +0.000298]$ | +0.000094 | +0.000096 | +0.000124 |
| | TA-IB | +0.000069 | $[-0.000336, +0.000468]$ | +0.000232 | -0.000344 | +0.000319 |
| | AV-IB | +0.000359 | $[-0.000171, +0.000908]$ | +0.000730 | -0.000271 | +0.000618 |
| | CLIP | -0.000086 | $[-0.000247, +0.000079]$ | -0.000109 | -0.000089 | -0.000059 |
| | CLAP | -0.000026 | $[-0.000553, +0.000514]$ | +0.000282 | -0.000213 | -0.000147 |
| | AVHScore | +0.000293 | $[-0.000204, +0.000819]$ | +0.000762 | -0.000197 | +0.000314 |
| | JavisScore | +0.000156 | $[-0.000293, +0.000622]$ | +0.000571 | -0.000060 | -0.000042 |
| C5 | TV-IB | +0.000105 | $[-0.000070, +0.000279]$ | -0.000007 | +0.000010 | +0.000312 |
| | TA-IB | -0.000219 | $[-0.000601, +0.000183]$ | -0.000017 | -0.000561 | -0.000079 |
| | AV-IB | -0.000160 | $[-0.000698, +0.000378]$ | -0.000127 | +0.000112 | -0.000464 |
| | CLIP | -0.000105 | $[-0.000252, +0.000047]$ | +0.000133 | -0.000301 | -0.000147 |
| | CLAP | -0.000052 | $[-0.000484, +0.000395]$ | -0.000355 | +0.000117 | +0.000082 |
| | AVHScore | +0.000159 | $[-0.000381, +0.000682]$ | -0.000063 | +0.000598 | -0.000060 |
| | JavisScore | +0.000301 | $[-0.000161, +0.000777]$ | -0.000458 | +0.000977 | +0.000386 |
| C6 | TV-IB | -0.000012 | $[-0.000212, +0.000188]$ | -0.000274 | -0.000018 | +0.000255 |
| | TA-IB | -0.000200 | $[-0.000707, +0.000324]$ | -0.000607 | +0.000047 | -0.000041 |
| | AV-IB | -0.000561 | $[-0.001292, +0.000173]$ | -0.000337 | -0.000539 | -0.000807 |
| | CLIP | +0.000077 | $[-0.000073, +0.000229]$ | +0.000286 | +0.000021 | -0.000075 |
| | CLAP | -0.000127 | $[-0.000800, +0.000561]$ | +0.000075 | -0.000247 | -0.000209 |
| | AVHScore | -0.000324 | $[-0.001002, +0.000357]$ | +0.000083 | -0.000472 | -0.000583 |
| | JavisScore | -0.000122 | $[-0.000725, +0.000470]$ | -0.000159 | -0.000424 | +0.000217 |

Table 30: Complete common-suite results for C7–C9 (Table 24), including all three seed differences. The 95% intervals are marginal and unadjusted; corrected primary intervals appear in Table 25.

| ID | Metric | Difference | 95% interval | Seed 42 | Seed 43 | Seed 44 |
|---|---|---|---|---|---|---|
| C7 | VQ | +0.004436 | $[-0.003158, +0.011926]$ | +0.006302 | -0.003733 | +0.010738 |
| | MQ | +0.002379 | $[-0.003041, +0.007767]$ | +0.000080 | +0.004076 | +0.002981 |
| | TA | +0.000408 | $[-0.013063, +0.014001]$ | +0.005367 | -0.005800 | +0.001658 |
| | CLIP | +0.000029 | $[-0.000088, +0.000145]$ | +0.000026 | +0.000031 | +0.000031 |
| | CLAP | +0.003672 | $[+0.002823, +0.004556]$ | +0.003992 | +0.003082 | +0.003942 |
| | PQ | -0.014350 | $[-0.019184, -0.009528]$ | -0.012682 | -0.015461 | -0.014908 |
| C8 | VQ | -0.004074 | $[-0.011781, +0.003562]$ | -0.004158 | +0.002294 | -0.010358 |
| | MQ | -0.004845 | $[-0.010456, +0.000729]$ | -0.005143 | -0.001839 | -0.007552 |
| | TA | -0.006459 | $[-0.020409, +0.007933]$ | -0.008902 | -0.018531 | +0.008055 |
| | CLIP | +0.000035 | $[-0.000068, +0.000141]$ | -0.000024 | +0.000002 | +0.000127 |
| | CLAP | +0.003626 | $[+0.002726, +0.004552]$ | +0.003640 | +0.003315 | +0.003924 |
| | PQ | -0.010735 | $[-0.015605, -0.005918]$ | -0.012156 | -0.008799 | -0.011249 |
| C9 | VQ | -0.012479 | $[-0.020531, -0.004387]$ | -0.016337 | -0.002670 | -0.018430 |
| | MQ | -0.008363 | $[-0.014308, -0.002500]$ | -0.009395 | -0.006093 | -0.009601 |
| | TA | -0.006510 | $[-0.020988, +0.007965]$ | -0.005102 | -0.008686 | -0.005742 |
| | CLIP | +0.000027 | $[-0.000090, +0.000142]$ | +0.000073 | -0.000090 | +0.000096 |
| | CLAP | +0.000722 | $[+0.000252, +0.001201]$ | -0.000401 | +0.001251 | +0.001315 |
| | PQ | -0.001275 | $[-0.004192, +0.001685]$ | +0.001030 | -0.001075 | -0.003781 |

Table 31: Complete official-suite results for C7–C9 (Table 24), including all three seed differences. The 95% intervals are marginal and unadjusted; corrected primary intervals appear in Table 25.

| ID | Metric | Difference | 95% interval | Seed 42 | Seed 43 | Seed 44 |
|---|---|---|---|---|---|---|
| C7 | TV-IB | -0.000038 | $[-0.000179, +0.000105]$ | -0.000155 | +0.000019 | +0.000021 |
| | TA-IB | +0.000851 | $[+0.000095, +0.001627]$ | +0.000820 | +0.001073 | +0.000659 |
| | AV-IB | +0.002566 | $[+0.001576, +0.003595]$ | +0.002599 | +0.003155 | +0.001944 |
| | CLIP | +0.000042 | $[-0.000067, +0.000148]$ | +0.000131 | +0.000039 | -0.000044 |
| | CLAP | +0.003372 | $[+0.002504, +0.004280]$ | +0.003427 | +0.003275 | +0.003414 |
| | AVHScore | +0.002290 | $[+0.001356, +0.003261]$ | +0.002583 | +0.002851 | +0.001435 |
| | JavisScore | +0.001779 | $[+0.000986, +0.002584]$ | +0.001989 | +0.001861 | +0.001486 |
| C8 | TV-IB | -0.000026 | $[-0.000169, +0.000119]$ | +0.000119 | +0.000036 | -0.000234 |
| | TA-IB | +0.001051 | $[+0.000279, +0.001842]$ | +0.001427 | +0.001026 | +0.000700 |
| | AV-IB | +0.003127 | $[+0.002115, +0.004177]$ | +0.002936 | +0.003693 | +0.002751 |
| | CLIP | -0.000035 | $[-0.000140, +0.000065]$ | -0.000155 | +0.000018 | +0.000031 |
| | CLAP | +0.003500 | $[+0.002589, +0.004402]$ | +0.003353 | +0.003523 | +0.003624 |
| | AVHScore | +0.002614 | $[+0.001649, +0.003598]$ | +0.002500 | +0.003322 | +0.002018 |
| | JavisScore | +0.001901 | $[+0.001095, +0.002741]$ | +0.002148 | +0.002285 | +0.001270 |
| C9 | TV-IB | +0.000066 | $[-0.000084, +0.000215]$ | +0.000140 | +0.000050 | +0.000009 |
| | TA-IB | +0.000414 | $[+0.000041, +0.000796]$ | +0.000230 | +0.000798 | +0.000215 |
| | AV-IB | +0.000639 | $[+0.000081, +0.001212]$ | +0.000055 | +0.001596 | +0.000267 |
| | CLIP | -0.000048 | $[-0.000172, +0.000076]$ | -0.000028 | -0.000108 | -0.000008 |
| | CLAP | +0.000856 | $[+0.000386, +0.001328]$ | +0.000266 | +0.001029 | +0.001274 |
| | AVHScore | +0.000595 | $[+0.000070, +0.001132]$ | -0.000294 | +0.001729 | +0.000348 |
| | JavisScore | +0.000413 | $[-0.000020, +0.000847]$ | -0.000276 | +0.001497 | +0.000017 |

Table 32: Separate isolated prefix-one timing on one monitored C500, with four development prompts, seed 42, and three interleaved repetitions (12 measurements per row). SD is the sample standard deviation. Memory is in GiB. Native is a local timing reference, not an additional quality cell or a new main-table speed measurement.

| Cell | Gen. (s) | SD (s) | Denoise (s) | With export (s) | Allocated | Reserved |
|---|---|---|---|---|---|---|
| Native | 34.6435 | 0.0330 | 32.9449 | 36.0541 | 17.400 | 28.523 |
| $B_1$ | 16.1467 | 0.0270 | 14.4493 | 17.5604 | 17.400 | 28.523 |
| $A_1$ | 17.2873 | 0.0466 | 15.5898 | 18.6805 | 20.080 | 28.523 |
| $J_1$ | 17.4272 | 0.0139 | 15.7295 | 18.8168 | 17.400 | 28.523 |

Table 33: Complete per-sample scores for the three illustrated cases and the two additional cases C and R defined above. Cases (a–c) correspond to percussion, birdsong, and rotor noise in Figure 2. All six metrics are higher-is-better and use the common-suite scorers on the confirmation cohort.

| Case | Method | VQ | MQ | TA | CLIP | CLAP | PQ |
|---|---|---|---|---|---|---|---|
| (a) | Native | 1.2422 | -0.1865 | 3.5312 | 0.3398 | 0.3490 | 7.1413 |
| | Joint22 | 1.0469 | -0.2041 | 2.9688 | 0.3293 | 0.3430 | 7.2334 |
| | AsymRefresh | 1.0312 | -0.2158 | 3.2188 | 0.3293 | 0.3570 | 7.2609 |
| (b) | Native | 4.8125 | 1.4688 | 4.1250 | 0.3042 | 0.4845 | 7.4756 |
| | Joint22 | 4.8750 | 1.3594 | 4.1562 | 0.3042 | 0.4610 | 7.4599 |
| | AsymRefresh | 5.2188 | 1.1250 | 4.2500 | 0.3037 | 0.4851 | 7.4559 |
| (c) | Native | 1.0469 | -0.1104 | 2.1250 | 0.3738 | 0.1673 | 5.5364 |
| | Joint22 | 0.7539 | -0.3750 | 4.1875 | 0.3704 | 0.1997 | 5.5563 |
| | AsymRefresh | 0.8164 | -0.3594 | 4.0312 | 0.3711 | 0.1943 | 5.4232 |
| C | Native | 1.3672 | -0.0005 | 4.4688 | 0.3315 | 0.3473 | 6.2836 |
| | Joint22 | 1.4375 | 0.0981 | 4.1250 | 0.3306 | 0.3371 | 6.2746 |
| | AsymRefresh | 1.6016 | 0.1523 | 4.1250 | 0.3303 | 0.3370 | 6.3352 |
| R | Native | 1.2656 | -0.1953 | -0.7852 | 0.3044 | 0.3039 | 4.5236 |
| | Joint22 | 1.2891 | -0.0078 | -0.4219 | 0.3076 | 0.3950 | 4.5388 |
| | AsymRefresh | 1.4062 | 0.0713 | 0.3652 | 0.3096 | 0.3300 | 4.5403 |