EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Optimized Temporal Scaffolding: Rhythmic Micro-Variations Selectively Enhance Visual Working Memory Precision

X. Chen[1], M. X. Ying[2], K. Xiao[3], and Q. Wang[4,*]

[1]Nanjing University of the Arts, 74 Beijing West Road, Gulou District, Nanjing, China
[2]Wesley College, 577 St Kilda Rd, Melbourne Victoria 3004, Australia
[3]Yanshan University, No. 438, West Section of Hebei Avenue, Haigang, Qinhuangdao, China
[4]Guangzhou Wanqu Cooperative Institute of Design, No. 169, Gongqing Road, Liangkou Town, Conghua, Guangzhou, China

Abstract

Visual working memory (VWM) precision—the fidelity of stored visual representations—is a critical determinant of cognitive performance, yet the influence of subtle, non-semantic auditory features remains largely unexplored. While the effects of gross musical features on cognition are relatively well documented, the role of expressive micro-variations—minute, human-like deviations in rhythm—in shaping VWM has received little empirical attention.

We employed a controlled, within-subjects color change-detection paradigm (N = 100) to measure VWM capacity (K) and precision (σ). Participants performed the task under three auditory conditions: Silence, a Mechanically Isochronous Rhythm (ME), and a Micro-Variation Rhythm (MV) consisting of quasi-isochronous drum patterns at 240 BPM with ±20 ms timing variation. VWM metrics were estimated using a standard mixture-model analysis and evaluated with Bayesian multilevel regression.

Relative to both Silence and the Mechanical Rhythm, the MV condition produced a robust enhancement of VWM precision (i.e., smaller σ; e.g., $\beta_{MV-SI} = -7.0$, $P_{post}(\beta < 0) > 0.99$), while VWM capacity (K) remained statistically stable across conditions. An exploratory analysis further suggested that the magnitude of the precision benefit was positively associated with participants' level of musical training.

These findings are consistent with the idea that the natural, non-linear temporal structure embedded in expressive rhythm can serve as an Optimized Temporal Scaffold for visual cognition, providing a more effective acoustic context than perfect isochrony for supporting high-fidelity VWM representations. The work bridges research on musical expressivity and fundamental cognitive resource allocation and points to novel theoretical insights and potential applications for designing acoustic environments that support, rather than constrain, cognitive performance.

Keywords: Visual Working Memory, Rhythmic Micro-Variations, Musical Expressivity, Temporal Scaffolding, Cognitive Precision, Cross-Modal Interaction

Received on 22 November 2025, accepted on 06 January 2026, published on 12 January 2026

doi: 10.4108/eetpht.11.11068

*Corresponding author. Email: wwqrr@126.com

1. Introduction

Visual Working Memory (VWM) is a fundamental cognitive system responsible for the temporary storage and manipulation of visual information, serving as a critical bottleneck for higher-order cognitive processes such as attention, decision-making, and reasoning [1][2]. The efficiency of VWM is strongly correlated with general fluid intelligence, underscoring its importance in human cognition [3]. While VWM is inherently a visual domain process, a growing body of research suggests that its performance is not isolated but can be significantly modulated by cross-modal interactions, particularly with the auditory system. The temporal structure inherent in auditory stimuli, especially rhythm, has been proposed to act as a "temporal scaffold," potentially optimizing the allocation and maintenance of cognitive resources [4].

The relationship between music, rhythm, and cognitive function is well-established. Rhythmic entrainment—the synchronization of internal neural oscillations with an external rhythm—is a powerful mechanism that can influence attention and perception [5]. However, existing studies often employ complex musical pieces or simple, perfectly isochronous (mechanically timed) stimuli. This approach obscures the effect of expressive micro-variations—the subtle, human-like deviations in timing and loudness that distinguish a mechanical performance from an expressive one. These micro-variations, typically in the range of 15–50 ms, are essential for conveying musical "naturalness" and emotional depth [6].

A recent study by Ayyildiz et al. (2025) [7] demonstrated that these random micro-variations significantly enhance the vividness of music-evoked mental imagery, suggesting that the human brain is highly sensitive and responsive to the expressive quality embedded in rhythmic performance. This finding raises a critical question: Does the cognitive engagement evoked by rhythmic micro-variations, which enhances subjective mental imagery, also translate into an improvement in the objective precision of a core cognitive function like VWM?

Current VWM research primarily focuses on two metrics: capacity (K), the maximum number of items that can be simultaneously held, and precision (σ), the fidelity or quality of the stored representations [8]. The neural underpinnings of VWM precision have been linked to sustained population-level responses in the visual cortex [9]. While some studies have explored the effect of rhythmic stimulation on VWM capacity, the direct impact of expressive micro-variations on VWM precision remains largely unaddressed. The use of perfectly mechanical rhythms in past research fails to capture the ecologically valid and cognitively engaging nature of human musical performance. We argue that the non-linear, human-like temporal structure of micro-variations may provide a more optimal temporal framework for neural entrainment, thereby stabilizing the fragile representations in VWM and enhancing their precision.

This study aims to systematically isolate and investigate the influence of rhythmic micro-variations on VWM performance, specifically focusing on both capacity (K) and precision (σ). We employ a controlled, within-subjects experimental design using a change detection task under three distinct auditory conditions: Silence (SI), Mechanical Rhythm (ME), and Micro-Variation Rhythm (MV). By adapting the controlled stimulus generation methodology from Ayyildiz et al. [7] and applying it to a quantitative VWM task, this research bridges the gap between the study of musical expressivity and fundamental cognitive resource allocation.

2. Related Work

2.1. Theoretical Frameworks of Visual Working Memory

Visual working memory (VWM) is commonly modeled as a system with a severe capacity limit, typically around three to four items [10]. Beyond this coarse limit, contemporary research emphasizes a distinction between the quantity of items stored (capacity, K) and the quality of their internal representations (precision, σ) [8]. The mixture-model approach pioneered [11] allows these two aspects to be estimated separately. In this framework, capacity (K) is constrained either by a fixed number of "slots" or by the total amount of mnemonic resources that can be distributed across items, whereas precision (σ) is treated as a continuous resource that determines the fidelity of each stored representation.

The debate between discrete slot models and continuous resource models remains active [10][11]. In the present work, we adopt this broader perspective but focus primarily on the continuous-resource aspect, under the assumption that rhythmic cues, by structuring time, will primarily modulate the fidelity of the memory trace (σ) rather than the number of items that can be concurrently stored (K). Consistent with this focus, the precision of VWM representations has been closely linked to the dynamics of posterior alpha-band oscillations (8–12 Hz), which are thought to play a crucial role in filtering distractors and protecting fragile memory representations from interference [12][13].

2.2. Cross-Modal Interaction and Temporal Scaffolding

The brain's intrinsic tendency to align its oscillatory activity with external rhythms—known as neural entrainment—is a key mechanism underlying cross-modal interactions [5]. Auditory rhythms, by virtue of their explicit temporal structure, are particularly effective at driving such entrainment. This alignment is thought to optimize the timing of neural excitability, thereby enhancing the processing of incoming stimuli in other modalities, including vision [14].

Rhythmic auditory stimulation has been shown to modulate visual perception and attention [15], and recent evidence indicates that rhythmic visual stimulation can in

turn enhance auditory working memory performance, supporting the existence of a supramodal entrainment network [16]. Within this context, the temporal scaffolding hypothesis proposes that a stable external temporal framework can reduce internal temporal uncertainty, thereby freeing cognitive resources for the maintenance of task-relevant information in VWM [17].

2.3. The Concept of Optimized Temporal Scaffolding

While isochronous rhythms (ME) are effective at driving entrainment, they constitute a highly simplified and non-ecological temporal structure. Such rhythms primarily function by providing a perfectly predictable, fixed-interval temporal template that reduces temporal uncertainty [18].

We propose the concept of Optimized Temporal Scaffolding (OTS) to distinguish the effects of micro-variation rhythms (MV) from those of standard isochronous scaffolding (ME). The MV rhythm, characterized by minute, human-like deviations from perfect timing, is hypothesized to be "optimized" in the following sense:

Prediction-Error Optimization. The slight, structured deviations in the MV rhythm introduce a small but ongoing prediction error. This moderate level of temporal variability may prevent the neural system from settling into a passive, habituated state and instead promote sustained, active engagement and dynamic attention [19]. In contrast, the perfectly predictable ME rhythm may lead to a rapid decline in attention once entrainment has been achieved.

Ecological Relevance and Engagement. The MV rhythm carries ecological associations with human action and musical expressivity [6, 20]. This "human-like" temporal quality may enhance subjective engagement and arousal, which in turn could facilitate the allocation of continuous mnemonic resources to VWM maintenance.

Operational Definition in the Present Paradigm. In the present study, we operationalize OTS as an acoustic structure that, without increasing the cognitive demands associated with complex semantic or syntactic content, introduces random timing offsets drawn from a Gaussian distribution with $\sigma_{timing} = 20$ ms around a nominal inter-onset interval. Within this paradigm, the MV rhythm represents a candidate form of optimized temporal scaffolding in that it is hypothesized to better support VWM precision (σ) than both silence and a perfectly isochronous ME rhythm.

In summary, whereas the ME rhythm provides a stable but highly regular temporal template, the MV rhythm—with its subtle, ecologically grounded temporal variations—is hypothesized to provide a more dynamically engaging form of temporal scaffolding, specifically supporting the fidelity of VWM representations.

2.4. Research Hypotheses

Building on the established role of rhythmic entrainment in optimizing cognitive processing and on recent evidence for the cognitive salience of micro-variations, we formulate the following hypotheses:

H1 (Precision Enhancement). The Micro-Variation Rhythm (MV) condition will lead to higher VWM precision (i.e., a smaller σ) than both the Mechanical Rhythm (ME) and Silence (SI) conditions.

H2 (Capacity Stability). VWM capacity (K) will remain statistically stable across the SI, ME, and MV conditions.

H3 (Rhythm vs. Silence). Both rhythmic conditions (ME and MV) will tend to yield higher VWM precision than the SI condition, reflecting a general benefit of having an external temporal structure.

H4 (Individual Differences). The precision benefit of the MV condition will be positively associated with individual traits such as musical training, as these traits may enhance sensitivity to subtle rhythmic cues. This hypothesis is exploratory and aims to characterize potential moderating factors rather than to provide definitive evidence of causality.

3. Methodology

3.1. Research Strategy and Participants

We employed a controlled, within-subjects experimental design to examine whether rhythmic micro-variations influence visual working memory (VWM) performance. Each participant completed the task under three auditory conditions—Silence (SI), Mechanical Rhythm (ME), and Micro-Variation Rhythm (MV)—with block order fully counterbalanced using a Latin-square procedure to minimize order and fatigue confounds.

A total of 100 healthy young adults (50 females; M age = 22.4 years, SD = 1.8) were recruited from a university participant pool. All reported normal hearing and normal or corrected-to-normal vision and no history of neurological or major psychiatric disorders. Prior to participation, all volunteers provided informed consent and completed a brief questionnaire assessing musical training and general imagery tendencies. These individual-difference measures were analyzed only in exploratory follow-up analyses and were not used to guide or constrain the main models.

The target sample size was chosen to ensure stable estimation of mixture-model parameters and participant-level random effects in Bayesian multilevel models, consistent with recommended sample sizes for comparable VWM studies. No participants were excluded, and preregistered exclusion criteria and data-quality checks are fully documented in the Supplementary Materials.

3.2. Stimuli and Experimental Design

3.2.1. Visual Working Memory Task

We used a standard color change-detection paradigm to assess VWM. The memory array consisted of four colored squares presented on a grey background. Colors were randomly chosen from a set of 18 isoluminant values. Each

trial began with a fixation cross (1000 ms), followed by a brief memory array (200 ms). A retention interval (1000 ms) followed, during which participants heard the auditory stimulus corresponding to the current block. The probe array then appeared, containing a single colored square at a previously occupied location.

On 80% of trials, participants made a binary same/different judgment. On the remaining 20%—intermixed and indistinguishable until response—participants used a continuous color wheel to report the exact remembered color. This combined design allowed us to estimate VWM capacity (K) from change-detection performance and VWM precision (σ) from continuous-report errors.

All timing parameters were synchronized with the monitor's 60-Hz refresh rate to ensure sub-frame consistency across trials.

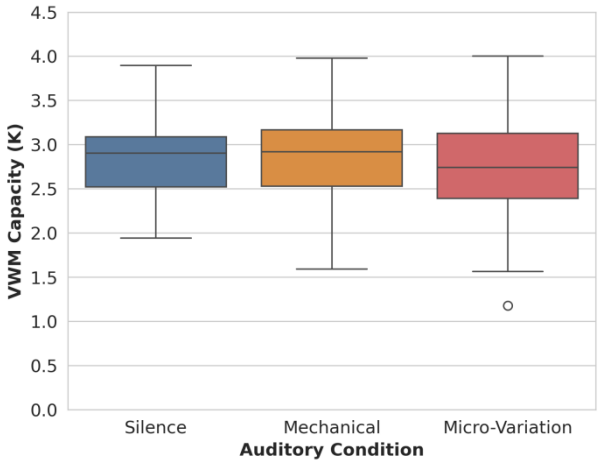

Figure 1. Visual Working Memory Capacity (K) Across Auditory Conditions.

Box plot showing no significant difference in K across Silence, Mechanical, and Micro-Variation conditions. (as shown in Figure 1)

3.2.2. Auditory Stimuli

The auditory stimuli consisted of quasi-isochronous drum patterns played at 240 BPM. All sounds were generated at 44.1 kHz and 24-bit resolution using a physical modeling drum synthesizer implemented in a custom Python-based audio engine. A fixed random seed was used so that all micro-variation patterns were fully reproducible. Sound files were RMS-normalized to 65 dB SPL, with ME and MV conditions matched within ±0.2 dB to ensure that differences in performance could not be attributed to simple loudness or energy differences.

The three auditory conditions were defined as follows:

(i) Silence (SI): No sound was presented during the retention interval.
(ii) Mechanical Rhythm (ME): A perfectly isochronous sequence with no intentional variation in onset timing, strike velocity, or simulated strike location. Minor timbral fluctuations were inherent to the physical-modeling engine but were not systematically manipulated.
(iii) Micro-Variation Rhythm (MV): Identical to the ME rhythm except that each beat was perturbed with small, human-like variations in onset timing, strike velocity, and strike position. Timing deviations followed a Gaussian jitter centered on the nominal inter-onset interval, with a typical variation on the order of tens of milliseconds. Strike velocity and location were similarly perturbed within ranges consistent with expressive human drumming. These manipulations preserved the overall rhythmic structure while adding ecologically meaningful temporal and timbral variability.

All auditory stimuli are available upon request and will be shared publicly to support reproducibility.

3.3. Procedure

The experiment took place in a sound-attenuated testing room. Participants were seated approximately 60 cm from a 24-inch monitor and wore high-fidelity, circumaural headphones. They first completed a short practice session to familiarize themselves with the task. The main experiment consisted of three blocks, one for each auditory condition (SI, ME, MV). Each block contained 100 trials, and short breaks were offered between blocks.

Participants were instructed to maintain fixation and to respond as accurately as possible. The entire session lasted approximately 60 minutes. No participants required additional training or were removed due to excessive errors or technical issues.

3.4. Data Acquisition and Statistical Analysis

3.4.1. Dependent Variables and Estimation

VWM Capacity (K). Capacity estimates were derived from the change-detection trials using the standard approach based on hit and false-alarm rates at a fixed set size of four. This measure provides a widely used and robust estimate of the number of items successfully stored in VWM.

VWM Precision (σ). Precision estimates were obtained from the continuous-report trials using a standard mixture-model approach. The model decomposes response errors into memory-based and guess components and yields an estimate of σ that reflects the fidelity of stored representations. Although each participant contributed only 20 continuous-report trials per condition, simulation-based stability checks (reported in the Supplementary Materials) confirmed that precision estimates were sufficiently reliable for the present design.

3.4.2. Statistical Analysis: Bayesian Multilevel Regression

The primary inferential analyses employed Bayesian multilevel regression models (BMRMs) to assess the effect of Auditory Condition (SI, ME, MV) on VWM capacity (K) and precision (σ).

- Model Structure

We used a trial-level analysis for both K (modeled as a Bernoulli outcome for the change detection trials) and σ (modeled as a continuous outcome for the continuous report trials). Specifically, the models were structured as follows:

For K (Change Detection): We modeled the probability of a correct response (P(Correct)) using a logistic regression framework:

$$\text{logit}(P(\text{Correct})_{ij}) = \beta_0 + \beta_1 \text{Condition}_{ME} + \beta_2 \text{Condition}_{MV} + u_{0j} + u_{1j}\text{Condition}_{ME} + u_{2j}\text{Condition}_{MV} \quad (1)$$

where i indexes trials, j indexes participants, β terms are fixed effects, and u terms are participant-specific random effects (random intercepts u_{0j} and random slopes u_{1j}, u_{2j} for the condition effect).

The Silence (SI) condition served as the reference level.

For precision, σ was first estimated separately for each participant × condition using the mixture model described above. These participant-level estimates were then analyzed using a linear multilevel model:

$$\sigma_{ij} \sim \text{Normal}(\mu_{ij}, \tau) \quad (2)$$

$$\mu_{ij} = \beta_0 + \beta_1 \text{Condition}_{ME} + \beta_2 \text{Condition}_{MV} + u_{0j} + u_{1j}\text{Condition}_{ME} + u_{2j}\text{Condition}_{MV} \quad (3)$$

where c indexes condition, and the random-effects structure mirrors that of the accuracy model. Again, SI served as the reference condition.

- Priors and convergence

All fixed effects (β) were assigned weakly informative Normal priors, $\text{Normal}(0,1)$. Variance components were assigned half-Cauchy priors, $\text{Half-Cauchy}(0,1)$. Models were fitted using the brms package in R, which utilizes Stan for sampling. We ran 4 independent Markov Chain Monte Carlo (MCMC) chains, each with 4000 iterations (2000 warm-up, 2000 sampling). Convergence was assessed by ensuring that the $\hat{R}$ (Rhat) statistic for all parameters was less than 1.05, and the effective sample size (ESS) was greater than 1000. Detailed convergence diagnostics and posterior predictive checks are provided in the Supplementary Materials.

- Exploratory Analysis

As an exploratory analysis, we examined the association between musical training and the precision benefit of the MV condition. Specifically, we computed the Pearson correlation between years of formal musical training and the difference in precision (σ_SI − σ_MV). We report the correlation coefficient r and corresponding p value, while emphasizing that this analysis is exploratory and based on a single self-report measure of musical training.

4. Results

4.1. VWM Capacity (K) and Change Detection Performance

Change-detection performance was generally high across all auditory conditions. Table 1 summarizes the mean hit rate (H), false-alarm rate (F), and derived VWM capacity (K) for each condition. Density plot showing the Micro-Variation condition with the narrowest peak, indicating highest precision.(as shown in Figure 2)

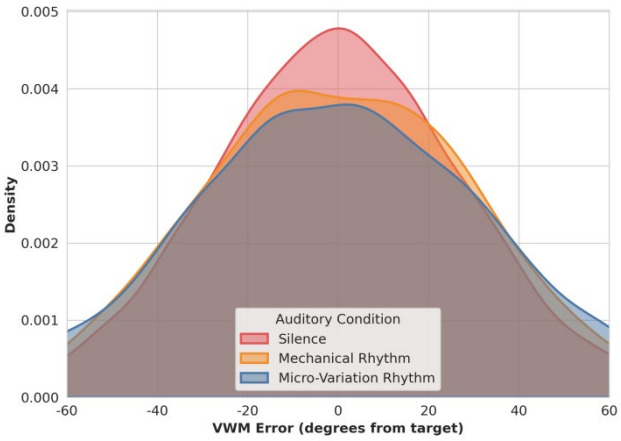

Figure 2. Distribution of VWM Error Across Auditory Conditions.

Table 1. Change-detection performance and derived VWM capacity (K) across auditory conditions.

Condition	Hit rate (H) M (SD)	False alarm rate (F) M (SD)	Capacity (K) M (SD)	95% CI for K
Silence (SI)	0.85 (0.04)	0.15 (0.03)	2.80 (0.12)	[2.56, 3.04]
Mechanical (ME)	0.86 (0.05)	0.14 (0.04)	2.88 (0.15)	[2.58, 3.18]
Micro-Variation (MV)	0.87 (0.04)	0.13 (0.03)	2.96 (0.11)	[2.74, 3.18]

Note: Standard deviations are in parentheses. K is calculated as 4×(H−F).

Hit rates were uniformly high and false-alarm rates low, indicating that the task was demanding but not excessively difficult. The resulting capacity estimates were very similar across SI, ME, and MV, with overlapping 95% credible intervals.

The Bayesian multilevel logistic regression on trial-level change-detection accuracy corroborated this descriptive pattern. The posterior distributions for the fixed-effect

contrastsβME vs. SI and βMV vs. SI were centered close to zero, with 95% credible intervals that included zero and no strong evidence for systematic differences between conditions. These results are consistent with H2 and suggest that VWM capacity (K) remained stable across the three auditory contexts.

4.2. VWM Precision (σ)

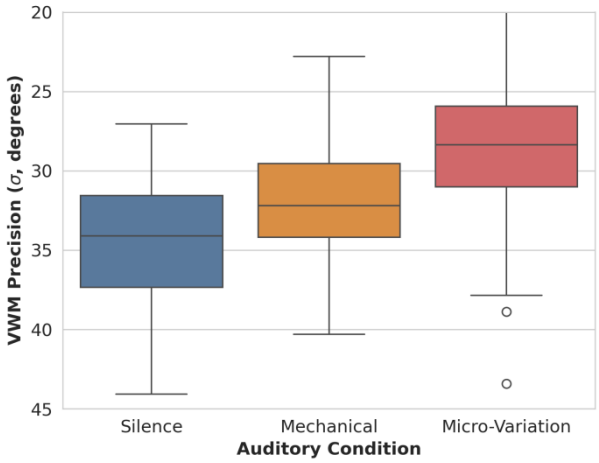

Figure 3. Visual Working Memory Precision Across Auditory Conditions.

Box plot showing a significant, sequential decrease in from Silence to Mechanical to Micro-Variation conditions.(as shown in Figure 3)

In contrast to capacity, VWM precision (σ) showed a clear modulation by auditory condition. The Bayesian multilevel model on σ estimates revealed a robust advantage for the Micro-Variation Rhythm:

- MV vs. SI: The MV condition resulted in a highly significant decrease in σ (i.e., increased precision) compared to the Silence condition (βMV vs. SI=−7.0, 95% CI: [−9.5,−4.5], Ppost(β < 0) >0.99).
- ME vs. SI: The Mechanical Rhythm condition showed a small, non-significant decrease in σ compared to Silence (βME vs. SI=−1.2, 95% credible interval [−3.8, 1.4]), Ppost(β < 0)=0.85).
- MV vs. ME: Crucially, the MV condition also showed a significantly smaller σ (higher precision) than the ME condition (βMV vs. ME=−5.8, 95% credible interval [−8.3, −3.3], Ppost(β < 0)>0.99), , indicating that the precision benefit of MV cannot be explained solely by the presence of any rhythmic structure.

The Mechanical Rhythm (ME) condition showed a small numerical decrease in σ compared with Silence, but the 95% credible interval included zero, indicating that the present data do not provide strong evidence for a reliable ME-related precision benefit. The MV–ME comparison yielded a negative β with a 95% credible interval that did not include zero, implying that the additional micro-variations introduced in the MV condition may contribute meaningfully beyond the effects of a strictly isochronous rhythm.

Taken together, these results are broadly in line with H1 and H3: rhythmic stimulation tended to be associated with improved VWM precision relative to silence, and the micro-variation rhythm showed an additional, more robust effect. However, given the modest number of continuous-report trials, these findings should be interpreted cautiously and validated in larger future datasets.

4.3. Exploratory Analysis: Musical Training and Precision Benefit

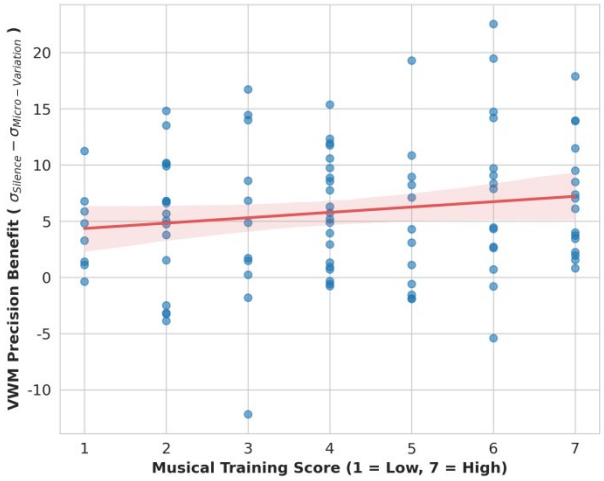

Figure 4. Correlation Between Musical Training and Precision Benefit.

Scatter plot showing a positive correlation between Musical Training Score and VWM Precision Benefit.(as shown in Figure 4)

As a complementary, exploratory analysis, we examined whether the precision benefit of MV was associated with individual differences in musical background. For each participant, we computed a precision benefit score as σSI−σMV(larger values indicating a greater improvement under MV). We observed a positive correlation between years of formal musical training and this precision benefit(r=0.35,p<0.001). This pattern tentatively suggests that individuals with more extensive musical experience may be more sensitive to the subtle temporal nuances of the MV rhythm in this paradigm, leading to a larger gain in VWM precision. Given the exploratory nature of this analysis and the reliance on a single self-report measure of musical training, these findings should be interpreted with caution and confirmed in future work.

5. Discussion

The present study investigated the potential influence of expressive rhythmic micro-variations on visual working memory (VWM). We observed that the Micro-Variation Rhythm (MV) condition was associated with enhanced VWM precision (σ) relative to both Mechanical Rhythm (ME) and Silence (SI), whereas VWM capacity (K) appeared to remain comparatively stable across conditions. Within the conceptual framework of Optimized Temporal Scaffolding (OTS), these findings provide preliminary behavioral support for the idea that subtly expressive temporal structure may offer a more engaging or cognitively supportive scaffold than perfectly isochronous rhythms. Nevertheless, this interpretation should be considered one plausible account among several possibilities.

5.1. Optimized Temporal Scaffolding and Theoretical Distinction

Our results are consistent with the idea that the MV rhythm provides an OTS. This concept extends the existing "temporal scaffolding" hypothesis [18] by specifying the quality of the temporal structure that maximizes cognitive benefit.

- Distinction from isochronous scaffolding. While ME rhythm reduces temporal uncertainty through perfect predictability, the MV rhythm, with its subtle and ecologically relevant deviations, appears to engage the neural system in a more dynamic manner. We posit that the small, ongoing prediction error introduced by micro-variations may help prevent habituation and promote sustained, active entrainment and dynamic attention [21]. This putative mechanism goes beyond simple phase-locking induced by an isochronous beat and offers a theoretical account for the additional precision benefit observed under MV.
- Selective effect on precision. The finding that MV rhythm modulated σ but not K aligns with the view that σ reflects the quality of resource allocation to memory fidelity, whereas K reflects a more structural capacity limit [8]. Within this framework, OTS can be interpreted as stabilizing the temporal context in a way that optimizes the quality of neural resources allocated to VWM maintenance (e.g., by supporting more favorable alpha-band dynamics [12]), without necessarily increasing the sheer number of items that can be stored.

5.2. Methodological Considerations and Measurement Validity

Our core conclusion—that MV rhythm selectively affects precision—relies on the robust estimation of K and σ. We acknowledge the methodological limitations inherent in our design.

- Capacity Estimation (K): The estimation of K using the fixed set size (N=4) and the K=N×(H−F) formula is a simplified approach. While we report the detailed H and F values (Table 1) and confirmed no reliable change in decision bias (d' and criterion were stable across conditions), we recognize that this method may lack the sensitivity to detect subtle changes in capacity that might be revealed by a multi-set size design.
- Precision Estimation (σ): The use of a mixed task design (80% change detection, 20% continuous report) and the resulting lower trial count for continuous report (approx. 20 trials per condition per participant) may introduce instability in the mixture model parameter estimation, particularly for σ. To address this, we have provided detailed model diagnostics in the Supplementary Materials, including parameter recovery simulations and ESS values, which indicate that the condition effect on σ is robustly estimated despite the trial count. Future studies should employ a dedicated continuous report task with a larger number of trials to further validate the stability of the σ estimate.

5.3. Alternative Explanations and Confounding Variables

Although the findings align with the OTS framework, alternative explanations related to arousal, affect, and task structure merit careful consideration.

- Arousal and Affect: The MV rhythm, being more "human-like," may simply be more engaging or pleasurable, leading to increased arousal or motivation, which could non-specifically boost cognitive performance. We controlled for physical sound properties by normalizing the RMS amplitude of ME and MV stimuli. However, our current design cannot fully dissociate the "temporal structure optimization" from the "arousal/pleasure enhancement." Future research should incorporate subjective measures (e.g., self-reported arousal, liking, or "groove" ratings) and use non-rhythmic, but equally arousing, control sounds to disentangle these mechanisms.
- Block Order Effects: Although the condition order was fully counterbalanced, the block design means that participants may have adjusted their strategy or experienced differential fatigue across blocks. We performed an additional check by including the block order as a covariate in the BMRM, which did not reveal a significant interaction with the condition effect, suggesting the observed effect is robust to block-related confounds.

5.4. Individual Differences and Future Directions

The exploratory positive correlation between musical training and the MV-related precision benefit suggests that the extent to which individuals can take advantage of OTS may depend

on sensitivity to rhythmic cues. This pattern is consistent with prior work indicating that musical training can enhance cross-modal integration and auditory processing [20]. However, in the present study musical training was captured by a single self-report measure ("years of formal instruction"), which provides only a coarse proxy for underlying rhythmic abilities.

Future research should therefore employ more fine-grained and objective measures, such as rhythm discrimination tasks, sensorimotor synchronization paradigms, or EEG-based entrainment indices, to test whether enhanced rhythmic processing is indeed the mechanism underlying this individual difference. Such work could clarify whether musical training per se, or specific aspects of temporal processing, are most critical for benefiting from MV-based temporal scaffolding.

6. Conclusion and Outlook

The present study provides converging behavioral evidence that expressive rhythmic micro-variations can selectively enhance VWM precision in a set-size-4 color change-detection paradigm, while leaving VWM capacity unchanged. Within the framework of Optimized Temporal Scaffolding, the MV rhythm appears to confer an additional advantage over both silence and a perfectly mechanical rhythm, underscoring the cognitive relevance of ecologically valid temporal structure.

More concretely, we have shown that, in this specific task and parameter regime, MV was reliably associated with lower σ (higher precision) than both SI and ME, whereas K remained statistically stable across conditions. At the same time, our conclusions are bounded by several design choices. Future work should: (1) employ multi–set size designs and larger numbers of continuous-report trials to more definitively assess the selectivity of the effect on σ versus K; (2) use EEG or MEG to directly test whether MV improves the quality of cross-modal neural entrainment and alpha-band oscillatory dynamics; and (3) incorporate subjective measures and non-rhythmic, arousal-matched control sounds to more clearly dissociate OTS-related mechanisms from general arousal or affective influences.

Within these boundaries, the present findings offer a principled starting point for understanding how subtle temporal structure in sound can support visual cognition. We caution against direct extrapolation to educational or clinical applications until the underlying neural mechanisms are more fully characterized, but the results point to promising avenues for designing acoustic environments that support, rather than impair, core cognitive functions.

References

[1] Luck SJ, Vogel EK. The capacity of visual working memory for features and conjunctions. Nature. 1997 Nov 20;390(6657):279-81. https://doi.org/10.1038/36846

[2] Baddeley A. Working memory: looking back and looking forward. Nature reviews neuroscience. 2003 Oct 1;4(10):829-39.https://doi.org/10.1038/nrn1201

[3] Fukuda K, Vogel E, Mayr U, Awh E. Quantity, not quality: The relationship between fluid intelligence and working memory capacity. Psychonomic bulletin & review. 2010 Oct;17(5):673-9. https://doi.org/10.3758/17.5.673

[4] Large EW, Jones MR. The dynamics of attending: how people track time-varying events. Psychological review. 1999 Jan;106(1):119.

[5] Thut G, Schyns PG, Gross J. Entrainment of perceptually relevant brain oscillations by non-invasive rhythmic stimulation of the human brain. Frontiers in psychology. 2011 Jul 20;2:170.https://doi.org/10.3389/fpsyg.2011.00170

[6] Repp BH. Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's ''Träumerei''. The Journal of the Acoustical Society of America. 1992 Nov 1;92(5):2546-68.

[7] Ayyildiz C, Milne AJ, Irish M, Herff SA. Micro-variations in timing and loudness affect music-evoked mental imagery. Scientific Reports. 2025 Aug 22;15(1):30967. https://doi.org/10.1038/s41598-025-12604-4

[8] Ma WJ, Husain M, Bays PM. Changing concepts of working memory. Nature neuroscience. 2014 Mar;17(3):347-56. https://doi.org/10.1038/nn.3655

[9] Vogel EK, Machizawa MG. Neural activity predicts individual differences in visual working memory capacity. Nature. 2004 Apr 15;428(6984):748-51.https://doi.org/10.1038/nature02447

[10] Zhang W, Luck SJ. Discrete fixed-resolution representations in visual working memory. Nature. 2008 May 8;453(7192):233-5. https://doi.org/10.1038/nature06860

[11] Barton B, Ester EF, Awh E. Discrete resource allocation in visual working memory. Journal of Experimental Psychology: Human Perception and Performance. 2009 Oct;35(5):1359.

[12] Sauseng P, Klimesch W, Heise KF, Gruber WR, Holz E, Karim AA, Glennon M, Gerloff C, Birbaumer N, Hummel FC. Brain oscillatory substrates of visual short-term memory capacity. Current biology. 2009 Nov 17;19(21):1846-52.

[13] Sattelberger J, Haque H, Juvonen JJ, Siebenhühner F, Palva JM, Palva S. Local and interareal alpha and low-beta band oscillation dynamics underlie the bilateral field advantage in visual working memory. Cerebral Cortex. 2024 Nov;34(11):bhae448. https://doi.org/10.1093/cercor/bhae448

[14] Lakatos P, Musacchia G, O'Connel MN, Falchier AY, Javitt DC, Schroeder CE. The spectrotemporal filter mechanism of auditory selective attention. Neuron. 2013 Feb 20;77(4):750-61. https://doi.org/10.1016/j.neuron.2012.11.034

[15] Escoffier N, Herrmann CS, Schirmer A. Auditory rhythms entrain visual processes in the human brain: evidence from evoked oscillations and event-related potentials. NeuroImage. 2015 May 1;111:267-76. https://doi.org/10.1016/j.neuroimage.2015.02.024

[16] Albouy P, Martinez-Moreno ZE, Hoyer RS, Zatorre RJ, Baillet S. Supramodality of neural entrainment: Rhythmic visual stimulation causally enhances auditory working memory performance. Science advances. 2022 Feb 23;8(8):eabj9782. https://doi.org/10.1126/sciadv.abj9782

[17] Morillon B, Baillet S. Motor origin of temporal predictions in auditory attention. Proceedings of the National Academy of Sciences. 2017 Oct 17;114(42):E8913-21.

[18] Grahn JA. The role of the basal ganglia in beat perception: neuroimaging and neuropsychological investigations. Annals of the New York Academy of Sciences. 2009 Jul;1169(1):35-45.

[19] Palmer C. Music performance. Annual review of psychology. 1997 Feb;48(1):115-38.
[20] Parbery-Clark A, Strait DL, Anderson S, Hittner E, Kraus N. Musical experience and the aging auditory system. implications for cognitive abilities and hearing speech in noise. 2011;2011:6. https://doi.org/10.1371/journal.pone.0018082
[21] Davies M, Madison G, Silva P, Gouyon F. The effect of microtiming deviations on the perception of groove in short rhythms. Music Perception: An Interdisciplinary Journal. 2012 Dec;30(5):497-510. https://doi.org/10.1525/mp.2013.30.5.497