

GIMFlow: An LLM-Guided Dialogue-to-Music System for GIM-Inspired Well-Being Experiences

Ruijia Huang^{1,*} Yao Yao^{2,3,*} Hai Zhao^{2,3,†}

¹University of Toronto

²AGI Institute, School of Computer Science, Shanghai Jiao Tong University

³Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University

ruijia.huang@mail.utoronto.ca, yaoyao27@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

*Equal contribution. †Corresponding author.

Abstract

Guided Imagery and Music (GIM) organizes verbal preparation, guided listening, imagery exploration, and reflection into a structured, music-centered process. Existing LLM-based well-being agents remain text-centered, and emotion-aware music systems operate outside such multi-phase workflows, conditioning music on static mood descriptions. We present *GIMFlow*, an interactive, non-clinical GIM-inspired system that translates reflective dialogue into a structured musical trajectory through four phases: Prelude, Induction, Music & Imaging, and Postlude. *GIMFlow* estimates current and target affective states in continuous valence–arousal space, constructs a deterministic four-waypoint transition, and maps waypoints to sections of an LLM-generated symbolic music blueprint for deterministic rendering; the system also supports metadata-based retrieval and unified session-level logging. A formative within-subject study (N=23 participants) of the initial prototype revealed that generated music captured only the endpoint state; we then aligned waypoints with intrapiece sections and conducted a preliminary within-subject evaluation of the refined system (N=10). Results show descriptively favorable trends in valence and negative affect while positive affect and arousal are mixed; all findings are exploratory. *GIMFlow* is released as an open-source, installable research demonstration with example sessions and reproducible logging utilities.

Installable demo package / code:

[https://github.com/](https://github.com/ruijiahuang2006-collab/gimflow)

[ruijiahuang2006-collab/gimflow](https://github.com/ruijiahuang2006-collab/gimflow)

Video: <https://youtu.be/xfEQDTup9pg>.

This research was supported by the Shanghai Jiao Tong University 2030 Initiative and The Major Program of Chinese National Foundation of Social Sciences under Grant “The Challenge and Governance of Smart Media on News Authenticity” [No. 23&ZD213].

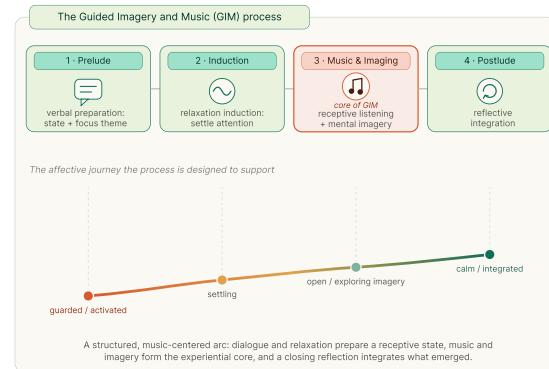


Figure 1: The Guided Imagery and Music (GIM) process. *GIMFlow* is inspired by GIM’s four receptive phases (Prelude, verbal preparation; Induction, relaxation; Music & Imaging, receptive listening with mental imagery; and Postlude, reflective integration) and by the affective journey they are designed to support, from an initial guarded state toward a calmer, more integrated one.

1 Introduction

Guided Imagery and Music (GIM) (Bonny, 1989; Grocke, 2019) is a structured music-and-imagery experience combining verbal preparation, relaxation induction, music listening with imagery exploration, and post-listening integration (Figure 1). Its structured, multi-phase, music-centered nature makes it a compelling reference model for well-being LLM systems, yet to our knowledge no existing LLM agent implements this full workflow.

Building such an agent poses three technical challenges: (i) *phase-scripted dialogue*, requiring the agent to recognize which phase the participant is in and produce contextually appropriate prompts; (ii) *affective trajectory planning* from the participant’s current state to a target state, rather than reacting to a single static mood label; and (iii) *cross-modal mapping* of that trajectory to internal music structure, so the emotional arc that guides dialogue also shapes music progression.

Existing work addresses these only partially. Mental well-being chatbots (Fitzpatrick et al., 2017; Fulmer et al., 2018; Inkster et al., 2018) and empathetic dialogue systems (Rashkin et al., 2019; Poria et al., 2019) engage in supportive dialogue but do not treat music as a core medium. Emotion-aware music generation and retrieval (Agostinelli et al., 2023; Copet et al., 2024; Liu et al., 2024; Melechovsky et al., 2024; Yang and Chen, 2012) outputs music matching a mood description, but operates outside a multi-phase workflow and conditions on a single static mood without modeling a current-to-target trajectory. Recent works (Kar et al., 2025; Hotta et al., 2025) show that dialogue systems can be enriched with affect- or cognition-aware components for adjacent applications such as tutoring and coaching, but none realize the phase-scripted, trajectory-conditioned music workflow that GIM entails.

We present *GIMFlow*, an LLM-guided GIM-inspired well-being agent, offered as a non-clinical research demonstration for the NLP and affective-computing community and for developers of music-centered interactive systems. GIMFlow (i) structures the interaction into four phases (Prelude, Induction, Music & Imaging, Postlude) coordinated by an LLM state classifier; (ii) estimates current and target affective states in continuous valence-arousal (VA) space; (iii) constructs a deterministic four-waypoint trajectory; (iv) translates it into a section-level symbolic music blueprint rendered procedurally to audio; and (v) supports a counter-balanced metadata-based retrieval condition. **Contributions:** (1) a phase-scripted workflow agent architecture coupling dialogue, VA planning, and music; (2) two within-subject user studies (formative pilot N=23; preliminary evaluation of refined system N=10) with paired affective measures and an exploratory blind pairwise LLM-as-judge analysis using position-swapped inputs and keyword-based leakage checks; (3) an open-source release with prompts, example sessions, aggregate data, reproducible logging utilities, and exploratory analysis scripts.

2 Related Work

GIM and music therapy. The Bonny Method combines relaxation induction, evocative music, and guided verbal exploration (Bonny, 1989; Grocke, 2019). Prior empirical work and systematic reviews report favorable affective and physiological

outcomes in some populations (McKinney et al., 1997; McKinney and Honig, 2016; De Witte et al., 2022), with recent work examining remote and technology-mediated settings (Honig and Hannibal, 2022). GIMFlow operationalizes selected workflow elements of GIM and evaluates usability rather than therapeutic outcomes.

Conversational and emotion-aware systems. Chatbot systems for mental well-being include Woebot (Fitzpatrick et al., 2017), Wysa (Inkster et al., 2018), and Tess (Fulmer et al., 2018). LLM-based mental-health systems have been surveyed by Guo et al. (2024). Empathetic and emotion-aware dialogue is studied via EmpatheticDialogues (Rashkin et al., 2019) and the emotion-recognition-in-conversation literature (Poria et al., 2019; Zhang et al., 2025), with recent EMNLP demonstrations for tutoring (Kar et al., 2025) and coaching (Hotta et al., 2025). These systems deliver text-based dialogue; GIMFlow’s core interaction medium is music. Music emotion recognition and emotion-aware recommendation frame selection in valence-arousal space (Russell, 1980; Posner et al., 2005; Yang and Chen, 2012; Tran et al., 2023; Schedl et al., 2018). Controllable text-to-music generation includes MusicLM (Agostinelli et al., 2023), MusicGen (Copet et al., 2024), and Mustango (Melechovsky et al., 2024). GIMFlow’s “generation” is symbolic blueprint-to-MIDI rendering, not neural text-to-audio synthesis.

3 System Design

Figure 2 shows the architecture. A browser-facing client and GIMFlow server coordinate six modules: a *Session Manager*, a *Facilitator Agent* implementing a four-phase LLM dialogue, an *Emotion/VA Planner*, a *Music Module* with two counterbalanced conditions, a *Questionnaire Module*, and a *Logger* that exports full session-level logs. We describe each module below.

Session Manager. The Session Manager maintains the shared session state and coordinates within-subject session progression: it randomly assigns the condition order (persisted across sessions), tracks session completion, and enforces the 5-minute washout interval before the second session begins.

Facilitator Agent. The Facilitator Agent is a prompt-driven LLM wrapper that implements the four-phase dialogue mirroring the receptive-style GIM arc: **Prelude** elicits current state and focus

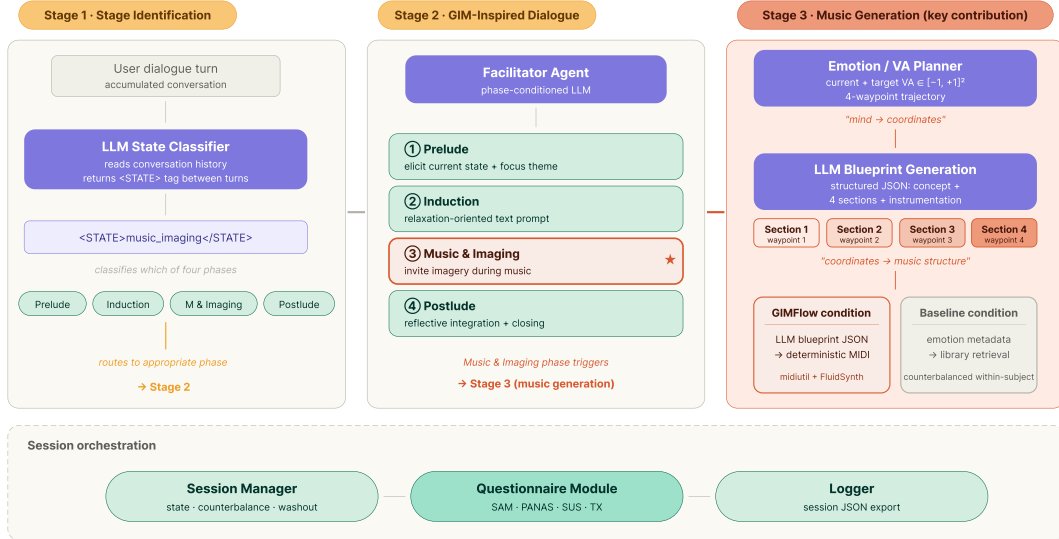


Figure 2: GIMFlow system architecture. Stage 1 identifies the current dialogue phase using an LLM state classifier; Stage 2 produces phase-conditioned facilitation dialogue across Prelude, Induction, Music & Imaging, and Postlude; Stage 3 translates current and target VA states into a four-waypoint trajectory and maps the waypoints onto four sections of a symbolic music blueprint, rendered in the GIMFlow condition or replaced by metadata-based retrieval in the Baseline condition. The Session Manager, Questionnaire Module (SAM/PANAS/SUS/TX), and Logger coordinate the shared session state and export one log record per session.

theme; **Induction** provides a relaxation-oriented text prompt; **Music & Imaging** invites playback of a selected or generated piece and the description of arising imagery; **Postlude** integrates the experience into a brief reflective summary. Phase transitions are governed by an LLM state classifier that reads the accumulated conversation and returns the current phase between turns as a `<STATE>` tag. Each phase concatenates a base system prompt with a phase-specific guidance block (released prompts in Appendix A).

Emotion/VA Planner. The Emotion/VA Planner infers the participant’s current and target affective states in continuous valence–arousal space from the Prelude dialogue. A VA estimation prompt returns a rationale, focus theme, and (current VA, target VA) pair with continuous values in $[-1, +1]$. A deterministic interpolator constructs a four-waypoint linear trajectory $w_i = \text{current} + \frac{i}{3}(\text{target} - \text{current})$ for $i \in \{0, 1, 2, 3\}$, aligning waypoints with the four sections of the music blueprint. The trajectory is planned once per session and drives the Music Module.

Music Module. In the GIMFlow condition,¹ the planner’s four waypoints seed an LLM-generated symbolic music blueprint (concept, four-section

¹The deployed system, session logs, and repository retain the legacy internal label `kimusic` for the GIMFlow generation condition.

structure with per-section VA waypoints, and instrumentation) that a procedural renderer using `midutil`, `fluidsynth`, and a General MIDI soundfont converts to MIDI and mp3 audio, played through a browser music player. For comparison, a retrieval Baseline (§4) forms the second within-subject condition.

Questionnaire Module. The Questionnaire Module delivers the Self-Assessment Manikin (SAM) and Positive and Negative Affect Schedule (PANAS) as pre- and post-session measures within each of the two sessions, and the System Usability Scale (SUS) together with a custom Therapy Experience (TX) instrument as post-experiment measures after both sessions are complete. Item texts are provided in the released repository.

Logger. The Logger exports one JSON per session with a unified schema covering session identifiers, pre/post questionnaires, phase-labeled dialogue turns, Emotion/VA Planner outputs, Music Module traces (GIMFlow blueprint or Baseline retrieval record), and timestamps. Appendix B summarizes the top-level field groups.

4 Implementation

GIMFlow is a Python web application using Gradio. Dialogue turns, emotion estimation, and blueprint generation are handled by prompting

GPT-5.5 through an OpenAI-compatible API; the GIMFlow condition renders audio with midiutil, fluidsynth, and a General MIDI soundfont. To assess the contribution of blueprint-conditioned generation, we constructed a retrieval Baseline for within-subject comparison: it draws MP3 tracks from an indexed library of roughly 2,000 instrumental tracks with structured metadata (title, genre, mood, tags, and audio features such as tempo and dynamics), mapping the LLM’s mood keywords to five heuristic mood clusters (restorative, energetic, tense, melancholic, neutral), requesting up to 12 candidates with relaxed and then general fallback criteria, and randomly selecting among eligible unused candidates. Track duration is not matched across conditions; the two conditions share every component except the Music Module, so condition differences reflect music delivery rather than the dialogue agent. The Session Manager enforces a 5-minute countdown between the two counter-balanced sessions in a single visit. The GIMFlow source code is released under the MIT license; the baseline audio library is not redistributed (§7). We release the source code, session logging schema, illustrative session-JSON exports, demo audio, aggregate analysis scripts, and exploratory AI-evaluation utilities; user-level raw logs and questionnaire records are not released for privacy reasons. Figure 3 shows representative UI panels from a Session 1 flow.

5 Evaluation

Study 1 (N=23) surfaced a Music Module implementation gap; a refined implementation was evaluated in Study 2 (N=10). We also report an exploratory LLM-based qualitative audit of anonymized session logs. All numeric results are descriptive; p-values are exploratory and uncorrected. Both studies use a two-session within-subject counterbalanced protocol: each session collects pre/post SAM and PANAS around the four-phase dialogue, with a system-enforced 5-minute washout between sessions. SUS and TX are administered once after Session 2.

Study 1 (formative pilot; N=23 participants). Of 24 participants, one was excluded for incomplete post-questionnaires, leaving 23 for the paired analysis (12 Baseline→GIMFlow, 11 GIMFlow→Baseline). The session used a Chinese-language deployment; the workflow, VA planner, questionnaires, and logging are language-agnostic.

Table 1 reports paired SAM and PANAS changes. Direction-consistent descriptive trends favor the GIMFlow condition on SAM Valence and PANAS negative affect; arousal changes are reported as activation. Overall usability (SUS = 73.30 ± 18.83 , N=22; one participant missed the post-experiment response) is in the acceptable range (Brooke, 1996); overall experience (TX = 5.85 ± 0.93 on a 7-point scale) is favorable.

Measure	Baseline Δ	GIMFlow Δ	d	p
SAM Valence	$+0.30 \pm 2.40$	$+1.00 \pm 1.41$	+0.23	.35
SAM Arousal	-0.26 ± 2.47	$+0.35 \pm 2.10$	+0.25	.25
PANAS PA	-0.70 ± 9.19	$+2.52 \pm 5.36$	+0.37	.18
PANAS NA	$+0.26 \pm 6.59$	-1.96 ± 4.61	-0.26	.38

Table 1: Study 1 paired SAM/PANAS Δ (N=23). Cohen’s d and Wilcoxon p are exploratory.

Refinement. Post-hoc examination of Study 1 session logs showed that although the VA Planner produced a four-waypoint sequence and the LLM produced a matching blueprint, the initial Music Module rendered a single piece aligned only with the endpoint waypoint. We refined the Music Module so that the blueprint’s four structural sections are deterministically mapped onto the four waypoints (section_i.waypoint_index = i, section_i.waypoint_va = waypoint_i.va). The LLM is informed of the trajectory when generating the blueprint’s overall parameters; the renderer executes each section with a per-key chord progression and a section-index-based dynamics profile. All other GIMFlow components (Session Manager, Facilitator Agent, VA Planner, Questionnaire Module, Logger) are unchanged.

Study 2 (preliminary evaluation of refined system, N=10). Ten new participants who had not participated in Study 1 completed the same protocol with the refined Music Module (4 Baseline→GIMFlow, 6 GIMFlow→Baseline). Direction-consistent descriptive trends again favor GIMFlow on SAM Valence ($+1.50 \pm 1.35$ vs. $+1.20 \pm 0.79$, $d = +0.20$) and PANAS negative affect (-3.30 ± 4.45 vs. -2.10 ± 2.56 , $d = -0.30$); PANAS positive affect descriptively favors Baseline in this small sample. Overall SUS = 75.00 ± 14.19 ; overall TX = 5.50 ± 0.96 . Full pairwise Δ values are in Appendix C.

Exploratory blind pairwise LLM-based analysis. For each participant, condition labels were stripped and the LLM judge (Claude Sonnet 4.6, temperature 0) received structured pre/post affect-

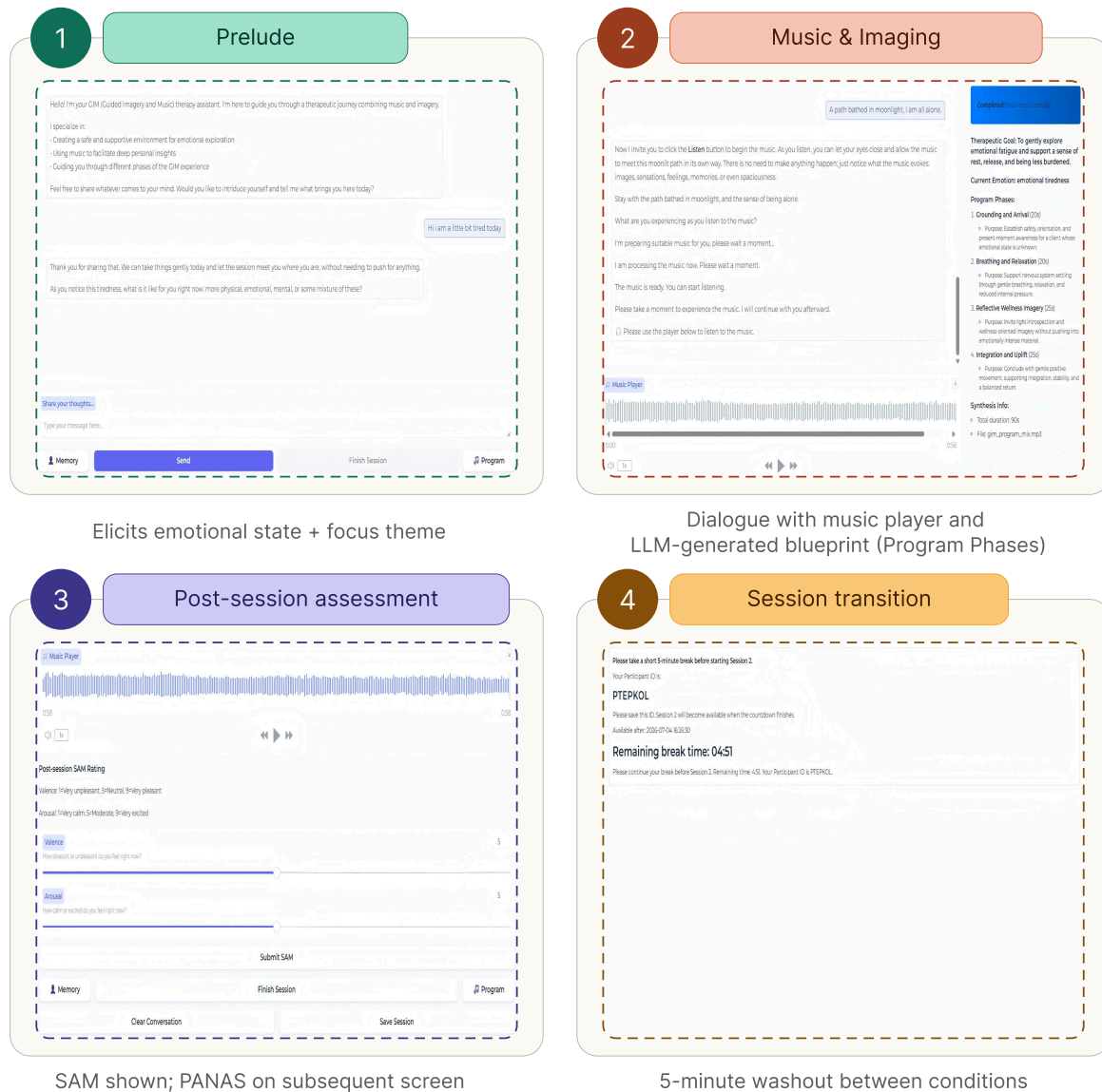


Figure 3: Representative UI panels from a Session 1 flow. ① Prelude elicits current state and focus theme; ② Music & Imaging presents dialogue with music player and the LLM-generated blueprint (Program Phases); ③ Post-session assessment with SAM (PANAS on the subsequent screen); ④ 5-minute washout between conditions with participant ID.

tive outcomes and deltas, music-profile metadata, participant questionnaire responses, and phase-stratified representative dialogue excerpts in both A/B position orders; the judge did not receive or analyze audio. The judge rated the two sessions on five criteria (*emotional alignment*, *facilitation coherence*, *music-profile alignment*, *engagement*, *safety*; 5-point scale) and chose an overall winner with a confidence value in [0, 1]. Each participant's two position-swapped evaluations were aggregated to the participant level (per-criterion scores averaged across positions; a winner assigned only when both positions agreed), and statistical tests used

these participant-level scores rather than individual calls. The protocol included a keyword-based blind check for explicit condition-identifying strings (0 hits across 33 participant-level pair inputs prior to position swapping), A/B position swaps, and phase-stratified excerpts with length caps. Study 1: 11 GIMFlow wins, 11 Baseline wins, 1 tie (22/23 position-swap agreement, 95.65%); Study 2: 5 GIMFlow, 3 Baseline, 2 ties (8/10 agreement, 80.0%). Overall GIMFlow–Baseline mean rating differences on the 5-criterion overall mean were +0.23 (Study 1) and +0.20 (Study 2); Wilcoxon p on GIMFlow – Baseline overall ratings were

Component	MH chatbots (Fitzpatrick et al., 2017; Inkster et al., 2018; Ful- mer et al., 2018)	Emotion dialogue (Rashkin et al., 2019; Po- ria et al., 2019; Zhang et al., 2025)	Music rec. (Tran et al., 2023; Wang et al., 2021; Schedl et al., 2018)	Text-to-music (Agostinelli et al., 2023; Copet et al., 2024; Mele- chovsky et al., 2024; Liu et al., 2024)	GIMFlow ours
Multi-turn dialogue	✓	✓	✗	✗	✓
Structured GIM work- flow	✗	✗	✗	✗	✓
Music as core interac- tion medium	✗	✗	✓	partial	✓
Explicit current→target VA planning	✗	✗	partial	✗	✓
Supports generation and retrieval	–	–	retrieval	generation	✓
Session-level logging	partial	partial	partial	✗	✓

Table 2: Functional components across representative related system categories. Cell judgments reflect published feature coverage of the specific papers cited per column.

.30 and .45. Although criterion-level mean ratings descriptively favored GIMFlow on all five criteria (Appendix D), participant-level overall preferences were tied in Study 1 and only weakly favored GIMFlow in Study 2. We therefore interpret the LLM-judge results as an exploratory structured qualitative audit rather than as independent evidence of condition-level user benefit.

Case study. To illustrate the refined Music Module (Study 2), we describe participant P408J7U, whose 21.9-minute GIMFlow session cleanly exemplifies the refinement described above. The session opened the Prelude with a Chinese description of academic pressure, from which the VA Planner extracted focus theme “*tension under academic pressure*” with current VA = (−0.6, 0.35) and target VA = (0.25, −0.25). The LLM blueprint’s four sections carried waypoint_va matching the planner’s four waypoints. Pre/Post SAM (Valence, Arousal) moved from (2, 2) to (6, 6); Pre/Post PANAS (PA, NA) moved from (16, 37) to (25, 25). Extended session content is in Appendix E.

6 Comparison with Existing Systems

Table 2 summarizes functional coverage across representative categories. GIMFlow differs from these categories in coupling a phase-scripted GIM dialogue workflow, explicit current-to-target VA planning, and blueprint-conditioned symbolic music generation within one reproducible pipeline: mental well-being chatbots and empathetic dialogue systems do not treat music as the core interaction medium, and emotion-aware music systems condition on a single static mood without a multi-phase, trajectory-driven workflow.

7 Conclusion

GIMFlow demonstrates that a phase-scripted GIM-inspired dialogue workflow, LLM-based valence-arousal estimation with waypoint planning, and blueprint-conditioned symbolic music rendering can be integrated into a single LLM-based system that produces traceable dialogue, planning, and music-structure trajectories within a session. GIMFlow is released under the MIT license as a non-clinical demonstration; session-level JSON logs, example sessions, and analysis utilities support reproducible offline inspection and future evaluation. Future work includes larger studies powered for null-hypothesis tests, richer waypoint-to-structure couplings in the renderer, and non-clinical prompt-level terminology in future deployments (§8).

8 Limitations

The current evaluation is based on two small, Chinese-language studies, so all findings should be considered preliminary. Future work will examine larger and more diverse populations and more fully control for order effects; an exploratory split (Appendix G) already showed lower post-experiment SUS/TX when GIMFlow occurred second, which, given single post-session administration, we read as a likely order or recency effect rather than condition preference. The current studies compare the full GIMFlow generation condition against a retrieval baseline, but do not causally isolate the explicit VA trajectory mechanism itself. The VA estimates and waypoint-to-section mappings currently serve as inspectable planning representations; their alignment with self-reported affect and perceived musical emotion will be evaluated through human anno-

tation, listening studies, and audio analysis. Several confounds remain uncontrolled, including novelty, perceived personalization, duration mismatch, order effects, and quality differences between curated retrieval and symbolic generation. Future ablations should compare target-only generation, shuffled waypoints, and no-trajectory generation to separate trajectory-planning from personalization and novelty effects. We do not claim state-of-the-art audio generation quality: the current symbolic renderer is intended to make the planned affective structure inspectable and reproducible; future work should compare perceived audio quality against neural text-to-music systems and conduct human listening studies. The exploratory LLM-judge analysis also has important limitations. The judge received structured pre/post affective outcomes and questionnaire summaries in addition to dialogue excerpts and music-profile metadata, so its ratings are not fully independent of the quantitative outcomes reported in the user studies. In addition, our blind check focused on explicit condition-identifying strings and cannot rule out more subtle structural cues in the input representation. We therefore treat the LLM-judge results as a structured qualitative audit rather than as independent evidence of condition-level user benefit. Future work should use independently validated human judgments, listening studies, or judge protocols that exclude outcome summaries and better control condition-specific metadata. We will also evaluate open-source LLMs, validated experience measures, and standardized non-clinical prompts to improve reproducibility.

9 Ethics Statement

No formal institutional IRB/REB approval was obtained for this pilot demonstration study. Study 1 used a written recruitment notice that described the two-session protocol, session length, and operational instructions; it did not include a formal written consent form, explicit pre-participation exclusion criteria, or crisis-support resources. For Study 1, several mitigating factors reduced potential ethical risk in the absence of a formal consent form: (a) the recruitment notice explicitly framed the study as a research prototype of an AI-guided music experience; (b) participation was voluntary and could be terminated at any time by closing the browser or ending the session; (c) no direct personally identifying information was requested, and participant identifiers are pseudonymous 7-character system-

generated codes with no mapping to real-world identities; (d) individual sessions were designed to last approximately 15–25 minutes, though actual duration varied with participant interaction; and (e) no adverse experience was spontaneously reported to the research team during the study. These factors do not substitute for a formal consent process, and Study 2 was designed specifically to address the procedural gaps identified in Study 1. Study 2 introduced a written informed consent form that explicitly stated the non-clinical, research-prototype nature of the system; pre-participation exclusion criteria (current treatment for depression or anxiety, severe recent distress or self-harm ideation within three months, history of dissociative or psychotic symptoms, and current medications affecting mood or consciousness); an explicit right to withdraw at any point with optional deletion of collected data; and crisis-support hotline resources embedded in the consent form. Participants in both studies were informed that the study collected pre/post questionnaires, dialogue logs, and session metadata, and that no personally identifying content was required. Participant identifiers are pseudonymous 7-character strings; questionnaire records are stored separately from dialogue logs, keyed by pseudonymous ID. Raw data are retained for 5 years for research verification purposes and are not publicly released; only aggregate analytical outputs (means, standard deviations, effect sizes, and exploratory AI-evaluation summaries) accompany this paper. Participants were free to withdraw at any time. GIMFlow is a technology demonstration and research prototype, not a clinical intervention or therapeutic service. It is not a substitute for professional mental-health care. The released system prompts (Appendix A) explicitly frame the system as GIM-inspired well-being facilitation rather than therapy. The deployed system used therapist-oriented wording, whereas the released prompts (Appendix A) use non-clinical facilitator terminology, so participant responses may not fully generalize to the released wording.

References

- Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. MusicLM: Generating Music From Text, January 2023.
- H. L. Bonny. Sound as Symbol: Guided Imagery and Music in Clinical Practice. *Music Therapy Perspectives*, 6(1):7–10, January 1989.
- Margaret M. Bradley and Peter J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, March 1994.
- John Brooke. SUS: A quick and dirty usability scale. In Patrick W. Jordan, Bruce Thomas, Bernard A. Weerdmeester, and Ian L. McClelland, editors, *Usability Evaluation in Industry*, pages 189–194. Taylor & Francis, 1996.
- Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and Controllable Music Generation, January 2024.
- Martina De Witte, Ana Da Silva Pinho, Geert-Jan Stams, Xavier Moonen, Arjan E.R. Bos, and Susan Van Hooren. Music therapy for stress reduction: a systematic review and meta-analysis. *Health Psychology Review*, 16(1):134–159, January 2022.
- Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health*, 4(2):e19, June 2017.
- Russell Fulmer, Angela Joerin, Breanna Gentile, Lysanne Lakerink, and Michiel Rauws. Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial. *JMIR Mental Health*, 5(4):e64, December 2018.
- Denise Erdonmez Grocke. *Guided Imagery and Music: The Bonny Method and Beyond*. Barcelona Publishers, Dallas, TX, 2nd ed edition, 2019.
- Zhijun Guo, Alvina Lai, Johan Hilge Thygesen, Joseph Farrington, Thomas Keen, and Kezhi Li. Large Language Models for Mental Health Applications: A Systematic Review. *JMIR Mental Health*, 11:e57400, 2024.
- Timothy J. Honig and Niels Hannibal. Client experiences of shifting from in-person to telehealth formats of Guided Imagery and Music (GIM) sessions. *The Arts in Psychotherapy*, 81:101954, November 2022.
- Hajime Hotta, Huu-Loi Le, Manh-Cuong Phan, and Minh-Tien Nguyen. Metamo: Empowering Large Language Models with Psychological Distortion Detection for Cognition-aware Coaching. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 862–872, Suzhou, China, 2025. Association for Computational Linguistics.
- Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR mHealth and uHealth*, 6(11):e12106, November 2018.
- Debanjana Kar, Leopold Böss, Dacia Braca, Sebastian Maximilian Dennerlein, Nina Christine Hubig, Philipp Wintersberger, and Yufang Hou. MathBuddy: A Multimodal System for Affective Math Tutoring. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 821–838, Suzhou, China, 2025. Association for Computational Linguistics.
- Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024.
- Cathy H. McKinney and Timothy J. Honig. Health Outcomes of a Series of Bonny Method of Guided Imagery and Music Sessions: A Systematic Review. *Journal of Music Therapy*, page thw016, December 2016.
- Cathy H. McKinney, Michael H. Antoni, Mahendra Kumar, Frederick C. Tims, and Philip M. McCabe. Effects of Guided Imagery and Music (GIM) Therapy on Mood and Cortisol in Healthy Adults. *Health Psychology*, 16(4):390–400, 1997.
- Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. Mustango: Toward Controllable Text-to-Music Generation, June 2024.
- Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances. *IEEE Access*, 7:100943–100953, 2019.
- Jonathan Posner, James A. Russell, and Bradley S. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3):715–734, September 2005.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y.-Lan Boureau. Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset, August 2019.
- James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, December 1980.

- Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, 7(2):95–116, June 2018.
- Hieu Tran, Tuan Le, Anh Do, Tram Vu, Steven Bogaerts, and Brian Howard. Emotion-Aware Music Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13):16087–16095, 2023.
- Shu Wang, Chonghuan Xu, Austin Shijun Ding, and Zhongyun Tang. A Novel Emotion-Aware Hybrid Music Recommendation Method Using Deep Neural Network. *Electronics*, 10(15):1769, July 2021.
- David Watson, Lee Anna Clark, and Auke Tellegen. Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. *Journal of Personality and Social Psychology*, 54(6):1063–1070, 1988.
- Yi-Hsuan Yang and Homer H. Chen. Machine Recognition of Music Emotion: A Review. *ACM Transactions on Intelligent Systems and Technology*, 3(3):1–30, May 2012.
- Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. DialogueLLM: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. *Neural Networks*, 192:107901, December 2025.

A Released Prompt Excerpts

Prompts below are **public-release versions** (non-clinical facilitator wording); deployed versions used therapist-oriented wording (§8).

A.1 Base system prompt (release version).

You are a GIM-inspired well-being facilitator supporting a participant through a four-phase session: Prelude (elicit current state and focus theme), Induction (scaffold entry into imagery), Music & Imaging (invite the participant to listen and describe arising imagery), Postlude (integrate the experience). Between turns, silently return a state tag <STATE>{phase_name}</STATE>. This session is a research demonstration, not clinical care.

A.2 Music blueprint generation prompt (GIM-Flow condition, excerpt).

Given focus theme, current VA (v, a), target VA (v', a'), and the 4-waypoint sequence, return a JSON blueprint with keys: tempo_bpm, key, time_signature, duration_seconds, harmonic_language, melodic_motion, dynamic_curve, texture, instrument_roles, and a 4-item structure array. Each item: {name, bars, character, section_index (0-3), waypoint_index (0-3), waypoint_va {v, a}}.

A.3 VA estimation prompt (excerpt).

Given the dialogue so far, estimate the participant's current valence-arousal (both in [-1.0, +1.0]) and produce a target VA consistent with the stated concern. Return ONLY a JSON object:

```
{
  "current_state_va": {"valence": 0.0, "arousal": 0.0},
  "target_state_va": {"valence": 0.0, "arousal": 0.0},
  "focus_theme": "short string",
  "rationale": "brief string"
}
```

Continuous values preferred; near-zero if ambiguous.

A.4 Baseline music-retrieval prompt (excerpt).

You are a music-selection facilitator. Translate the participant's current focus into structured retrieval criteria. Return ONLY JSON with keys: tempo_preference ("slow"|"medium"|"fast"), dynamics_preference ("soft"|"moderate"|"intense"), mood_keywords (array from DB mood options), genre_keywords (array from DB genres), avoid_keywords (elements to avoid).

The remaining prompts (state classification prompt, phase-specific dialogue prompts, and music-metadata injection template) are released with the code (§4). The VA estimation prompt requests continuous valence and arousal values in $[-1, +1]$ (rather than a categorical label), together with a short natural-language focus theme and a brief rationale. Exploratory LLM-judge utilities and prompts are also included in the repository for transparency, but the judge analysis is interpreted only as a structured qualitative audit.

B Session Logging Schema

Field group	Representative fields	Purpose
identifiers	user_id, session_id, condition, condition_order, session_number	Session provenance
pre/post questionnaires	pre_sam, post_sam, pre_panas, post_panas	Paired affective measures
dialogue log	conversation_log (turn: role, phase, content, timestamp)	Phase-labeled dialogue
VA + music	current_state_va, target_state_va, waypoint_sequence, music_sequence	Planner + music
usability + TX status	sus_score, tx_items, completion_status, session_duration_seconds	Post-experiment Bookkeeping

Table 3: Session JSON top-level field groups.

GIMFlow sessions add generation_params with the LLM blueprint; Baseline sessions add retrieval_keywords and emotion_label, with track-level VA null. Full definitions, an example session, exploratory analysis utilities, demographics, and questionnaire items (SAM (Bradley and Lang, 1994), PANAS (Watson et al., 1988), SUS (Brooke, 1996), TX) are in the repository (§4).

C Study 2 Full SAM/PANAS Table

Measure	Baseline Δ	GIMFlow Δ	d	p
SAM Valence	+1.20 $\pm$ 0.79	+1.50 $\pm$ 1.35	+0.20	.47
SAM Arousal	+0.20 $\pm$ 2.10	+0.90 $\pm$ 3.21	+0.30	.53
PANAS PA	+5.20 $\pm$ 4.73	+4.20 $\pm$ 7.32	-0.10	.80
PANAS NA	-2.10 $\pm$ 2.56	-3.30 $\pm$ 4.45	-0.30	.43

Table 4: Study 2 paired SAM/PANAS changes (within-subject, N=10).

D LLM Judge Configuration and Criterion Results

Setting	Value
Judge model	Claude Sonnet 4.6
Temperature	0
Max output tokens	2048
Retry policy	up to 3 attempts on parsing failure
Position swaps	A/B evaluated per participant pair
Output format	JSON with 5-criterion ratings + winner + confidence

Table 5: LLM judge configuration used for the exploratory structured qualitative audit in §5.

Criterion	Study 1			Study 2		
	G	B	Δ	G	B	Δ
Emot. alignment	3.50	3.20	+0.30	3.70	3.35	+0.35
Facil. coherence	3.48	3.41	+0.07	3.60	3.50	+0.10
Music-profile align.	3.57	3.09	+0.48	3.65	3.30	+0.35
Engagement	3.39	3.26	+0.13	3.45	3.35	+0.10
Safety	4.11	3.93	+0.17	4.10	4.00	+0.10
Overall	3.61	3.38	+0.23	3.70	3.50	+0.20

Table 6: Exploratory blind pairwise LLM-judge 5-criterion means (5-point scale). Criterion-level means descriptively favor GIMFlow, but Study 1 overall win counts were tied; these ratings are reported as a structured qualitative audit rather than independent evidence of condition-level user benefit.

E Extended Case Study Details

Session identification. Participant P408J7U, Session 1 (GIMFlow condition), condition_order = [kimusic, baseline], session duration = 21.9 min.

Pre/post questionnaires. Pre SAM (V, A) = (2, 2); Post = (6, 6). Pre PANAS (PA, NA) = (16, 37); Post = (25, 25); Δ PA = +9, Δ NA = -12.

Planning and blueprint. Focus theme (translated from Chinese): “tension under academic pressure”. VA trajectory $(-0.6, 0.35) \rightarrow (-0.317, 0.15) \rightarrow (-0.033, -0.05) \rightarrow (0.25, -0.25)$ maps onto four blueprint sections *orientation*, *transition*, *development*, *target alignment*, with matching waypoint_va.

F Therapy Experience Questionnaire Items

The custom 6-item Therapy Experience (TX) questionnaire was administered once after both conditions were completed. Each item was rated on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). Items were designed to capture different facets of the session experience relevant to a GIM-inspired workflow, but the instrument is not psychometrically validated (§8). **T1 (Empathy):** “I felt the system understood my emotional state during the session.” **T2 (Music match):** “The music generated or selected during the session matched what I was feeling.” **T3 (Imagery facilitation):** “The session helped me notice or explore mental images and inner experiences.” **T4 (Interpretation quality):** “The system’s responses reflected what I actually shared, rather than generic replies.” **T5 (Flow / comfort):** “The overall session flow felt smooth and comfortable.” **T6 (Reuse intention):** “I would consider using this kind of system again in the future.” The TX mean reported in §5 is the per-participant mean of the six item

scores.

G Split of Post-Experiment Scores by Second-Session Condition

Descriptive only. SUS/TX are single post-experiment measures collected after both conditions, so this split cannot be interpreted as condition-specific preference; pattern is more consistent with recency/ordering than with condition-level judgment.

Measure	Study 1 (N=22)		Study 2 (N=10)	
	B-as-S2 (n=11)	G-as-S2 (n=11)	B-as-S2 (n=6)	G-as-S2 (n=4)
SUS score	75.00 $\pm$ 18.23	71.59 $\pm$ 20.13	77.92 $\pm$ 14.95	70.63 $\pm$ 13.75
TX mean	6.02 $\pm$ 0.53	5.68 $\pm$ 1.21	5.94 $\pm$ 0.70	4.83 $\pm$ 0.97
T1 Empathy	6.00 $\pm$ 0.89	5.55 $\pm$ 1.29	6.00 $\pm$ 0.89	4.75 $\pm$ 0.96
T2 Music match	5.91 $\pm$ 1.04	5.91 $\pm$ 1.45	6.17 $\pm$ 0.75	4.75 $\pm$ 0.96
T3 Imagery	6.27 $\pm$ 0.65	5.55 $\pm$ 1.04	6.17 $\pm$ 0.75	4.75 $\pm$ 1.50
T4 Interpretation	5.91 $\pm$ 0.83	5.36 $\pm$ 1.57	5.50 $\pm$ 1.05	4.75 $\pm$ 0.96
T5 Flow	6.09 $\pm$ 1.04	5.82 $\pm$ 1.25	5.83 $\pm$ 0.75	5.00 $\pm$ 1.15
T6 Reuse	5.91 $\pm$ 1.45	5.91 $\pm$ 1.45	6.00 $\pm$ 1.55	5.00 $\pm$ 1.41

Table 7: Split of post-experiment SUS/TX scores by which condition was the second session.

In both studies, participants with GIMFlow as second session report descriptively lower post-experiment scores. Condition-focused interpretation therefore relies primarily on paired pre/post SAM/PANAS (§5), and post-experiment SUS/TX splits should be read only as descriptive order-related patterns.

H Participant Characteristics

Background survey completed by 21 of 23 Study 1 participants (PTL622R submitted but did not complete both sessions; PGFW2QK/PSD7IQP completed sessions without submitting survey) and all 10 Study 2 participants. Both studies drew from young Chinese undergraduates.

Characteristic	Category	S1 (N=21)	S2 (N=10)
Age	18–24	19 (90.5%)	10 (100%)
	25–34	2 (9.5%)	0 (0%)
Gender	Male	9 (42.9%)	2 (20%)
	Female	12 (57.1%)	8 (80%)
Education	UG in progress	18 (85.7%)	9 (90%)
	UG completed	2 (9.5%)	1 (10%)
	MSc in progress	1 (4.8%)	0 (0%)
Music training	Never	16 (76.2%)	6 (60%)
	< 1 yr	4 (19.0%)	2 (20%)
	1–3 yr	1 (4.8%)	2 (20%)
Prior music-therapy	Never	14 (66.7%)	7 (70%)
	Occasionally	6 (28.6%)	3 (30%)
	Multiple times	1 (4.8%)	0 (0%)

Table 8: Participant demographic distributions across both studies. Study 1’s survey offered only Male/Female for gender. Study 2’s survey added Non-binary/Other and Prefer-not-to-say options, but no participants selected those categories.