NEURAL INFORMATION PROCESSING SYSTEMS

PhaseDance: Capturing Rhythm and Expressivity in Dance Modeling

Meongeun Kim Taehui Lee Soomin Park Sung-Hee Lee

KAIST

meongeun.github.io/phasedance

Motivation: What Does It Mean to Dance to the Music?

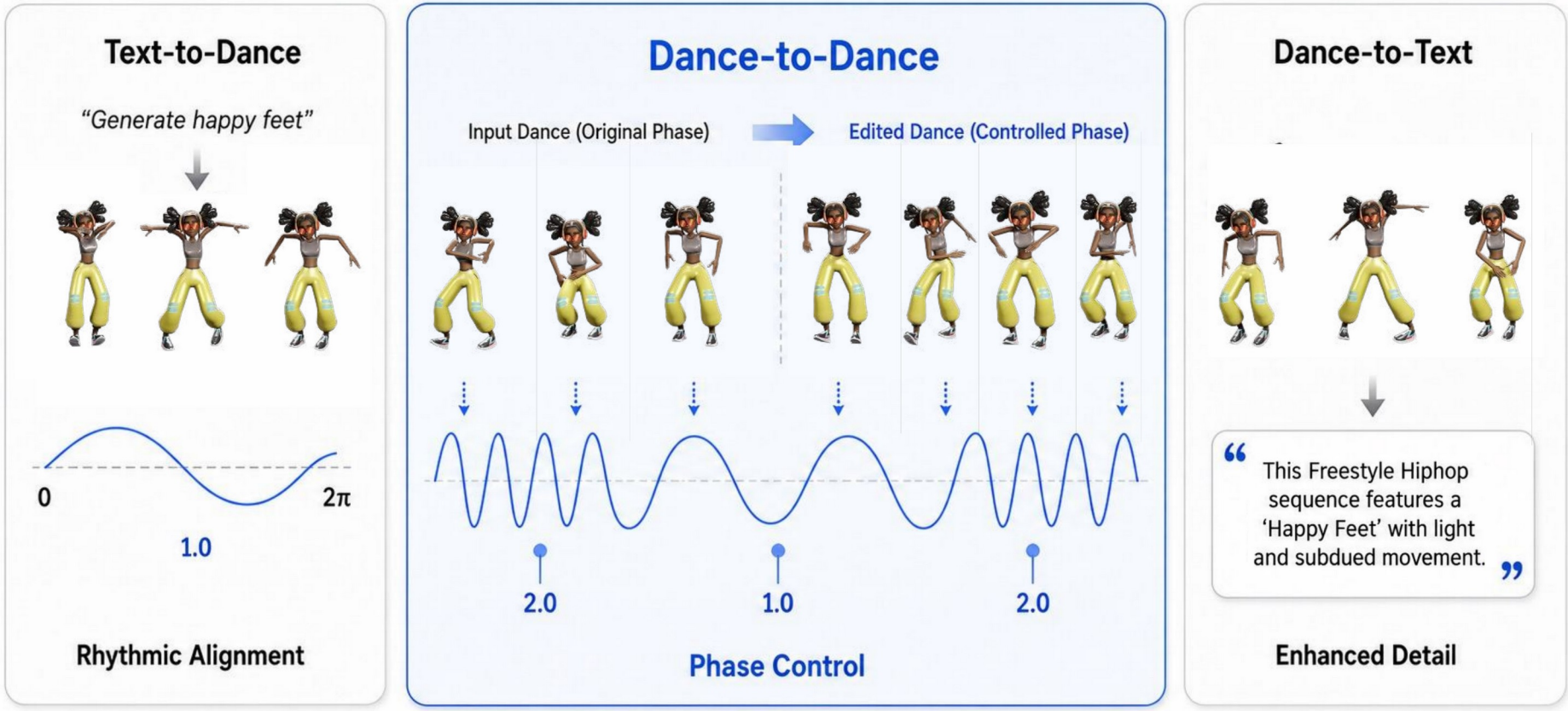
- **Existing view: beat matching.** Music-to-dance models align motion and music beat by beat. This captures local synchrony but not how rhythm is composed.
- **Music is periodic.** Choreography is laid out in beat units on an 8-count grid, so movement repeats with the period of the music.
- **Hitting every beat is not required.** Hip-hop *layback* and ballet *rubato* break one-to-one beat alignment, yet still feel musical because the period is kept.
- **Genre labels are coarse.** “Hip-hop” says nothing about how vigorous, smooth, or grounded a dance is.

Our view

Dancing to the music means matching its **period**, not coinciding with each beat.

→ Model dance on a periodic **phase** aligned to the music, and describe it with a vocabulary paired with music features.

PhaseDance at a Glance



A deterministic phase signal aligned with the music's period provides rhythmic alignment in **Text-to-Dance**, serves as a control knob in **Dance-to-Dance**, and supports detailed movement descriptions in **Dance-to-Text**.

Music-Aligned Kinematic Phase

A deterministic functional **$\Phi(J) = \varphi$** maps a joint motion sequence to a continuous phase aligned to the music's beat structure.

1 Kinetic curve

Mean inter-frame displacement over 22 SMPL joints, median + Gaussian smoothing → v_t.

2 Beat chain

Dynamic programming over local minima of v_t selects the chain whose period matches the music tempo; multipliers $m \in \{1, 2, ½, ¼, 4\}$ absorb half- / double-time.

3 Phase

S_t sweeps $0 \rightarrow 2\pi$ linearly within each beat segment; $\varphi_t = (\cos S_t, \sin S_t)$ on the unit circle.

No learned component, no encoder training. Beat-chain extraction succeeds on 100% of clips in the dance corpora.

The phase encodes where each frame sits within the beat cycle.

LMA-Grounded Dance Description

- **15 descriptors** derived from Laban Movement Analysis, each computed in closed form from a 22-joint motion sequence.
- **Paired with music.** Every descriptor tracks an audio feature with a shared meaning.
- **Four caption layers: genre · movement · description · texture**.
- **One vocabulary, both directions.** The same captions are the input of Text-to-Dance and the output of Dance-to-Text.

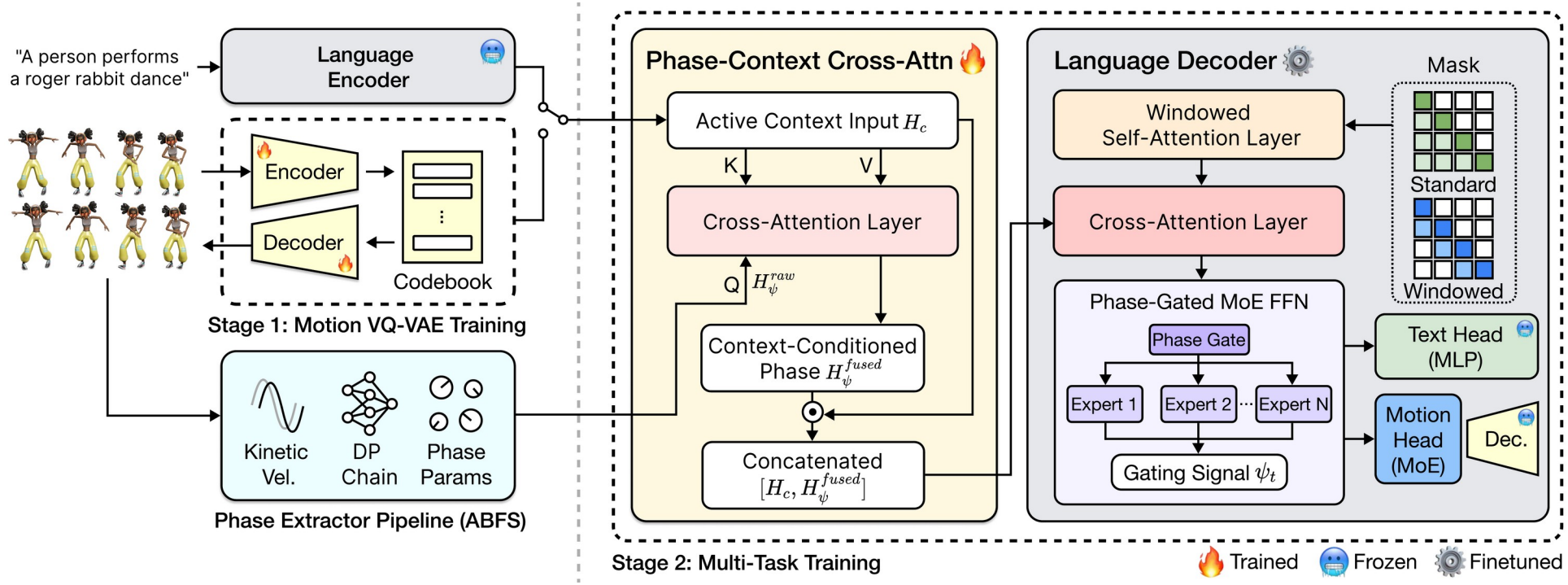
Motion descriptor	Music feature	Shared meaning
Vigor	RMS energy	Energy level
Accent density	Onset density	Rhythmic density
Groove depth	Swing ratio	Rhythmic depth
Movement tempo	Tempo (BPM)	Speed

Four of the fifteen motion–music pairs.

Example caption (House)

"In this house sequence, the dancer performs a farmer with punchy quality. Alternating hops transfer weight from one foot to the other …"

PhaseDance Architecture



- **Interpreted pathway.** Phase-Context Cross-Attention (PCCA) lets phase query the active context (text or motion).
- **Literal pathway.** Raw phase enters at three decoder depths: input embedding, phase-gated MoE FFN, multi-expert output head.
- **Windowed attention.** Each token attends to itself and the 4 preceding tokens.
- **Four tasks, one decoder.** Text-to-Dance, Dance-to-Text, Dance Prediction, Dance In-betweening.

Results: Text-to-Dance on AIST++

Method	FID_k↓	FID_g↓	BPS↑	Pearson↑	Spearman↑
Ground Truth	17.10	10.60	0.86	1.00	1.00
Bailando (music)	28.16	9.62	0.65	0.39	0.49
TM2D (music)	23.94	9.53	0.68	−0.47	−0.53
LMM (music)	22.08	21.97	0.70	−0.23	−0.36
UDE (music)	17.25	8.69	0.67	0.14	0.10
UniMuMo (text)	--	--	--	0.18	0.32
Ours	**24.53**	**11.37**	**0.67**	**0.76**	**0.52**
Ours (oracle phase)	**22.61**	**9.03**	**0.73**	**0.47**	**0.73**

Music-conditioned baselines are shown for reference; "--": not reported.

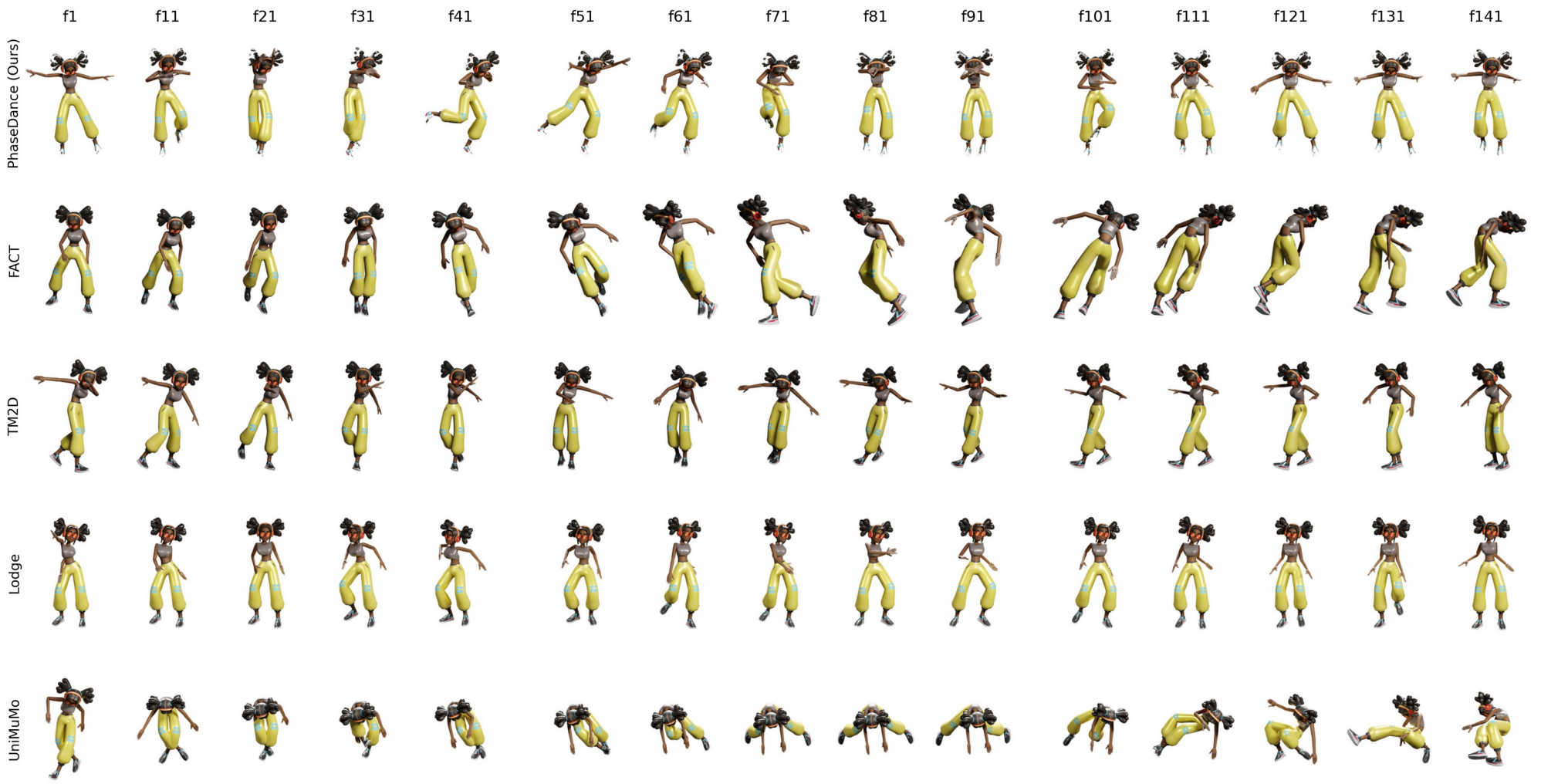
Beyond beat matching

BPS: tempo agreement between motion and music.

Pearson / Spearman: agreement of the per-genre BAS profile with the ground truth — does each genre keep its own rhythmic identity?

Best per-genre rhythm profile (Pearson 0.76) from text alone, without any audio input.

Qualitative Comparison: Text-to-Dance (Break)



Rows: PhaseDance (ours), FACT, TM2D, Lodge, UniMuMo. Frames sampled every 10 steps.

Results: Dance Prediction and In-betweening

	Dance Prediction				Dance In-betweening			
Model	**FID↓**	**DIV→**	**ADE↓**	**FDE↓**	**FID↓**	**DIV→**	**ADE↓**	**FDE↓**
Ground Truth	--	2.320	--	--	--	2.301	--	--
EDGE	8.936	1.954	5.495	5.425	5.086	1.484	5.156	5.508
Lodge	1.949	1.403	1.917	2.156	2.136	1.382	1.875	1.934
Bailando	0.973	2.552	4.930	4.999	0.612	1.936	4.952	5.019
MoMask	8.731	2.599	4.154	5.595	9.808	2.600	3.846	5.162
LMM	5.379	2.126	6.547	8.856	3.393	2.357	6.197	7.230
Ours	**2.077**	**2.315**	**2.952**	**4.127**	**1.534**	**2.392**	**2.735**	**3.070**
Ours (constant phase)	6.248	3.414	3.062	3.726	4.737	3.419	2.731	3.065
Ours (no phase)	3.190	1.804	4.307	9.591	3.291	1.652	2.805	3.529

- **Ours** keeps diversity closest to the ground truth on Dance Prediction (2.315 vs. 2.320); Lodge has lower trajectory error but much lower diversity (1.40).
- **Removing the phase hurts:** no phase raises FDE 4.13 → 9.59; a constant phase raises FID 2.08 → 6.25.

The model relies on a varying phase, so phase can act as an explicit rhythmic control.

Phase as a Control Knob (Dance-to-Dance)



Same motion prefix; phase supplied at 0.5× / 1.0× / 1.5× / 2.0× a nominal BPM.

Changing only φ yields rhythmically distinct continuations.

Results: Dance-to-Text on AIST++

70.58	62.84	23.93
CIDEr	**ROUGE-L**	**BLEU-4**
UniMuMo 12.80	UniMuMo 33.30	UniMuMo 7.20

- Compared with UniMuMo, the only released motion-to-text model for dance.
- Layered LMA captions name the genre, the move, and its quality, where UniMuMo falls back to generic motion descriptions.

GT: Lock · skeeter rabbit

Ours In this lock sequence, the dancer performs a skeeter rabbit with calm quality. One knee lifts and the lower leg kicks forward while the supporting foot hops.

UniMuMo The motion is that the person is doing the waltz dance.

GT: Pop · roll

Ours The ribcage shifts side to side and front to back in an isolated continuous loop. The movement quality is light.

UniMuMo The motion is that a person slowly walked forward in left direction.

Ablation Study (Text-to-Dance)

Variant	FID_k↓	FID_g↓	Div_k→	Div_g→
Ground Truth	17.10	10.60	8.19	7.45
PhaseDance (full)	**24.53**	**11.37**	**6.30**	**7.14**
Multi-task → single-task	27.77	16.72	4.41	7.51
Windowed attn (W = 4) → full causal	36.98	28.23	4.40	8.19
Multi-expert head → standard proj.	38.83	20.62	4.78	7.95
Phase-gated MoE → single FFN	55.06	33.26	2.75	6.09
PCCA → simple concatenation	80.44	55.01	1.71	3.04

Each variant replaces one component and is retrained from scratch.

Every component contributes; replacing PCCA with concatenation hurts the most.

Do the Rhythm Metrics Match Human Perception?

Study A · 29 viewers

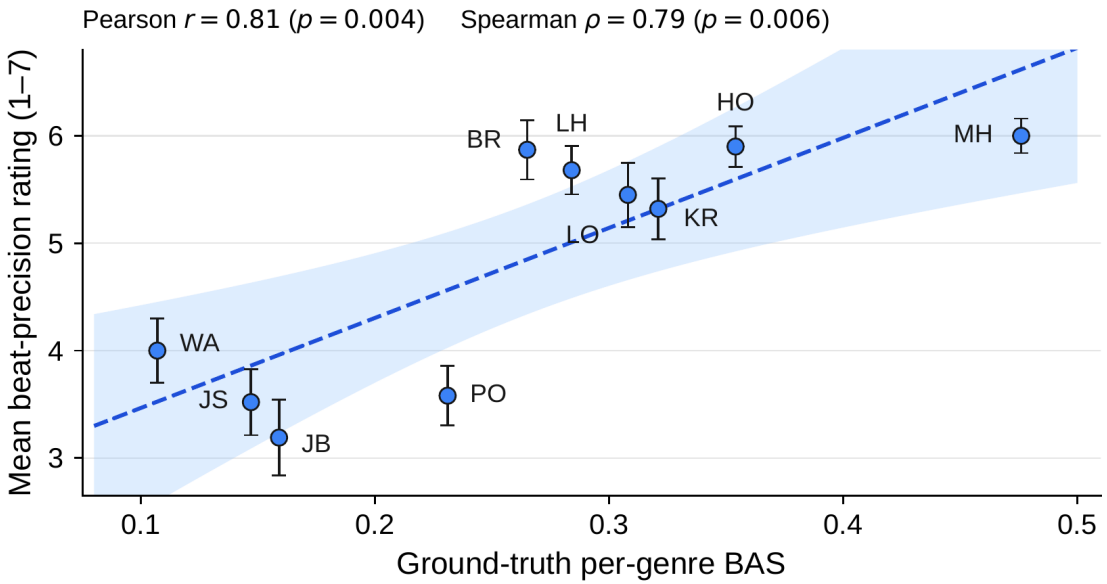
Same motion, different audio. When BAS and BPS disagree, viewers side with BPS:

111 vs. 3

Study B · 31 raters

Human beat-precision ratings track the ground-truth per-genre BAS profile:

Pearson r = 0.81



Study B: each point is one AIST++ genre (mean ± s.e.m.).

Conclusion

- **PhaseDance** is a phase-aware, text-augmented multi-task model for dance.
- **Deterministic, music-aligned phase** captures periodicity rather than beat coincidence and acts as a rhythmic control knob.
- **LMA-grounded captions** paired with music features describe expressivity beyond genre labels.
- **One decoder, four tasks:** Text-to-Dance, Dance-to-Text, Dance Prediction, Dance In-betweening.
- **Rhythm metrics** that agree with human perception in two user studies.

Project page & code
meongeun.github.io/phasedance

Thank you!