Timing In stand-up Comedy: Text, Audio, Laughter, Kinesics (TIC-TALK): Pipeline and Database for the Multimodal Study of Comedic Timing

Yaelle Zribi[1] **Florian Cafiero**[1,2] **Vincent Lépinay**[3] **Chahan Vidal-Gorène**[1]

[1]Centre Jean-Mabillon, École nationale des chartes, PSL, Paris, France
[2]Laboratoire de Recherche, EPITA, Paris, France
[3]médialab, Sciences Po, Paris, France
{yaelle.zribi, florian.cafiero, chahan.vidal-gorene}@chartes.psl.eu
vincent.lepinay@sciencespo.fr

Abstract

Stand-up comedy, and humor in general, are often studied through their verbal content. Yet live performance relies just as much on embodied presence and audience feedback. We introduce TIC-TALK, a multimodal resource with 5,400+ temporally aligned topic segments capturing language, gesture, and audience response across 90 professionally filmed stand-up comedy specials (2015–2024). The pipeline combines BERTopic for 60 s thematic segmentation with dense sentence embeddings, Whisper-AT for 0.8 s laughter detection, a fine-tuned YOLOv8-cls shot classifier, and YOLOv8s-pose for raw keypoint extraction at 1 fps. Raw 17-joint skeletal coordinates are retained without prior clustering, enabling the computation of continuous kinematic signals—arm spread, kinetic energy, and trunk lean—that serve as proxies for performance dynamics. All streams are aligned by hierarchical temporal containment without resampling, and each topic segment stores its sentence-BERT embedding for downstream similarity and clustering tasks. As a concrete use case, we study laughter dynamics across 24 thematic topics: kinetic energy negatively predicts audience laughter rate ($r = -0.75$, $N = 24$), consistent with a stillness-before-punchline pattern; personal and bodily content elicits more laughter than geopolitical themes; and shot close-up proportion correlates positively with laughter ($r = +0.28$), consistent with reactive montage.

1 Introduction

Humor is one of the most complex forms of human communication, involving timing, embodiment, and interaction. Stand-up comedy, in particular, constitutes an ideal case study for modeling how verbal, acoustic, and visual cues align to produce shared affective meaning.

The performer's gestures, pauses, and interaction with audience laughter are essential components of meaning and rhythm. Modeling these elements jointly raises both technical and conceptual challenges: how can we represent and evaluate *comedic timing* computationally? We take *comedic timing* to denote short-lag coordination across text, gesture, and audience response in live delivery.

Multimodal modeling of live performance is still mostly missing. Computational humor has largely centered on *textual* humor and its assessment (Kalloniatis and Adamidis, 2025): detecting humorous passages (Weller and Seppi, 2019; Annamoradnejad and Zoghi, 2024; Yang et al., 2015; Cafiero and Puren, 2025; Hossain et al., 2020), ranking by funniness (Potash et al., 2017), and predicting offensiveness or controversy (Meaney et al., 2021).

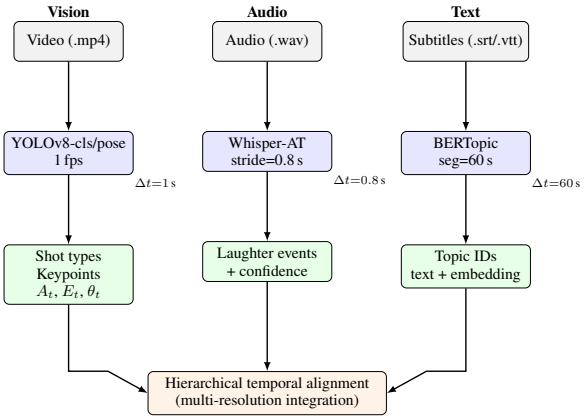

Figure 1: Multimodal processing pipeline. Each modality retains its native temporal resolution. Signals are merged by hierarchical temporal containment without resampling. Topic segments also store 384-dim sentence-BERT embeddings (`all-MiniLM-L6-v2`).

Distant viewing offers a framework to tackle the conceptual and technical challenge to simulate human vision for artistic analysis (Arnold and Tilton, 2019). Affective and social computing provide tools for modeling prosody, gesture, and emotion, though primarily in human–machine interaction.

Movement analysis has been applied to iconography (Impett and Moretti, 2017), ergonomic gesture learning (Glushkova et al., 2023), and pose clustering in archival theater footage (Rau et al., 2023) but without linking gesture to speech or audience response.

Recent work explores how language models might support stand-up writing, raising questions of prompt design, model bias, and cultural framing (Mirowski et al., 2024). The StandUp4AI dataset (Barriere et al., 2025) adds multilingual laughter labels, but focuses on

audio.

Beyond text-based generation, live performance has been identified as a critical setting for evaluating computational humor under real-world temporal and social constraints. Mirowski et al. (2025) argue that improvised and staged comedy offer unique testbeds for studying the interaction between human performers, AI systems, and audiences, emphasizing that timing, embodiment, and audience feedback are inseparable from humor evaluation. While our focus differs from these experimental setups, we contribute a reproducible multimodal dataset of stand-up performances, designed to operationalize these performative dimensions (timing, gesture, and laughter) as measurable signals for computational analysis.

Recent work has similarly proposed an automatic method for the analysis of stand-up comedy, with particular attention to timing and performance dynamics. Using repeated recordings of the same routines performed on different nights by two comedians, Pope et al. (2026) analyze timing structures in live comedy through matched speech sequences across performances, demonstrating that the apparent spontaneity of stand-up relies on a complex craft of pacing and adaptation to audience context. Their study shares our interest in timing, performance structure, and audience response, but differs in both scale and design: it focuses on prosody, speech, and laughter placement in unedited live performances, whereas we analyze multimodal coordination within a large corpus of professionally recorded and edited stand-up comedy specials, examining themes, laughter placement, as well as the visual and bodily dimensions of performance through shot composition and pose-derived movement signals. Taken together, these contrasting approaches underscore how central yet elusive timing remains in the analysis of stand-up comedy.

In this paper, we present TIC-TALK: a multimodal corpus of 90 Netflix stand-up specials and a documented processing pipeline aligning text, audio, and vision without resampling. We report available performance indicators for the shot classifier and descriptive coverage metrics for the other streams, and include corpus-level cross-modal findings—notably a negative relationship between kinetic energy and laughter rate across topics—that validate the temporal alignment.

Our main contributions are:

1. A reusable multimodal **corpus** of 90 professionally edited stand-up specials with temporally aligned text, audio, and vision streams, including raw pose keypoints and per-segment sentence-BERT embeddings;

2. A transparent, reproducible **processing pipeline** (BERTopic, sentence-BERT, Whisper-AT, YOLOv8-cls, YOLOv8s-pose) with documented training choices and performance indicators;

3. **Three kinematic signals** (A_t, E_t, θ_t) derived from raw skeletal coordinates; and a **cross-modal use case** (Section 3.3) on laughter dynamics across 24 thematic topics, demonstrating: (i) kinetic energy is the strongest kinematic predictor of laughter ($r = -0.75$, $N = 24$), consistent with a stillness-before-punchline pattern; (ii) personal/bodily themes elicit systematically more laughter than geopolitical content; (iii) belly laughs are virtually absent at topic granularity, motivating event-level annotation; and (iv) shot close-up proportion correlates weakly with laughter rate ($r = +0.28$), consistent with reactive montage;

4. A **short-horizon laughter onset prediction benchmark** (Section 3.4): given multimodal context up to t, predict whether a new laughter event will begin in $[t, t+2)$; ablation over five feature sets (N=285,916 anchors, 90 shows, show-level train/val/test split) shows that temporal laughter history dominates prediction (AUROC = 0.643), that multimodal fusion achieves the best performance (AUROC = 0.647, AUPRC = 0.277 vs. random baseline 0.170), and that shot and pose features contribute marginal but consistent gains.

We first detail the processing pipeline per modality and the temporal alignment strategy (Section 2), then describe the corpus as a reusable resource and its data structure (Section 3). Section 3.3 presents a descriptive use case—laughter dynamics across thematic topics; Section 3.4 presents a predictive use case—short-horizon laughter onset prediction. Limitations and biases are discussed in Section 4.

2 Processing Pipeline and Temporal Alignment

The dataset integrates four modality streams—text, audio, and vision (pose and shot)—each processed independently at its native temporal resolution before alignment into a unified hierarchical representation (Figure 1).

2.1 Audio: Laughter Detection

Laughter events were detected using **Whisper-AT**, a pretrained AudioSet-based audio tagging model (Gong et al., 2023). Inference was performed at a 0.8 s stride in a high-recall configuration. The model outputs class probabilities for multiple laughter types; contiguous positive windows were merged into continuous events. Each event is represented by start and end times in seconds, a label (`type`), and a confidence score. These events constitute high-resolution temporal anchors for audience response.

2.2 Text: Topic Segmentation

2.2.1 Data and time-based segmentation

We start from subtitle transcripts in `.srt` format, parsed with their start/end timestamps and normalized by removing markup and formatting codes, standardizing

apostrophes, and collapsing whitespace. We then construct contiguous time blocks by concatenating consecutive subtitle lines until a target duration is reached; a new block starts at the next subtitle line whose start time exceeds the current block limit. We remove stopwords using the union of a standard English stopword list and a curated set targeting fillers/discourse markers.

2.2.2 Sentence embeddings

Each block is embedded with a sentence-transformer encoder (`all-MiniLM-L6-v2`). Embeddings are computed in batches and L2-normalized: for each block i we obtain an embedding $\mathbf{e}_i \in \mathbb{R}^d$ and set

$$\tilde{\mathbf{e}}_i = \frac{\mathbf{e}_i}{\|\mathbf{e}_i\|_2}.$$

These normalized vectors are used both for training the topic model and for later topic assignment to 60-second blocks.

2.2.3 Topic modeling with BERTopic

We learn topics using BERTopic. We first apply UMAP to the normalized embeddings using cosine distance, with $n_{\text{neighbors}} = 15$, $n_{\text{components}} = 5$, and $min\_dist = 0.1$. The reduced representations are clustered with HDBSCAN ($min\_cluster\_size = 15$, $min\_samples = 5$). Topic representations are obtained from a unigram–bigram count vectorizer. The model is capped at 40 topics and we retain the top 10 words per topic for interpretation.

2.2.4 Model selection over training block size

Topic quality depends on the training granularity. We select the training block size from $\{120, 150, 180, 210, 240\}$ seconds. For each candidate size, we train a BERTopic model and compute three diagnostics based on the resulting topic assignments:

- the *number of discovered topics* K,
- the *largest-topic share* $s_{\max}$
- a *normalized topic entropy* H_{norm}.

Let n_k be the number of blocks assigned to topic $k \in \{1, \ldots, K\}$ and $N = \sum_{k=1}^{K} n_k$. Define $p_k = n_k/N$. We compute

$$s_{\max} = \max_{k \in \{1,\ldots,K\}} p_k$$

and

$$H_{\text{norm}} = \frac{-\sum_{k=1}^{K} p_k \log p_k}{\log K}.$$

We enforce two validity constraints to avoid degenerate solutions: $K \geq 10$ and $s_{\max} \leq 0.35$. Among valid candidates, we select the model maximizing a composite score:

$$S = H_{\text{norm}} + C_{\text{NPMI}} - 2\, s_{\max},$$

where C_{NPMI} is a topic coherence measure computed from the model's top words.

Topic coherence (NPMI). We compute coherence using normalized pointwise mutual information (NPMI) over a tokenized version of the preprocessed blocks (subsampling documents when necessary for efficiency). For a topic t, let W_t be its top-M words (here $M = 10$). For each pair $(w_i, w_j) \in W_t \times W_t$ with $i < j$, let $P(w)$ be the probability that a word appears in a document and $P(w_i, w_j)$ the probability that both appear in the same document (estimated from document co-occurrence counts). NPMI for a pair is

$$\text{NPMI}(w_i, w_j) = \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)}.$$

Topic coherence is the average over word pairs, and corpus-level coherence averages over topics:

$$C_{\text{NPMI}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left(\frac{2}{M(M-1)} \sum_{i<j} \text{NPMI}(w_i, w_j) \right),$$

where $\mathcal{T}$ is the set of non-outlier topics.

2.2.5 Topic assignment to 60-second segments

After selecting the training block size, we retrain the topic model on all blocks at that granularity. We then construct 60-second blocks for each show and assign a topic label to each block using the trained BERTopic model. Since HDBSCAN can label uncertain points as outliers (-1), we apply a three-step post-processing procedure to reduce outliers (1) BERTopic outlier reduction using an embedding-based strategy followed by a c-TF–IDF-based strategy, (2) reassignment of remaining outliers by nearest topic centroid in embedding space, and (3) within-show temporal gap filling: if a single outlier is flanked by two identical non-outlier topics, it is replaced by that topic.

For the centroid step, we compute a centroid for each non-outlier topic k by averaging the normalized training embeddings assigned to k and re-normalizing:

$$\mathbf{c}_k = \frac{\frac{1}{|I_k|} \sum_{i \in I_k} \tilde{\mathbf{e}}_i}{\left\| \frac{1}{|I_k|} \sum_{i \in I_k} \tilde{\mathbf{e}}_i \right\|_2},$$

where I_k is the set of training blocks assigned to topic k. For an outlier segment with embedding $\tilde{\mathbf{e}}$, we compute cosine similarities $\tilde{\mathbf{e}}^\top \mathbf{c}_k$ and reassign it to $\arg\max_k \tilde{\mathbf{e}}^\top \mathbf{c}_k$ when the maximum similarity is at least 0.30.

This procedure yields a topic sequence at 60-second resolution for each show, together with interpretable topic descriptors derived from c-TF–IDF.

2.3 Vision: Model Training and Feature Extraction

The visual pipeline combines two complementary YOLOv8 models operating hierarchically: (1) a **YOLOv8-cls** network fine-tuned for shot classification, and (2) a pre-trained **YOLOv8s-pose** estimator used to

extract body keypoints from full-body frames only (Maji et al., 2022). This design ensures both efficient processing and consistent framing for pose analysis.

Shot classification. Fine-tuned to recognize six shot types: *full shot, medium close-up, medium long shot, medium shot, other angles,* and *other*. The model was trained on 594 manually annotated frames and validated on 128 held-out samples (100 epochs, batch size 32, learning rate 1×10^{-3}). Validation yielded an average F1 = 0.91, with most confusion between adjacent framings (e.g., chest $\leftrightarrow$ waist). Predicted shot labels later serve both as contextual information and as filters for pose extraction, keeping only full-body and frontal views.

Pose estimation. Filtered frames are processed by the pre-trained **YOLOv8s-pose** model, producing 17 body keypoints (COCO skeleton) per detected performer at 1 fps. Raw pixel-normalized coordinates for all 17 joints are stored without prior discretization, preserving the full geometric information for downstream analysis. Joints with no detection confidence are recorded as $(0, 0)$ and excluded from derived computations via a validity filter.

Kinematic signals derived from raw keypoints. Raw keypoints enable the computation of three continuous scalar signals per frame, each capturing a distinct dimension of performance dynamics. Let $\mathbf{p}_j(t)$ denote the 2D coordinates of joint j at time t, and let $J(t)$ be the set of valid joints at t.

Arm spread measures the lateral extension of the performer's gesture relative to shoulder width:

$$A_t = \frac{\|\mathbf{p}_{\mathrm{W}_1}(t) - \mathbf{p}_{\mathrm{W}_2}(t)\|}{\|\mathbf{p}_{\mathrm{S}_1}(t) - \mathbf{p}_{\mathrm{S}_2}(t)\|},$$

where $\mathrm{W}_1, \mathrm{W}_2$ are the wrists and $\mathrm{S}_1, \mathrm{S}_2$ the shoulders. A_t=1 corresponds to a neutral stance; A_t>2 indicates open or emphatic gestures.

Kinetic energy quantifies total body movement between consecutive frames, normalized by performer height (bounding-box height h):

$$E_t = \frac{1}{h} \sum_{j \in J(t) \cap J(t-1)} \|\mathbf{p}_j(t) - \mathbf{p}_j(t-1)\|.$$

This serves as a proxy for performance intensity, capturing transitions between high-agitation delivery and still, high-confidence pauses.

Trunk lean encodes the signed angle of the torso axis relative to vertical:

$$\theta_t = \arctan\left(\frac{x_{\text{hip}}(t) - x_{\text{sho}}(t)}{y_{\text{hip}}(t) - y_{\text{sho}}(t)}\right) \times \frac{180}{\pi},$$

where $x_{\text{sho}}, y_{\text{sho}}$ and $x_{\text{hip}}, y_{\text{hip}}$ are the midpoints of the shoulder and hip pairs respectively. Lateral leans may, for instance, be characteristic of character-mimicry and asides, providing a posture-level complement to motion energy.

All three signals are smoothed with a sliding window of 30 s to suppress frame-level noise before analysis.

2.4 Hierarchical Temporal Alignment

Modalities operate on distinct temporal resolutions:

$$\Delta t_{\text{laugh}} = 0.8\,\text{s}, \quad \Delta t_{\text{pose}} = 1\,\text{s},$$

$$\Delta t_{\text{shot}} = 1\,\text{s}, \quad \Delta t_{\text{topic}} = 60\,\text{s}.$$

To preserve native granularity, temporal alignment is performed through hierarchical containment rather than resampling.

Let each modality m produce a sequence of temporal events or segments. Topic segments serve as the *anchor* level: each topic block $b_j = [s_j, e_j)$ defines a temporal window into which higher-frequency events are assigned by strict containment:

$$e \text{ is assigned to } b_j \iff t_e \in [s_j,\, e_j),$$

where t_e is the event timestamp. This applies uniformly to all nested streams: pose keyframes (Δt=1 s), shot labels (Δt=1 s), and laughter events (Δt=0.8 s). Kinematic signals (A_t, E_t, θ_t) are derived from the raw keypoints within each block after alignment. Each topic segment also stores the 384-dimensional sentence-BERT embedding $\tilde{\mathbf{e}}_j$ computed during topic modeling (Section 2.2.5), enabling direct use for similarity retrieval or cross-show clustering without re-encoding.

3 Corpus Output and Statistics

Following the pipeline in Section 2, we release only *derived, time-aligned annotations* for 90 stand-up performances. No audio, image or video is distributed.

3.1 Delivered outputs

- **Topics**: 60 s segments with topic id, aggregated text, and sentence-BERT embedding.
- **Laughter**: contiguous events with start/end times, type label, and confidence score.
- **Shots**: one label per frame (1 Hz) from YOLOv8-cls.
- **Poses**: raw 17-joint (x, y) coordinates at 1 fps with bounding-box dimensions; no prior clustering.

3.2 Summary statistics

Average runtime per show is 63 min (total ≈94 h). The unified dataset contains 5,416 topic segments across 90 shows, each storing a 384-dim embedding. The visual stream covers 322,973 frames at 1 fps; 22% are full-body frames yielding raw keypoint sequences for pose analysis. Text yields ≈3,100 one-minute segments with non-outlier topic assignments.

3.3 Cross-modal Analysis: Laughter Dynamics as a Use Case

We ask whether thematic content, kinematic profile, and shot composition co-vary with audience laughter across the 24 BERTopic topics and 5,416 aligned blocks.

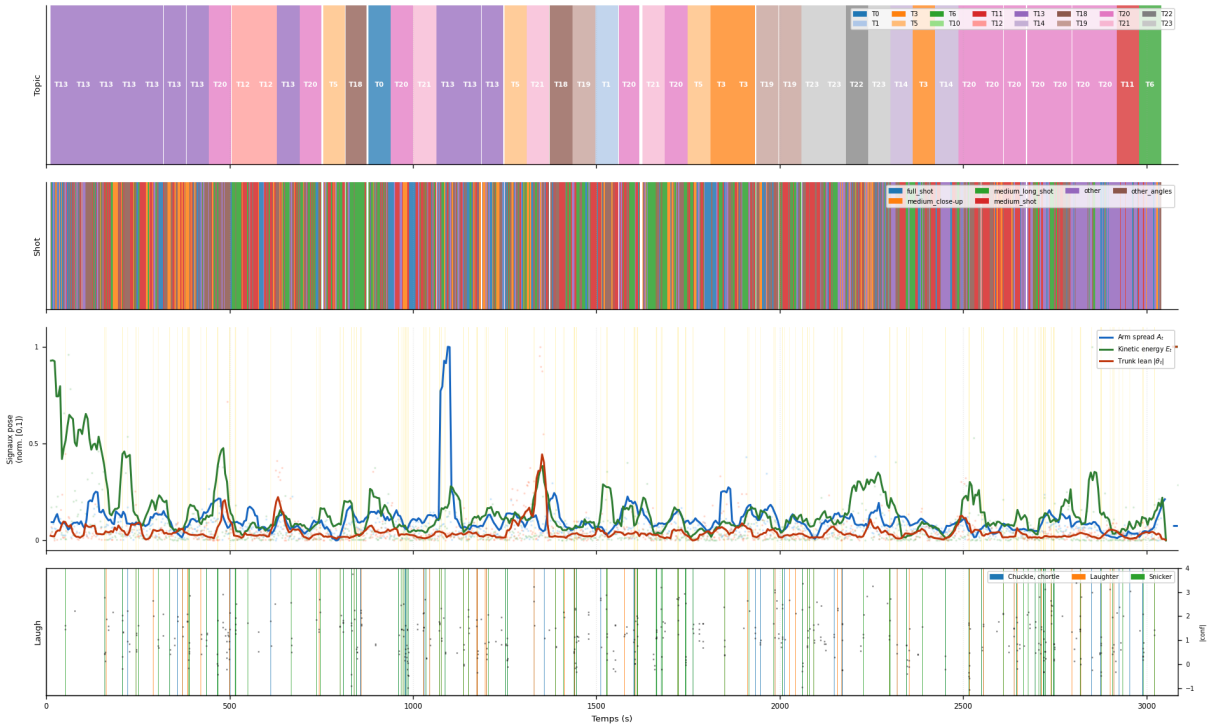

Figure 2: Example of aligned multimodal timelines for a show. Panels show from top to bottom: topic segments (BERTopic, 60 s blocks), shot-type predictions (1 Hz), three kinematic signals derived from raw pose keypoints (arm spread A_t, kinetic energy E_t, trunk lean θ_t, each normalized to $[0, 1]$; gold shading = laughter windows), and laughter events with confidence. The dense kinematic activity and prevalence of full-body shots reflect André's chaotic performance style; the alignment pipeline captures this without resampling across modalities.

Method. For each topic block, we compute (i) the laughter rate r_ℓ, defined as the share of block duration covered by detected laughter events (mean coverage: 17.8%, ≈1.2 events/10 s); (ii) per-frame kinematic features (kinetic energy E_t, arm spread A_t, trunk lean θ_t); and (iii) shot-type proportions (full-body, close-up, medium). These are averaged per topic, weighted by block count, to obtain topic-level profiles. Pearson correlations are then computed between each feature and the mean laughter rate $\bar{r}_\ell$ across the 24 topics. Figure 3 presents the complete 24 topics × 10 features matrix as a hierarchical clustermap; Table 1 reports the six most contrasted topics.

Topic	n	$\bar{r}_\ell$	$\bar{E}_t$	$\bar{A}_t$
T15: body / dress / boobs	97	.253	1.13	1.40
T22: baby / abortion	198	.248	1.05	1.28
T10: food / want / chef	210	.230	1.19	1.30
T3: indian / trump / india	197	.141	1.27	1.32
T1: iceland / icelandic	50	.085	1.27	1.40
T6†: sap / mae / hello	81	.051	**2.24**	1.34

Table 1: Six most contrasted topics by mean laughter rate $\bar{r}_\ell$ (share of block duration covered by laughter events). $\bar{E}_t$: mean kinetic energy (normalized joint displacement between consecutive frames). $\bar{A}_t$: mean arm spread (wrist-to-wrist / shoulder-to-shoulder ratio). †T6 is a structural artefact; see Section 4.

Finding 1 — Kinetic energy negatively predicts laughter rate ($r = -0.75$). The strongest cross-modal signal across the 24 topics — after the quasi-tautological event count — is a *negative* correlation between mean kinetic energy $\bar{E}_t$ and laughter rate: Pearson $r = -0.75$, $N = 24$. Topics with the highest laughter rates all exhibit markedly low kinetic energy: T15 (body/dress, $\bar{E}_t = 1.13$, $\bar{r}_\ell = 0.253$), T22 (baby/abortion, $\bar{E}_t = 1.05$, $\bar{r}_\ell = 0.248$), T10 (food, $\bar{E}_t = 1.19$, $\bar{r}_\ell = 0.230$), all below the corpus mean of 1.31. The low-laughter end is anchored by T11 (city/tonight/jewish, $\bar{E}_t = 1.65$, $\bar{r}_\ell = 0.121$) and the artefactual T6 ($\bar{E}_t = 2.24$, $\bar{r}_\ell = 0.051$).

This pattern is consistent with a *stillness-before-punchline* hypothesis: during high-laughter delivery, performers may reduce body movement and stabilize their posture to concentrate audience attention on the verbal content.

Finding 2 — A thematic hierarchy of funniness. Topic content stratifies audience laughter systematically. The four topics with the highest laughter rates ($\bar{r}_\ell > 0.20$) are all personal and bodily in register: physical appearance (T15, $\bar{r}_\ell = 0.253$), reproductive transgression (T22, $\bar{r}_\ell = 0.248$; also the highest *has_laughter* rate across the corpus at 0.869), everyday life (T10, $\bar{r}_\ell = 0.230$), and romantic relationships (T17,

Listing 1: Excerpt from the unified dataset (V2 structure).

```
{
  "ID_1": {
    "metadata": {
      "show_id": "1",
      "n_blocks": 62,
      "embedding_dim": 384,
      "keypoint_joints": ["Nez", "Epaule_1", "...
          ", "Cheville_2"]
    },
    "timeline": [{
      "block_id": 58,
      "start": 3480.0, "end": 3540.0,
      "topic_id": 6,
      "text": "marriage gender roles ...",
      "embedding": [0.021, -0.143, "..."],
      "laugh_events": [
        {"start": 3482.4, "end": 3485.6,
         "type": "laughter", "confidence": 0.92}
      ],
      "pose_keypoints": [{
        "time": 3483.0, "has_detection": true,
        "bbox": {"xmin": 412, "ymin": 28,
                 "xmax": 895, "ymax": 716},
        "keypoints": {
          "Epaule_1": [634.2, 182.5],
          "Poignet_1": [710.8, 480.3], "...": []
        }
      }],
      "shot_events": [
        {"time": 3483.0, "label": "full_shot",
         "class_id": 3, "score": 0.97}
      ]
    }]
  }
}
```

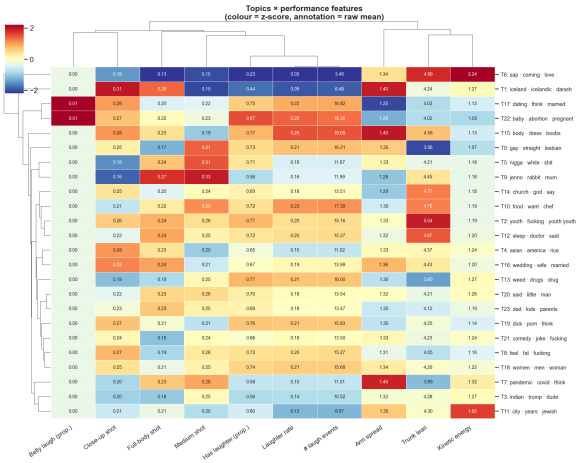

Figure 3: Hierarchical clustermap of 24 BERTopic topics × 10 performance features (z-scored row-wise). Colour encodes standardized deviation from the per-feature mean. The dominant pattern separates a low-$\bar{E}_t$ / high-$\bar{r}_\ell$ cluster (top: T15, T22, T10 — personal/bodily topics) from a high-$\bar{E}_t$ / low-$\bar{r}_\ell$ cluster (bottom: T6, T11). T6 (sap/mae/hello) is a structural artefact corresponding to subtitle encoding markers and on-stage entry sequences; it should be excluded from content-level comparisons.

$\bar{r}_\ell = 0.222$). By contrast, geopolitical and identity-framing topics generate substantially less laughter: T3 (trump/india, $\bar{r}_\ell = 0.141$) and T1 (iceland, $\bar{r}_\ell = 0.085$, 50 blocks concentrated in a single show). This replicates content-level funniness gradients documented in text-only humor classification (Yang et al., 2015; Annamoradnejad and Zoghi, 2024), anchoring them in live audience response at corpus scale.

Finding 3 — Belly laughs are quasi-absent at topic granularity. The deepest laughter category (*belly laugh*, AudioSet class 20) is effectively absent across all 24 topics: only T22 ($\hat{r}_{\text{belly}} = 0.0051$) and T17 ($\hat{r}_{\text{belly}} = 0.0061$) register non-zero values. Two non-exclusive explanations apply: (i) the Whisper-AT classifier may be conservative on this class (it represents fewer than 2 events per 74,000+ inference windows across the corpus); (ii) belly laughs are genuinely triggered by specific delivery moments rather than by sustained thematic content, making them invisible at 60 s block granularity. Either interpretation suggests that capturing deep, distinctive laughter requires event-level annotation at a finer temporal resolution.

Finding 4 — Shot composition and reactive montage ($r = +0.28$). Close-up shot proportion shows a weak positive correlation with laughter rate ($r = +0.28$, $N = 24$). High-laughter topics T15 and T22 both display above-average close-up ratios (0.278 and 0.265 respectively), consistent with *reactive montage*: directors increase close-up coverage during high-laughter passages, foregrounding the performer's facial expression at punchline delivery. A notable exception is T1 (iceland/icelandic), which exhibits the highest close-up ratio in the corpus (0.315) despite a low laughter rate ($\bar{r}_\ell = 0.085$), indicating that this association is partially confounded by performer- and show-level filming conventions.

3.4 Short-horizon Laughter Onset Prediction

Given the multimodal stream up to time t, can we predict whether a new laughter event will begin in the next $\delta = 2$ s?

Task and experimental setup. For each show, we sample anchor points at 1 s steps, excluding moments inside an ongoing laughter event. This yields 285,916 anchors across 90 shows; the positive rate is 17.0% (a new onset occurs in the next 2 s). Shows are split at the group level (GroupShuffleSplit): 62 shows for training, 14 for validation, 14 for test—no show appears in more than one split. A HistGradientBoostingClassifier is trained with balanced sample weights; the decision threshold is tuned on validation by maximizing F1.

Three feature groups are defined. **History** (10 scalars): from laughter events in the past 10 s—event count, rate, coverage, maximum event duration, mean and maximum confidence, coverage in the last 2 s and 5 s, time since last onset, time since last end. **Text** (64 scalars): the current topic block's 384-dim sentence-BERT embedding, reduced to 64 dimensions via PCA fitted on the training set only. **Vision** (20 scalars): shot proportion histogram over 6 classes, shot change rate,

mean shot confidence, and 12 pose scalars (arm spread mean/std/max/trend, trunk lean mean/std, kinetic energy mean/std/max/trend, detection rate)—all aggregated over the past 10 s window.

System	AUROC	AUPRC	F1	Prec	Rec
history-only	0.643	0.275	0.336	0.223	0.682
text-only	0.554	0.197	0.297	0.177	0.926
vision-only	0.538	0.187	0.291	0.178	0.806
text + vision	0.577	0.210	0.300	0.182	0.867
text + vision + history	**0.647**	**0.277**	**0.342**	**0.248**	0.553
Random (AUPRC baseline)	—	0.170	—	—	—

Table 2: Short-horizon laughter onset prediction: ablation over feature groups. Positive rate = 0.170. Test set: 45,894 anchors from 14 held-out shows. AUPRC of a random classifier equals the positive rate. Threshold tuned on validation for F1/precision/recall.

Findings. First, temporal laughter history is by far the strongest individual predictor (AUROC = 0.643), leaving only marginal room for the other modalities: the best multimodal system (text+vision+history) improves AUROC by only 0.004 over history alone. This reflects a *temporal auto-correlation* of audience laughter: a hot room stays hot, independently of what is being said or shown. Second, vision-only is the weakest unimodal system (0.538), but combining it with text (0.577) outperforms both text-only (0.554) and vision-only—a consistent, if small, multimodal synergy. Third, adding text and vision to history improves precision from 0.223 to 0.248 while moderately reducing recall—the full model is less trigger-happy, reducing false positives. Fourth, the overall performance level (AUPRC = 0.277 vs. random 0.170, a $1.6\times$ lift) indicates that laughter onset is predictable above chance from a 10 s window, but far from deterministic: the stochastic nature of audience response and the 60 s temporal granularity of topical context both limit the ceiling of short-horizon prediction.

3.5 Data Structure and Access

Annotations are serialized as a hierarchical JSON per show (Figure 1); each topic block stores its four aligned streams, enabling direct temporal queries without resampling.

3.6 Multimodal Visualization Examples

Figure 2 illustrates the aligned multimodal timeline for one show; analogous visualizations for all 90 specials are distributed with the corpus as an interpretability and consistency check for the alignment pipeline.

4 Discussion

The descriptive use case (Section 3.3) yields four findings. (1) The E_t–laughter anti-correlation ($r = -0.75$) is consistent with a *stillness-before-punchline* pattern, but remains correlational: filming conventions, performer mobility, and the artefactual T6 are plausible confounders; event-level replication (kinetic energy in the 5 s before vs. after laughter onset) would be a stronger test. (2) Personal/bodily topics outperform geopolitical ones on laughter rate, replicating text-only funniness gradients at the level of live audience response and suggesting thematic content is a first-order predictor independently of delivery style. (3) The near-absence of belly laughs at 60 s granularity motivates finer annotation: deep laughter is likely tied to specific delivery moments invisible at block level. (4) The shot composition signal ($r = +0.28$), coherent with reactive montage, is partially confounded by show-level filming conventions.

The predictive use case (Section 3.4) adds three observations. (5) Laughter auto-correlation (history-only AUROC = 0.643) accounts for most predictable variance, consistent with crowd contagion (Provine, 1992). (6) Text and vision contribute marginally to precision (0.248 vs. 0.223), limited by 60 s block granularity; sentence-level and frame-level features would likely yield stronger gains. (7) The modest ceiling (AUPRC = 0.277 vs. 0.170 random) reflects the inherent stochasticity of audience response, absent from our single-recording specials.

Together, these results show that hierarchical temporal alignment enables both descriptive and predictive cross-modal analyses while exposing their respective limits. The resource's value lies in enabling comparable, interpretable measurements of multimodal synchrony at corpus scale rather than absolute claims about funniness. Because all shows are professionally edited Netflix specials, platform conventions are embedded in the signal; comparisons are most reliable within similarly produced shows. Next steps include event-level annotation, prosodic features, and intra-comedian analyses to disentangle performer style from content.

5 Conclusion

This lightweight and modular pipeline provides an effective, reusable, and scalable framework for modeling the multimodal structure of stand-up comedy — a form whose apparent simplicity belies the artistry and craftsmanship of its performers. By operationalizing core dimensions such as gesture, timing, and audience response — a conceptual and technical challenge — and thus offering a model of stand-up performance, it supports both systematic analysis and critical reflection on what remains beyond computation.

Future work. While our pipeline was applied across a wide variety of performers and stand-up comedy specials, future experiments could focus on multiple video recordings by the same comedian, in order to track stylistic evolution over time and test the modularity of performance elements — a relevant hypothesis for stand-up, where long-form shows are often assembled from recombined short routines. In our current setup, editing is treated as part of the artistic object — and while some editing conventions recur, their variability across shows may obscure key aspects of the live performance.

Obtaining recordings from other sources would help generalization from this database. Capturing original data would also open new avenues toward the study of more local, ephemeral, or amateur practices, anchored in specific socio-geographic contexts.

6 Code Availability

All code used for processing and analysis is available at the following anonymous repository: `https://github.com/depotanonyme/16102025`.

7 Acknowledgements

We are thankful to participants at the AV in DH workshop (George Mason University), at the Sciences Po Medialab seminar and at the Bridging Computational Humanities and Computational Social Science Workshop (Ecole nationale des chartes - PSL) for their insightful remarks. We particularly thank Sylvaine Guiot and Jean-Philippe Cointet for their guidance. This work received support from the CultureLab: Computational Science of Culture Grand Research Programme of Université PSL, funded under the Investissements d'Avenir programme launched by the French Government and implemented by ANR, under reference ANR-10-IDEX-0001-02 PSL. Errors remain our own

8 Bibliographical References

References

Issa Annamoradnejad and Gohar Zoghi. 2024. Colbert: Using BERT sentence embedding in parallel neural networks for computational humor. *Expert Systems with Applications*, 249:123685.

Taylor Arnold and Lauren Tilton. 2019. Distant viewing: analyzing large visual corpora. *Digital Scholarship in the Humanities*, 34(Supplement_1):i3–i16.

Valentin Barriere, Nahuel Gomez, Leo Hemamou, Sofia Callejas, and Brian Ravenet. 2025. Standup4ai: A new multilingual dataset for humor detection in stand-up comedy videos. *arXiv preprint arXiv:2505.18903*.

Florian Cafiero and Marie Puren. 2025. A riddle in a haystack: Llm detection of intricate wordplays in colette and willy's novels for authorship attribution. In *Digital Humanities 2025*, Lisbon, Portugal.

Alina Glushkova, Dimitrios Makrygiannis, and Sotiris Manitsaris. 2023. Interactive sensorimotor guidance for learning motor skills of a glass blower. In Unknown, editor, *Culture and Computing*, volume 13933 of *Lecture Notes in Computer Science*, pages 29–43. Springer.

Yuan Gong, Sameer Khurana, Leonid Karlinsky, and James Glass. 2023. Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers. In *Proceedings of INTERSPEECH 2023*, pages 2798–2802, Dublin, Ireland. ISCA.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020. Semeval-2020 task 7: Assessing humor in edited news headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 746–758, Barcelona (online). International Committee for Computational Linguistics.

Leonardo Impett and Franco Moretti. 2017. Totentanz: Operationalizing aby warburg's pathosformeln. *New Left Review*, (107):68–97.

Antonios Kalloniatis and Panagiotis Adamidis. 2025. Computational humor recognition: a systematic literature review. *Artificial Intelligence Review*, 58:43. Article 43.

Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. 2022. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2637–2646.

J. A. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. Semeval 2021 task 7: Hahackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119, Online. Association for Computational Linguistics.

Piotr Mirowski, Kory Mathewson, and Boyd Branch. 2025. The theater stage as laboratory: Review of real-time comedy LLM systems for live performance. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 88–95, Online. Association for Computational Linguistics.

Piotr W. Mirowski, Juliette Love, Kory W. Mathewson, and Shakir Mohamed. 2024. A robot walks into a bar: Can language models serve as creativity support tools for comedy? an evaluation of LLMs' humour alignment with comedians. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, pages 1622–1636, Rio de Janeiro, Brazil. Association for Computing Machinery. See also arXiv:2405.20956 for open-access preprint.

Vanessa C Pope, Rebecca Stewart, and Elaine Chew. 2026. Timing structures in live comedy: A matched-sequence approach to mapping performance dynamics. *PNAS Nexus*, 5(1):pgaf394.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. Semeval-2017 task 6: #hashtagwars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 49–57, Vancouver, Canada. Association for Computational Linguistics.

Robert R. Provine. 1992. Contagious laughter: Laughter is a sufficient stimulus for laughs and smiles. *Bulletin of the Psychonomic Society*, 30(1):1–4.

Michael J. Rau, Peter Broadwell, Simon Wiles, and Vijoy Abraham. 2023. Ai-assisted performance analysis: Deep learning for live and archival theater. In *Digital Humanities 2023: Book of Abstracts*, Graz, Austria. Centre for Information Modelling — Austrian Centre for Digital Humanities, University of Graz. ADHO Digital Humanities Conference (DH2023), 10–14 July 2023.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3621–3625, Hong Kong, China. Association for Computational Linguistics.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376, Lisbon, Portugal. Association for Computational Linguistics.