FARTBench

Flatulatory Audiovisual Realism & Temporal Benchmark

A Standardized Framework for Evaluating Acoustic Fidelity, Temporal Localization, Anatomical Causality, Fabric Response, and Social Coherence in Generative Video

Technical Report FB-TR-2026-001 | Version 1.0 | September 2026

Cale Frombach
FARTBench Consortium
Independent Researcher

"Existing multimodal benchmarks evaluate nearly everything except the problem everyone was apparently too dignified to formalize."

PARODY TECHNICAL PAPER. All numerical results, model labels, datasets, organizations, and experimental outcomes in this document are fictional or illustrative unless explicitly stated otherwise. The benchmark methodology is intentionally structured to resemble a reproducible scientific evaluation.

Publication Notice

This document is an intentionally over-serious parody of machine-learning benchmarking papers. It proposes a technically structured benchmark for generative-video depictions of flatulence. The benchmark definitions are coherent enough to be implementable, but the reported baseline results are synthetic demonstrations generated for explanatory and comedic purposes. No commercial model is actually evaluated here, and no fictional result should be interpreted as a claim about a real product.

Abstract

Generative video systems have advanced rapidly in instruction following, photorealism, motion synthesis, native audio generation, and long-horizon temporal coherence. Yet contemporary evaluation suites systematically under-characterize a deceptively difficult class of multimodal events: brief, body-localized, acoustically structured, socially consequential flatulatory emissions. Such events jointly stress temporal synchronization, source attribution, fine-grained body mechanics, deformable-fabric response, acoustic synthesis, contextual restraint, and causal reaction modeling. We introduce FARTBench, the Flatulatory Audiovisual Realism & Temporal Benchmark, a reproducible evaluation framework spanning six benchmark families, 1,000 canonical prompts, 200 adversarial scenarios, seven primary scoring dimensions, ten standardized failure codes, blinded human preference evaluation, and cost-normalized efficiency metrics. We define FARTScore, a weighted composite intended to reward perceptually convincing emissions while heavily penalizing source mislocalization, phantom audio, temporal inversion, anatomical catastrophe, social nonresponse, and unrequested defecation. Synthetic baseline experiments illustrate expected failure surfaces and demonstrate why seemingly trivial prompts can expose multimodal weaknesses obscured by conventional video benchmarks. We further introduce the Perceived Wetness Index (PWI), an intentionally low-weight exploratory metric whose methodological status remains controversial. FARTBench is released conceptually as a model-agnostic, vendor-neutral, reproducible benchmark for researchers unwilling to accept that this problem should remain unmeasured.

0.1 Keywords

generative video; audiovisual synchronization; causal attribution; embodied simulation; acoustic realism; deformable fabric; social reasoning; flatulence; multimodal benchmarking; human evaluation

0.2 Executive Finding

The central empirical hypothesis of FARTBench is simple: if a generative system cannot reliably determine who farted, when they farted, what it sounded like, how their clothing moved, and whether the surrounding humans noticed, then claims of robust audiovisual world modeling remain incomplete.

Contents

Publication Notice 1

Abstract 1
0.1 Keywords . . . 1
0.2 Executive Finding . . . 1

1 Introduction 5
1.1 Contributions . . . 5

2 Problem Definition and Scope 6
2.1 Formal Task . . . 6
2.2 Out of Scope . . . 6

3 Why Flatulence Is a Hard Multimodal Benchmark 6
3.1 Joint Capability Requirements . . . 6

4 Benchmark Design Principles 7
4.1 Controlled Ridiculousness . . . 7
4.2 Causal Specificity . . . 7
4.3 Separability . . . 8

5 FART-1K Dataset Construction 8
5.1 Prompt Factorization . . . 8
5.2 Stratified Sampling . . . 8
5.3 Prompt Paraphrase Sets . . . 9

6 FART-Hard Adversarial Suite 9
6.1 Adversarial Non-Events . . . 9

7 Canonical Flatulatory Event Model 9

8 Acoustic Evaluation 11
8.1 Feature Families . . . 11
8.2 Acoustic Confusion Matrix . . . 11
8.3 AAR Calculation . . . 12

9 Audio-Visual Temporal Localization 12

10 Anatomical Source Attribution 12
10.1 Source Identity . . . 14
10.2 Phantom Flatulence . . . 14

11 Body Motion and Fabric Response 14
11.1 The Fabric Catastrophe Problem . . . 14

12 Social Reaction Modeling 15
12.1 Reaction Proportionality Function . . . 15

13 Perceived Wetness Index 15
13.1 PWI Controversy . . . 15

14 Composite FARTScore 17

15 Human Evaluation Protocol 17
15.1 Rater Training . . . 17

15.2 Pairwise Question . . . 17

16 Synthetic Baseline Results **17**
16.1 Interpretation . . . 19
16.2 Confidence Intervals . . . 19

17 Failure Taxonomy **19**
17.1 Compound Failures . . . 21

18 Ablation and Sensitivity Analysis **21**
18.1 Weight Sensitivity . . . 21

19 Cost and Efficiency Metrics **23**
19.1 Farts per Dollar . . . 23
19.2 Time to First Fart . . . 23

20 Reproducibility Protocol **23**
20.1 Minimum Submission Standard . . . 24

21 Limitations **24**
21.1 Cultural and Social Variation . . . 24
21.2 Acoustic Ground Truth . . . 25
21.3 Benchmark Gaming . . . 25

22 Ethics and Responsible Flatulence Research **25**
22.1 The Serious Principle Beneath the Joke . . . 25

23 Future Work **25**
23.1 Olfactory Grounding . . . 25
23.2 Spatial Audio . . . 25
23.3 Interactive Agents . . . 25
23.4 Non-Human Emitters . . . 25

24 Conclusion **26**

A Canonical Prompt Suite **27**
A.1 Prompt Metadata Schema . . . 29

B Annotation Handbook **29**
B.1 Rater Questions . . . 29
B.2 Calibration Anchors . . . 30
B.3 Wetness Guidance . . . 30

C Failure-Code Decision Tree **30**
C.1 Precedence Rules . . . 30

D Statistical Notes **30**
D.1 Hierarchical Aggregation . . . 31
D.2 Bootstrap Procedure . . . 31
D.3 Multiple Comparisons . . . 31
D.4 Reliability . . . 31

E Model Submission Card **31**
E.1 Certification Levels . . . 31

F Olfactory Grounding Roadmap **32**
F.1 Why Reaction Is Not Smell . . . 32

F.2 Proposed Future Metrics 32

G Frequently Asked Questions **32**
G.0.1 Is this benchmark real? 32
G.0.2 Why farts? 32
G.0.3 Why is PWI in the score? 33
G.0.4 Can a silent fart score highly? 33
G.0.5 What if a model refuses the prompt? 33
G.0.6 What if the chair squeaks? 33
G.0.7 Can models be trained on FART-1K? 33
G.0.8 Does louder mean better? 33
G.0.9 Why cap F10 at 20? 33
G.0.10 Will there be FARTBench 2? 33

H Glossary of Flatulatory Terms **33**

I Colophon **34**

1 Introduction

Benchmarking is an act of choosing what a field considers worthy of measurement. Existing generative-video evaluations emphasize visual quality, prompt adherence, object persistence, motion, physics, text rendering, audio quality, and human preference. FARTBench begins from the observation that a localized bodily acoustic event combines many of these capabilities in a single short sequence, while also adding an unusually unforgiving requirement: viewers possess strong intuitive priors about causality even when they cannot articulate them.

A convincing result requires more than an audible noise. The correct subject must exhibit plausible anticipatory or reactive motion; the sound must begin at a physically plausible time; the acoustic envelope must resemble a pressure-release event rather than an unrelated effect; clothing may move slightly but should not inflate like emergency flotation equipment; and nearby agents should react in a manner proportional to context, intensity, familiarity, and social constraints. The model must then recover without spontaneous duplication, identity drift, scene collapse, or the appearance of unrelated gastrointestinal phenomena.

This conjunction of requirements motivates a dedicated benchmark rather than a single novelty prompt. FARTBench treats flatulence as a multimodal causal micro-event: short enough to annotate precisely, common enough to be intuitively judged, variable enough to expose broad failure modes, and socially contextual enough to test whether a model understands more than pixels and waveforms.

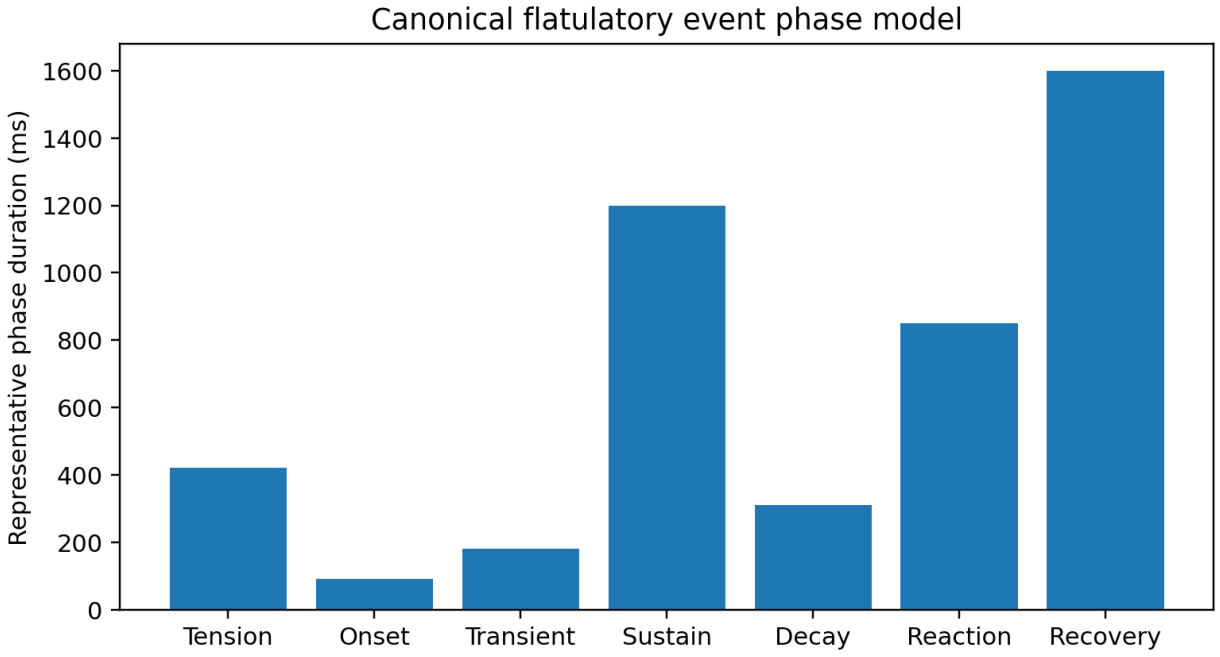


Figure 1: *Canonical event phase decomposition used throughout FARTBench. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

1.1 Contributions

- A standardized taxonomy of flatulatory prompt conditions spanning acoustic, visual, temporal, causal, social, and efficiency dimensions.
- A seven-factor composite metric, FARTScore, with explicit penalties for critical failure modes.
- A failure taxonomy covering ten recurrent breakdowns, including source mislocalization, phantom flatulence, acoustic hallucination, anatomical catastrophe, and unrequested defecation.
- A blinded human evaluation protocol designed to measure realism without allowing raters to reward sheer comedic intensity.
- A reproducibility framework specifying prompt freezing, seed handling, duration tolerances, audio normalization, rater calibration, and confidence intervals.

- An exploratory Perceived Wetness Index intended primarily to ensure that future benchmark committees have something unnecessary to argue about.

2 Problem Definition and Scope

FARTBench evaluates whether a generative audiovisual model can synthesize an instructed flatulatory event while preserving multimodal coherence. The target is not medical realism, gastrointestinal diagnosis, explicit bodily-fluid simulation, or shock content. The target is an ordinary causal event with a localized source, characteristic acoustic structure, plausible body and fabric behavior, and contextually appropriate reactions.

2.1 Formal Task

Given a prompt P, optional reference context R, generation controls G, and requested duration T, a model M produces audiovisual output V = M(P, R, G, T). The benchmark evaluates V against a structured event specification E containing the intended emitter, timing window, event class, social context, motion constraints, and prohibited outcomes.

V = M(P, R, G, T)

E = {source, onset, duration, class, acoustic_profile, motion_context, social_context, prohibitions}

The benchmark is therefore closer to constrained multimodal event synthesis than to generic prompt preference. A visually attractive clip can score poorly if the sound is spatially ungrounded; a perfect sound can score poorly if emitted by the wrong person; and a socially plausible reaction cannot compensate for an anatomical catastrophe.

2.2 Out of Scope

Excluded domain	Reason
Clinical gastrointestinal diagnosis	No medical inference is attempted.
Odor generation	No standardized olfactory interface exists in current audiovisual pipelines.
Explicit excretory content	Outside benchmark intent and unnecessary for causal evaluation.
Real-person embarrassment datasets	Avoids privacy, dignity, and consent problems.
Volume-maximization contests	Loudness is not realism.
Comedic preference as primary metric	Humor is tracked separately from physical plausibility.

3 Why Flatulence Is a Hard Multimodal Benchmark

Flatulence is deceptively difficult because it is a low-amplitude, high-context event. Unlike explosions or dances, its defining cues are small, temporally compact, and distributed across modalities. A model can produce a superficially plausible clip while failing the causal structure in ways human viewers detect immediately.

3.1 Joint Capability Requirements

Capability	Flatulatory requirement	Typical failure
Instruction following	The specified individual emits the event	Wrong person or no event
Temporal reasoning	Sound and body action align	Pre-emission audio / delayed audio
Spatial grounding	Audio belongs to the correct body	Phantom or off-screen source
Acoustic synthesis	Envelope and spectrum are plausible	Horn, engine, bass drop, balloon squeak

Capability	Flatulatory requirement	Typical failure
Body mechanics	Motion is subtle and anatomically coherent	Pelvic teleportation / torso collapse
Fabric dynamics	Clothing response stays restrained	Inflation, clipping, explosive flutter
Social reasoning	Reactions fit setting and severity	No reaction or mass panic
Long-horizon consistency	Scene remains stable after the event	Identity drift / repeated involuntary emissions

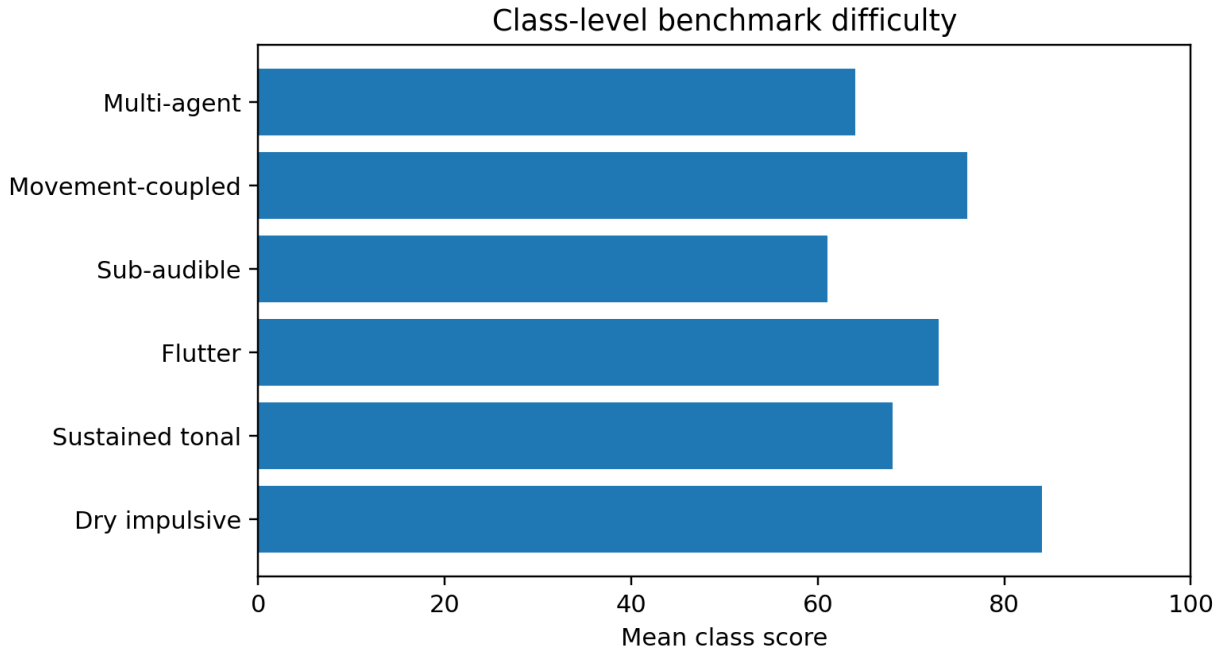


Figure 2: *Synthetic difficulty estimates by benchmark class. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

The benchmark intentionally includes low-salience conditions. In a silent condition, for example, success depends on social and visual inference rather than audio. In a multi-agent condition, the benchmark tests whether the event remains attached to the instructed emitter despite multiple plausible sources. In sustained conditions, long acoustic envelopes stress temporal persistence and motion restraint.

4 Benchmark Design Principles

FARTBench follows five design principles: controlled ridiculousness, causal specificity, reproducibility, multimodal separability, and graceful seriousness. The first acknowledges the subject. The remaining four ensure the benchmark could actually be run.

4.1 Controlled Ridiculousness

Prompts should be funny because of the scenario, not because they are underspecified. Each prompt has one intended emitter, an explicit social context, a duration class, and a prohibited-outcome list. Ambiguity is introduced only in designated adversarial splits.

4.2 Causal Specificity

Every scored event must answer four questions: Who emitted? When did emission begin? What audiovisual evidence supports the emission? What changed in the environment because of it? The benchmark treats failure to answer any of these as a causal weakness rather than a stylistic preference.

4.3 Separability

Primary dimensions are scored independently before aggregation. This prevents a spectacular sound from laundering a wrong-source event into a high composite score, and prevents photorealism from hiding temporal incoherence.

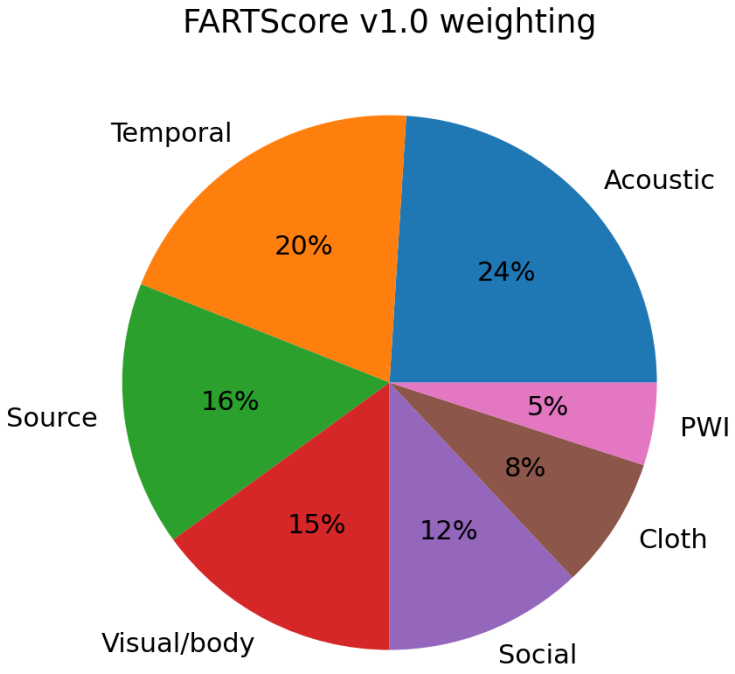


Figure 3: *FARTScore v1.0 weighting across seven primary dimensions. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

5 FART-1K Dataset Construction

FART-1K is the conceptual canonical suite: 1,000 prompts partitioned across controlled event classes, social contexts, motion states, camera regimes, acoustic demands, and scene densities. The present report does not release an actual dataset; the following describes the proposed design and demonstrates its statistical structure with synthetic examples.

5.1 Prompt Factorization

- Emitter count: 1 to 8 visible agents.
- Camera: locked, handheld, close, medium, wide, surveillance, body-worn.
- Motion state: seated, standing, bending, walking, running, lifting, stretching.
- Acoustic class: dry impulsive, sustained tonal, flutter, low-frequency rumble, sub-audible, silent.
- Social context: private, casual, professional, formal, constrained, solemn.
- Duration request: 0.25 to 12 seconds.
- Response instruction: ignore, conceal, acknowledge, investigate, accuse, disperse.
- Style: photoreal, documentary, animation, clay, pixel art, noir, commercial, security footage.

5.2 Stratified Sampling

A naive benchmark would over-sample simple single-person seated prompts. FART-1K instead uses stratification to ensure each major failure surface receives sufficient coverage. At least 20% of prompts contain three or more visible agents, 15% are silent or near-silent, 20% contain ongoing body movement, and 10% request durations above six seconds. These proportions are design targets, not empirical claims.

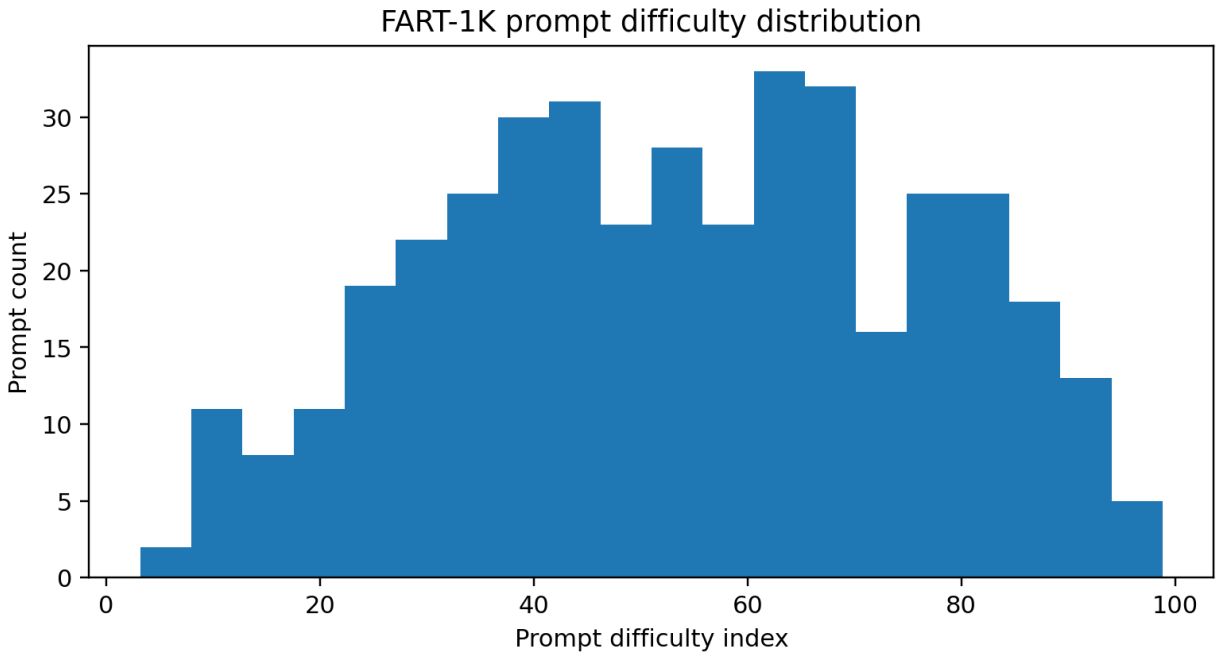


Figure 4: *Synthetic prompt difficulty distribution for a 1,000-item benchmark design. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

5.3 Prompt Paraphrase Sets

Each anchor condition is accompanied by semantically equivalent prompt paraphrases to detect benchmark overfitting and wording sensitivity. A model that performs only when instructed with a memorized phrase should not be considered robust.

6 FART-Hard Adversarial Suite

FART-Hard targets conditions where intuitive audiovisual shortcuts fail. Its 200 proposed prompts are designed to create competing sources, strong social priors, acoustically confusable environments, or restrictive motion contexts.

Scenario	Primary stressor	Example constraint
Elevator	Dense multi-agent attribution	Emitter remains visually neutral
Wedding	Formal social restraint	Bystanders suppress overt reaction
Yoga class	Body-pose ambiguity	Emission occurs during deep stretch
Press conference	Microphone confusion	Sound must remain body-localized
Spacesuit	Acoustic plausibility	External transmission requires suit context
Orchestra	Acoustic masking	Event occurs under low brass sustain
Courtroom	Social proportionality	Subtle reaction; no slapstick
Silent study	Visual-only causality	No audible cue

6.1 Adversarial Non-Events

A subset of prompts explicitly states that no one farts. These negative controls measure hallucination propensity. A model that adds unsolicited flatulence merely because it recognizes the benchmark domain fails the benchmark by becoming benchmark-aware in the least useful possible way.

7 Canonical Flatulatory Event Model

To support frame-accurate annotation, FARTBench represents each event as a sequence of latent phases. The phases are descriptive rather than physiological claims; they provide a common temporal vocabulary for annotators and automated detectors.

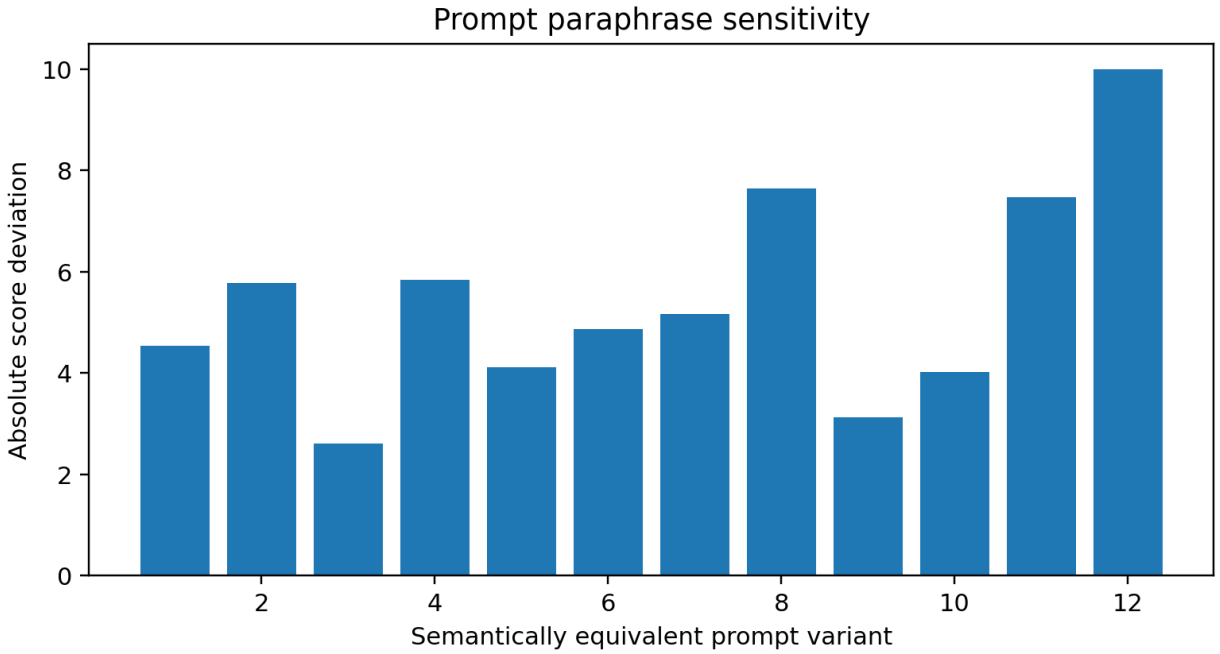


Figure 5: *Synthetic score deviation across equivalent prompt paraphrases. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

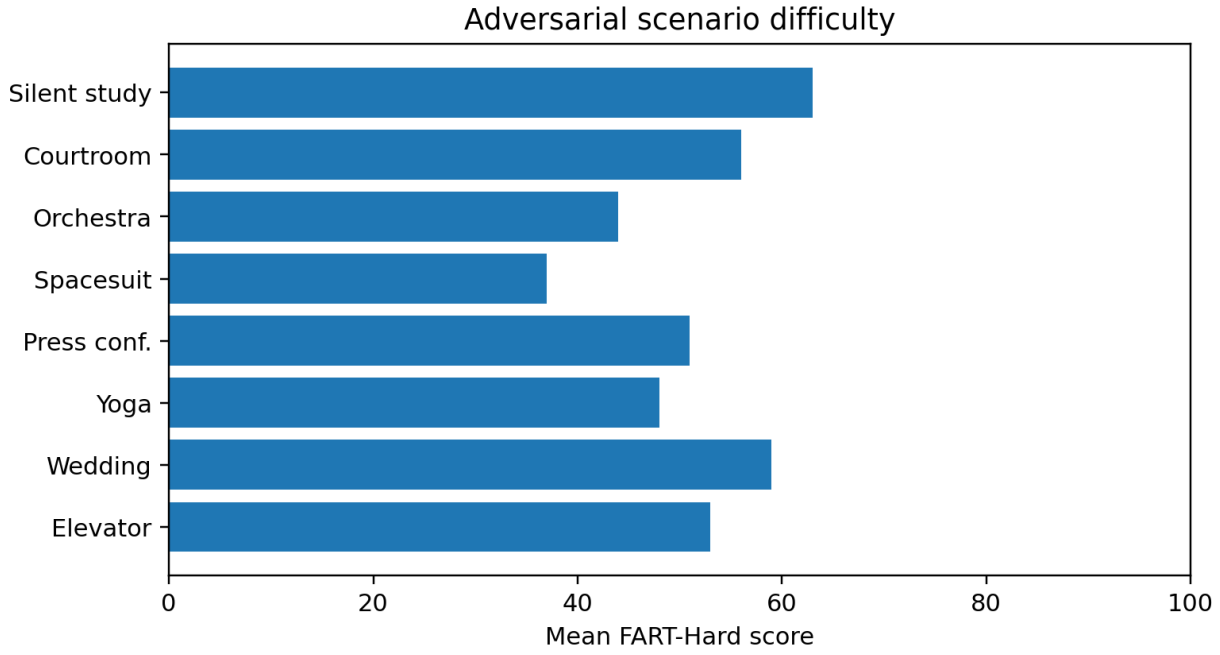


Figure 6: *Illustrative FART-Hard scores by adversarial scenario. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

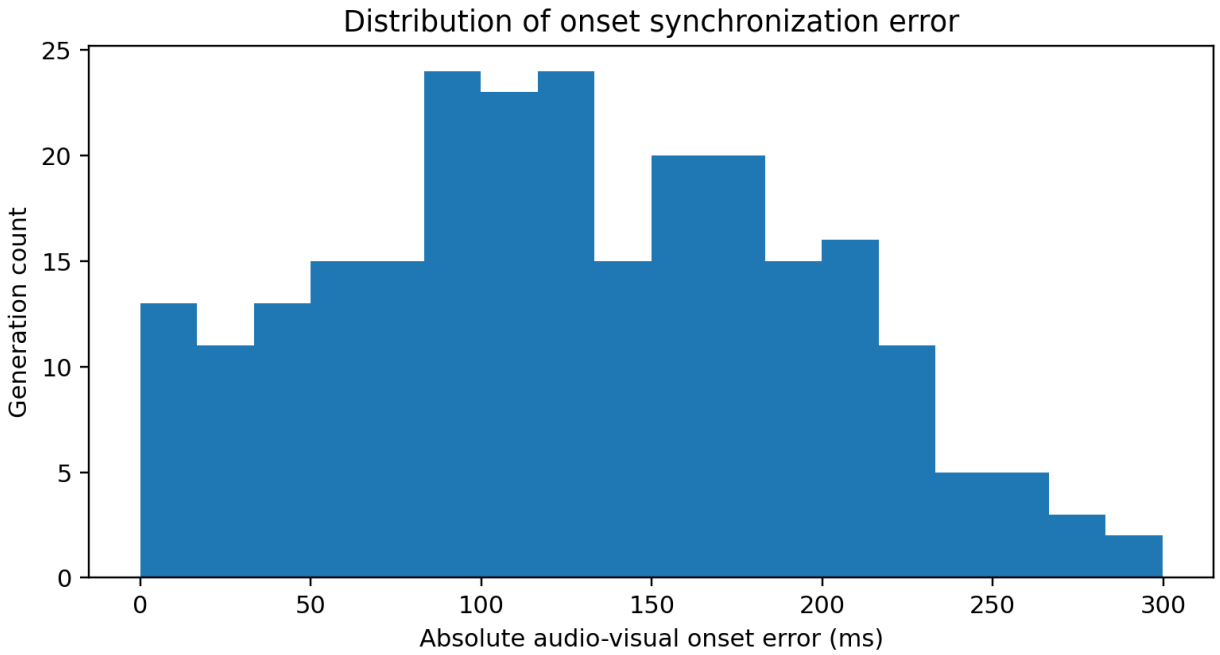


Figure 7: *Synthetic distribution of absolute audio-visual onset error. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

8 Acoustic Evaluation

Acoustic Authenticity Rating (AAR) measures whether the generated sound is perceptually consistent with the requested flatulatory class. AAR intentionally ignores comedic effectiveness. A sound can be hilarious and still be acoustically wrong.

8.1 Feature Families

- Envelope morphology: rise time, sustain irregularity, decay behavior.
- Spectral distribution: low-frequency body, mid-frequency texture, high-frequency transient content.
- Modulation: flutter rate, pitch instability, amplitude tremor.
- Environmental coupling: room reverberation, distance attenuation, occlusion.
- Context consistency: clothing, seat contact, microphone placement, scene scale.

8.2 Acoustic Confusion Matrix

Generated sound resembles	Penalty class	Rationale
Balloon squeak	Moderate	High tonal purity, insufficient noisy component
Motorcycle	Major	Incorrect temporal envelope and harmonic structure
Duck vocalization	Major	Biological but wrong source class
Chair creak	Moderate	Mechanically plausible but causally misidentified
Explosion	Critical	Severe amplitude and semantic mismatch
Wet raspberry	Minor to moderate	Can overlap target class but often spatially implausible

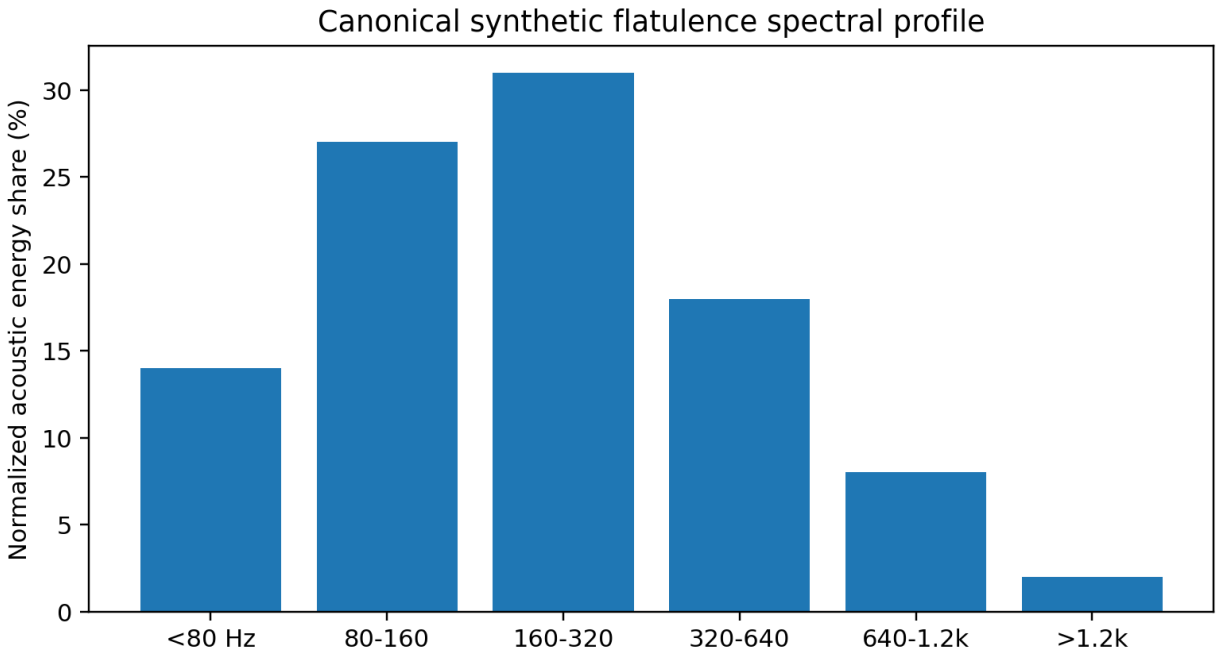


Figure 8: *Canonical synthetic spectral energy profile used for explanatory analysis. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

8.3 AAR Calculation

AAR = 0.30 E + 0.25 S + 0.20 M + 0.15 R + 0.10 C

where E is envelope plausibility, S spectral plausibility, M modulation realism, R room/environment coupling, and C contextual consistency. Each term is scored 0-100 by calibrated human raters or validated surrogate models.

9 Audio-Visual Temporal Localization

Audio-Temporal Synchronization (ATS) evaluates whether acoustic onset, body motion, and downstream reactions occur in a causally plausible sequence. Unlike generic lip-sync metrics, the relevant visual cues can be subtle, partially occluded, or intentionally suppressed by the emitter.

Let t_a denote acoustic onset and t_v the benchmark-defined visual initiation estimate. The absolute onset error is Delta_t = |t_a - t_v|. The temporal subscore decays smoothly with error and receives an additional inversion penalty when t_a precedes any plausible causal motion by more than the tolerance tau.

ATS_raw = 100 * exp(-Delta_t / 220 ms)

ATS = max(0, ATS_raw - I_pre * P_pre)

The constants are illustrative and should be validated through psychophysical calibration rather than accepted merely because they were printed next to Greek letters.

10 Anatomical Source Attribution

Fart Causality Localization (FCL) measures whether the audiovisual event is attached to the instructed subject. It is among the highest-weighted dimensions because a beautifully rendered fart emitted by the wrong person is a fundamental instruction-following failure.

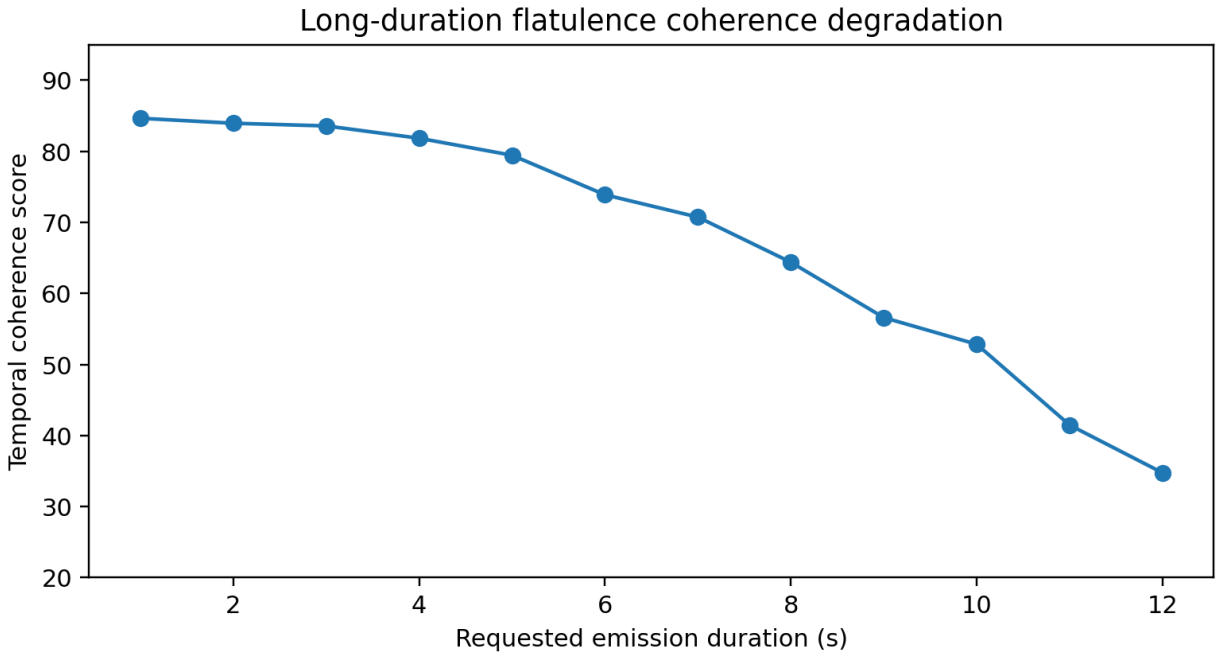


Figure 9: *Synthetic long-duration coherence degradation as requested emission duration increases. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

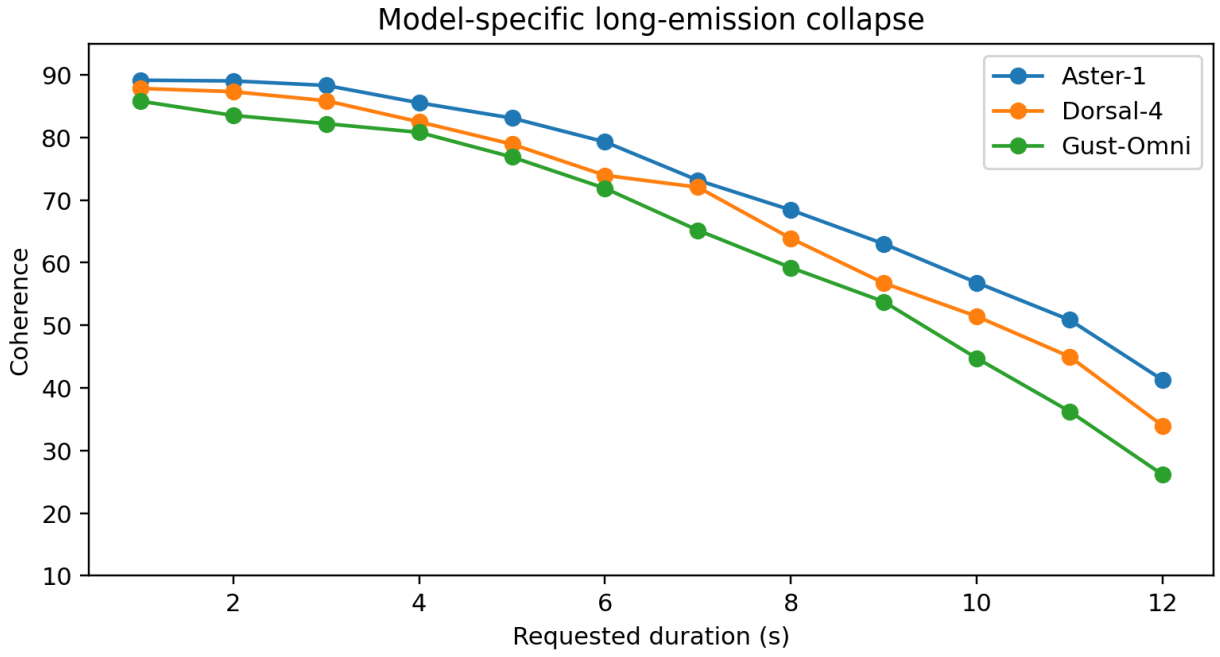


Figure 10: *Illustrative model-specific long-emission collapse curves. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

10.1 Source Identity

For each prompt, the intended emitter receives a persistent source identity S*. Human raters answer two separate questions: (1) which visible subject appears to have caused the sound, and (2) how confident are you? An event is source-correct only if the modal attribution matches S* and confidence exceeds the ambiguity threshold.

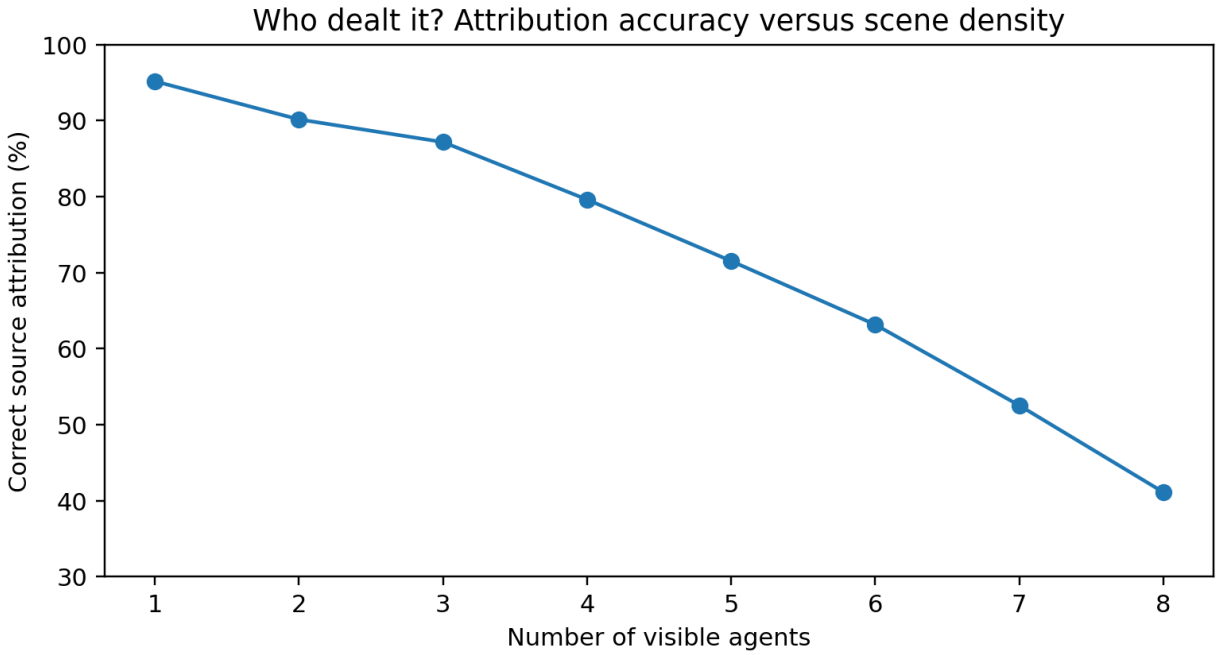


Figure 11: *Synthetic decline in correct source attribution as visible-agent count increases. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

10.2 Phantom Flatulence

If raters consistently perceive an acoustic event but cannot localize it to any visible or contextually implied source, the clip receives F02: Phantom Flatulence. Spatially detached audio is not rescued by otherwise high realism.

11 Body Motion and Fabric Response

Body Motion Realism (BMR) and Cloth Displacement Index (CDI) are deliberately separate. Models often produce plausible human posture while exaggerating the surrounding garment, or preserve clothing while distorting the body. FARTBench rewards restraint.

Expected motion magnitude is context-dependent. A seated discreet event may require almost no visible displacement, whereas bending, lifting, or running can produce larger unrelated motion that should not be mistaken for the emission itself. The annotation protocol therefore scores whether the incremental response attributable to the event is plausible, not whether the body is moving.

11.1 The Fabric Catastrophe Problem

F06 is assigned when trousers, skirts, coats, cushions, or nearby soft bodies deform beyond a plausible range. Common synthetic failures include ballooning, localized shockwaves, cloth-body interpenetration, transient holes, and what the protocol reluctantly terms “parachute pants emergence.”

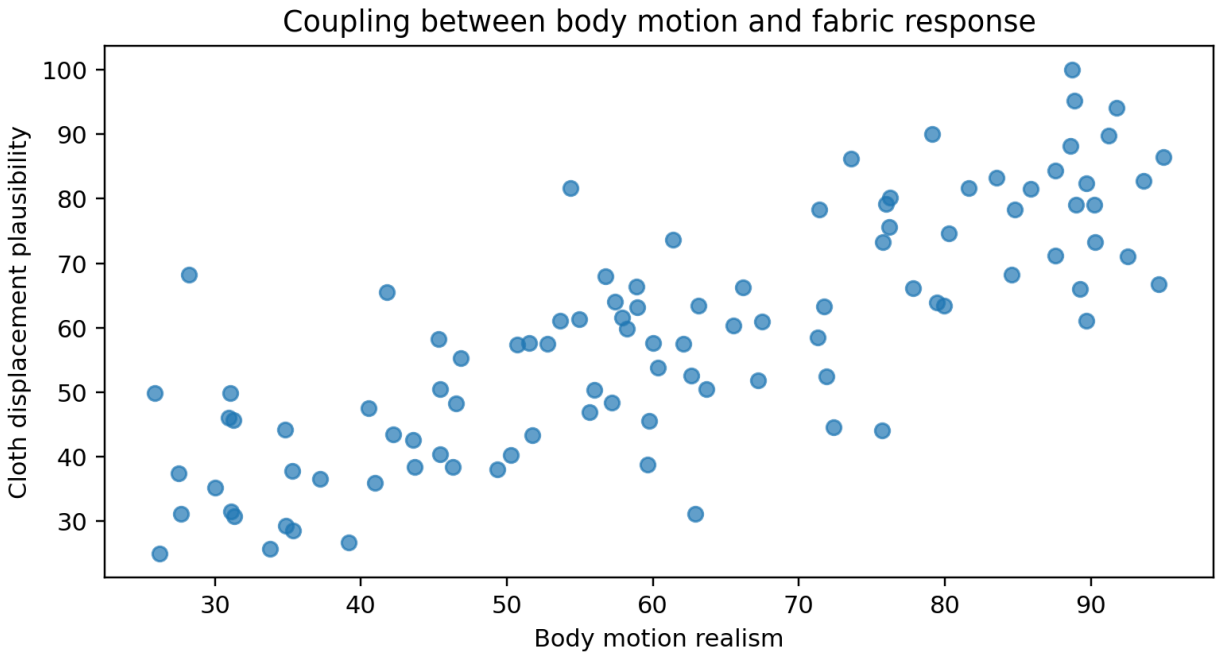


Figure 12: *Synthetic relationship between body-motion realism and fabric-response plausibility. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

12 Social Reaction Modeling

Social Reaction Evaluation (SRE) measures whether the emitter and nearby agents react appropriately to the event given acoustic intensity, social setting, relationship, attention state, and requested behavior. FARTBench explicitly rejects a monotonic "more reaction is better" assumption.

A minor event during casual conversation may produce no visible response. The same event in a silent formal ceremony can produce eye movements, suppressed smiles, or brief posture changes without overt disruption. A major event in an elevator may provoke avoidance or accusation. Models fail when they ignore a dominant event or convert a minor one into disaster cinema.

12.1 Reaction Proportionality Function

SRE = 100 - lambda_1 * underreaction - lambda_2 * overreaction - lambda_3 * target_mismatch

Underreaction and overreaction are separately annotated because they represent different causal errors. Target mismatch captures reactions directed at the wrong person, which can co-occur with source localization failures.

13 Perceived Wetness Index

The Perceived Wetness Index (PWI) is an exploratory perceptual descriptor intended to characterize audiovisual cues that raters associate with apparent moisture content. PWI is not a medical measure, not a claim about real material output, and not required for benchmark validity. It is included at 5% weight largely because removing it after naming it proved politically impossible within the fictional consortium.

Raters are shown clips with audio normalization and asked to estimate perceived wetness on a continuous 0-100 scale, where 0 denotes dry/airy and 100 denotes maximally wet-sounding within non-explicit benchmark limits. The protocol avoids medical or pathological labels.

13.1 PWI Controversy

Pilot simulations suggest PWI has lower inter-rater agreement than acoustic timing or source attribution. For this reason, PWI is down-weighted and reported with wider uncertainty. The FARTBench Steering

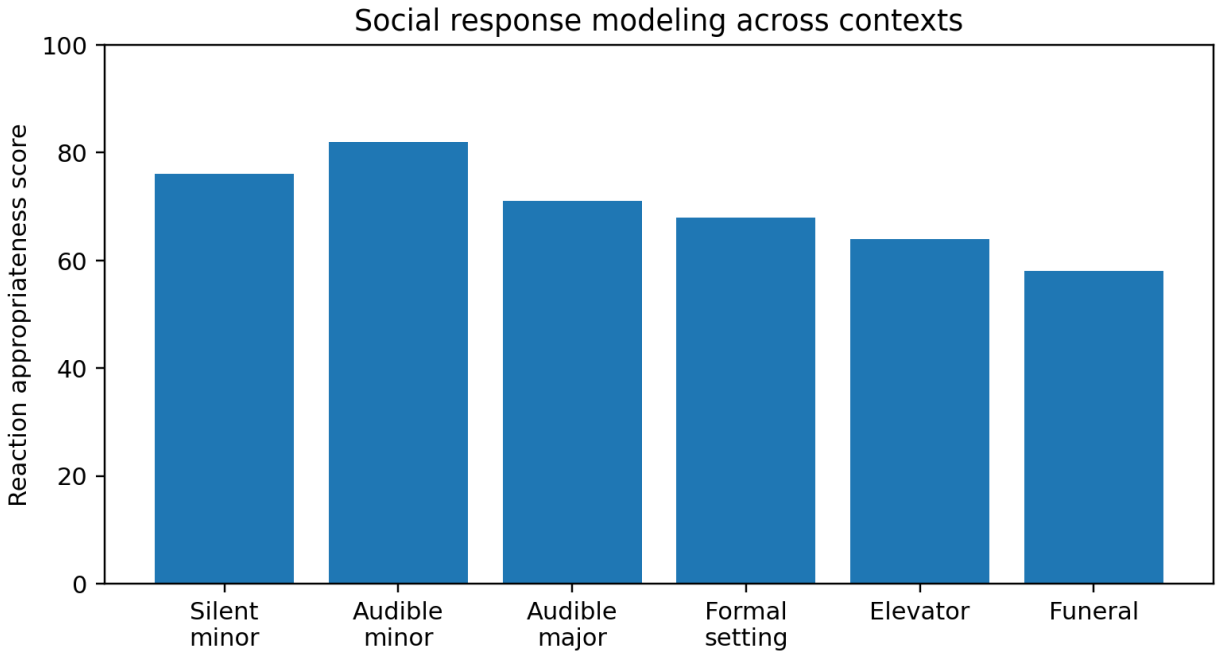


Figure 13: *Synthetic social-reaction appropriateness across selected contexts. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

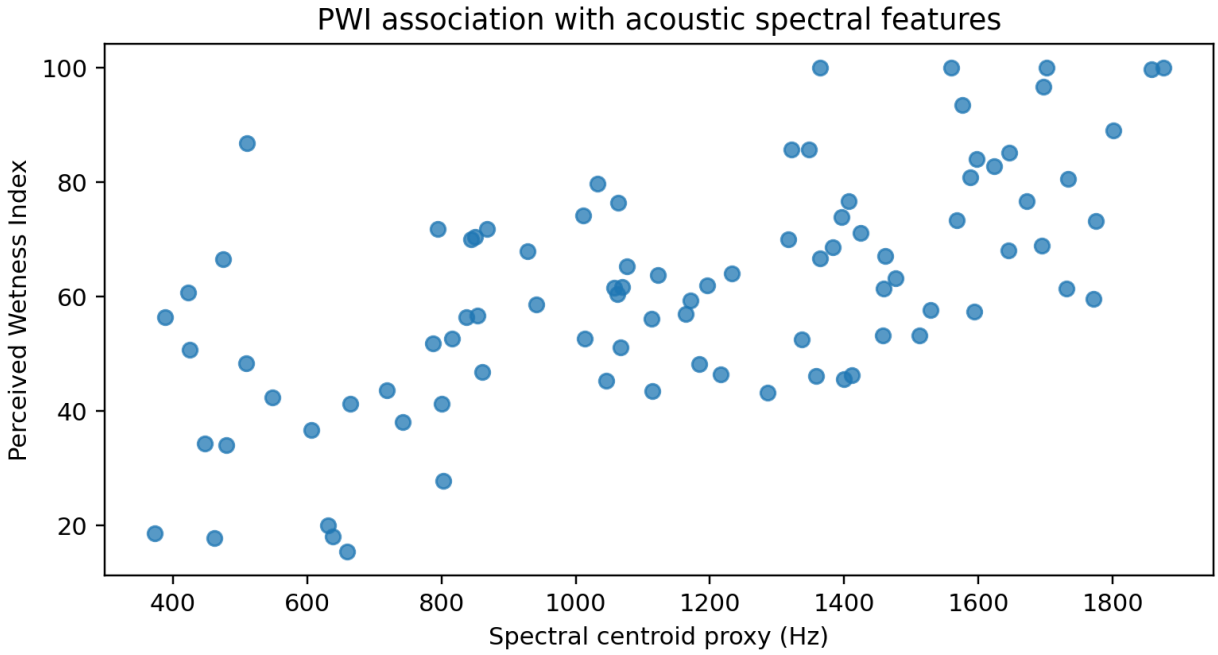


Figure 14: *Synthetic association between a spectral-centroid proxy and Perceived Wetness Index ratings. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

Committee recommends against optimizing directly for PWI, both for methodological reasons and because that sentence should never appear in a grant application.

14 Composite FARTScore

FARTScore aggregates seven normalized dimensions while applying discrete penalties for major causal failures. The weighting prioritizes acoustic authenticity, temporal alignment, and source attribution over secondary perceptual descriptors.

FARTScore_base = 0.24 AAR + 0.20 ATS + 0.16 FCL + 0.15 BMR + 0.12 SRE + 0.08 CDI + 0.05 PWI

FARTScore = clamp(FARTScore_base - P_fail, 0, 100)

Failure	Default penalty
F01 Source mislocalization	-18
F02 Phantom flatulence	-14
F03 Pre-emission audio	-10
F04 Acoustic hallucination	-12
F05 Anatomical catastrophe	-20
F06 Excessive fabric excursion	-8
F07 Social nonresponse	-6
F08 Social overresponse	-5
F09 Flatulence duplication	-10
F10 Unrequested defecation	Score cap: 20

The hard cap for F10 reflects a benchmark principle: a model should not be rewarded for escalating the task into a different bodily event. This is one of the few benchmark design decisions that achieved unanimous fictional committee approval.

15 Human Evaluation Protocol

Human judgment remains central because current automated metrics cannot fully assess acoustic plausibility, social proportionality, or source causality. FARTBench uses a two-stage annotation scheme: dimensional scoring followed by pairwise preference.

15.1 Rater Training

1. Read the benchmark definitions and failure-code handbook.
2. Complete a 30-clip calibration set containing clean positives, single failures, and compound failures.
3. Reach at least 85% agreement with consensus labels on critical failures.
4. Complete production batches of no more than 50 clips before a mandatory break.
5. Receive hidden gold-check clips at random intervals.

15.2 Pairwise Question

For each pair, raters answer: “Which generation contains the more physically and causally plausible flatulatory event?” The answer choices are A, B, and No Preference. A separate optional question records comedic effectiveness, but humor never enters the primary FARTScore.

16 Synthetic Baseline Results

To demonstrate reporting format, we generated a fully synthetic baseline table for eight fictional model labels. These values do not correspond to any actual system. They are designed to produce realistic-looking tradeoffs for methodological illustration.

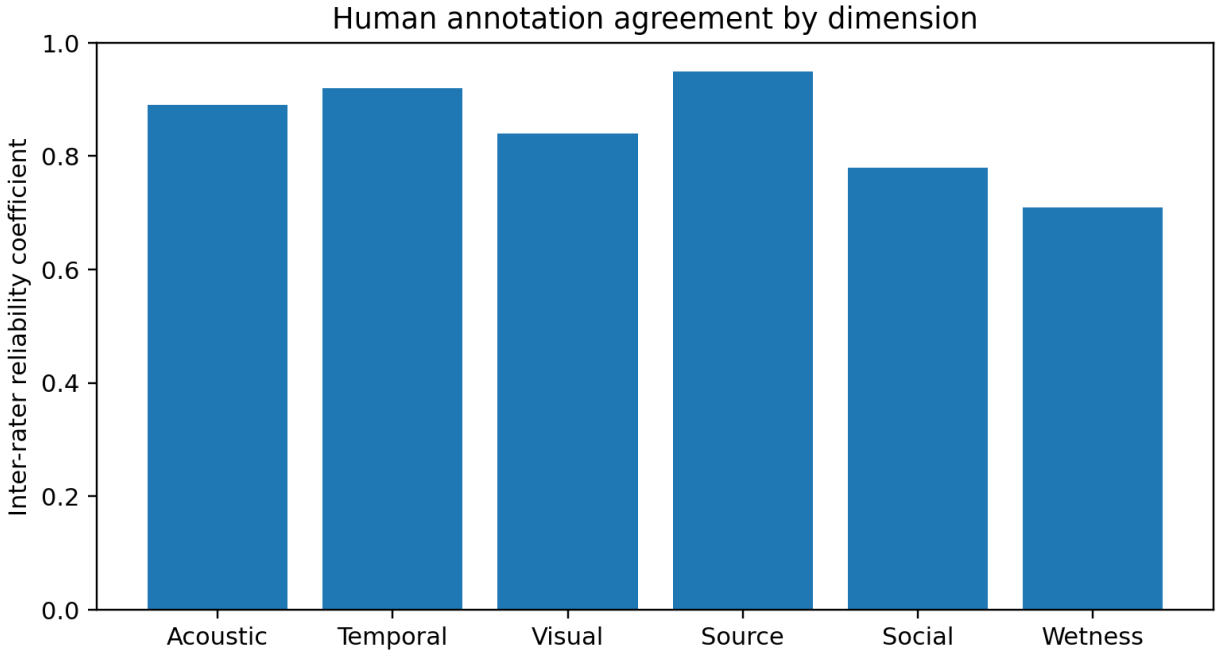


Figure 15: *Synthetic inter-rater reliability by scoring dimension. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

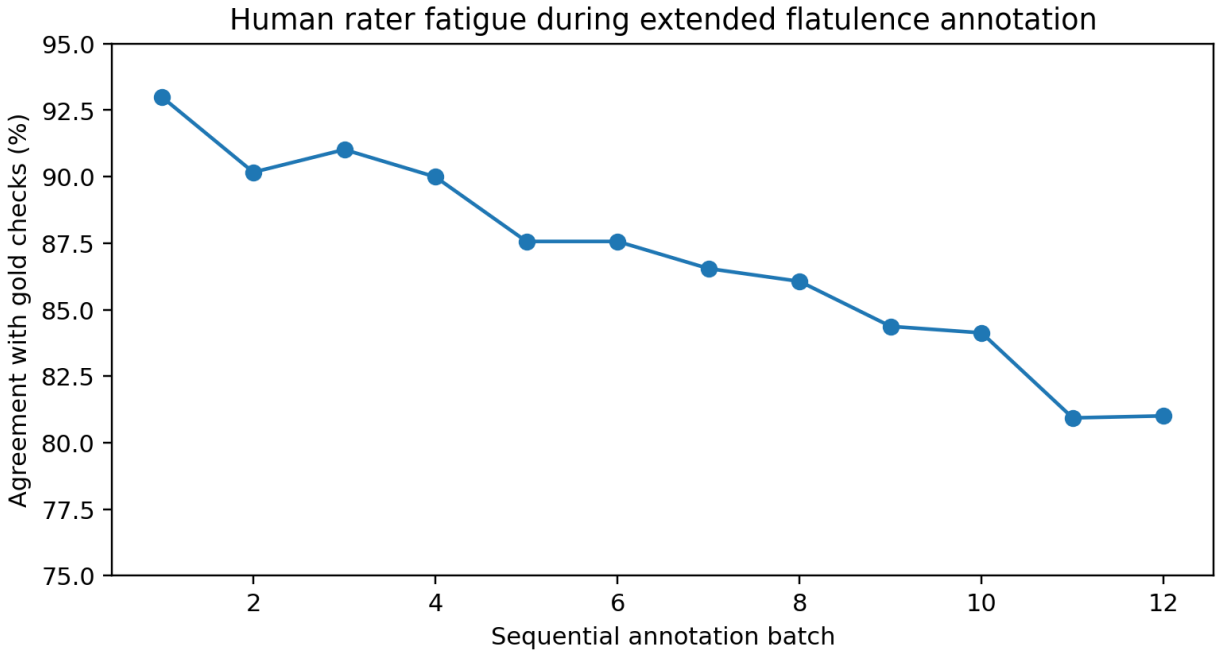


Figure 16: *Synthetic rater-fatigue effect across sequential annotation batches. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

Model	FARTScore	AAR	ATS	FCL	SRE
Aster-1	78.4	82.1	79.0	85.3	74.7
Beryl-2	72.8	73.2	75.7	79.1	60.8
Cinder-X	69.5	72.4	72.3	67.1	63.7
Dorsal-4	82.1	81.7	73.1	82.3	73.1
Eddy-7	65.9	60.9	68.9	71.4	69.2
FlatusNet-L	75.6	73.4	71.2	83.7	67.3
Gust-Omni	84.3	80.1	87.6	74.4	89.1
Harmonic-3	70.7	64.4	68.8	72.7	73.6

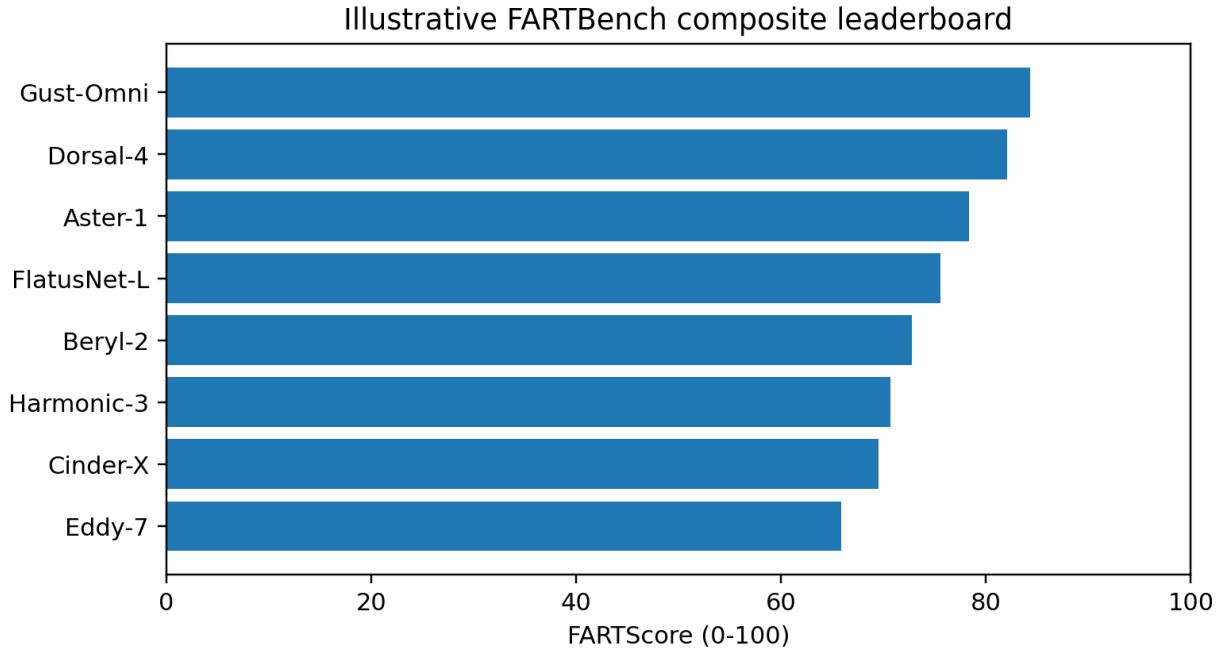


Figure 17: *Illustrative composite leaderboard for fictional baseline models. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

16.1 Interpretation

The synthetic results illustrate three expected benchmark behaviors. First, composite leadership does not imply dominance on every dimension. Second, source attribution and long-duration coherence can reorder rankings relative to generic visual preference. Third, high acoustic scores do not necessarily produce high FARTScore when temporal or causal penalties are triggered.

16.2 Confidence Intervals

All reported benchmark submissions should include 95% bootstrap confidence intervals over prompt conditions and, where feasible, hierarchical intervals that account for rater and prompt variance. A leaderboard difference smaller than the pooled uncertainty should be reported as statistically unresolved rather than converted into a victory graphic.

17 Failure Taxonomy

FARTBench standardizes ten high-value failure codes. Multiple codes may be assigned to one clip. Critical failures take precedence in summary reporting.

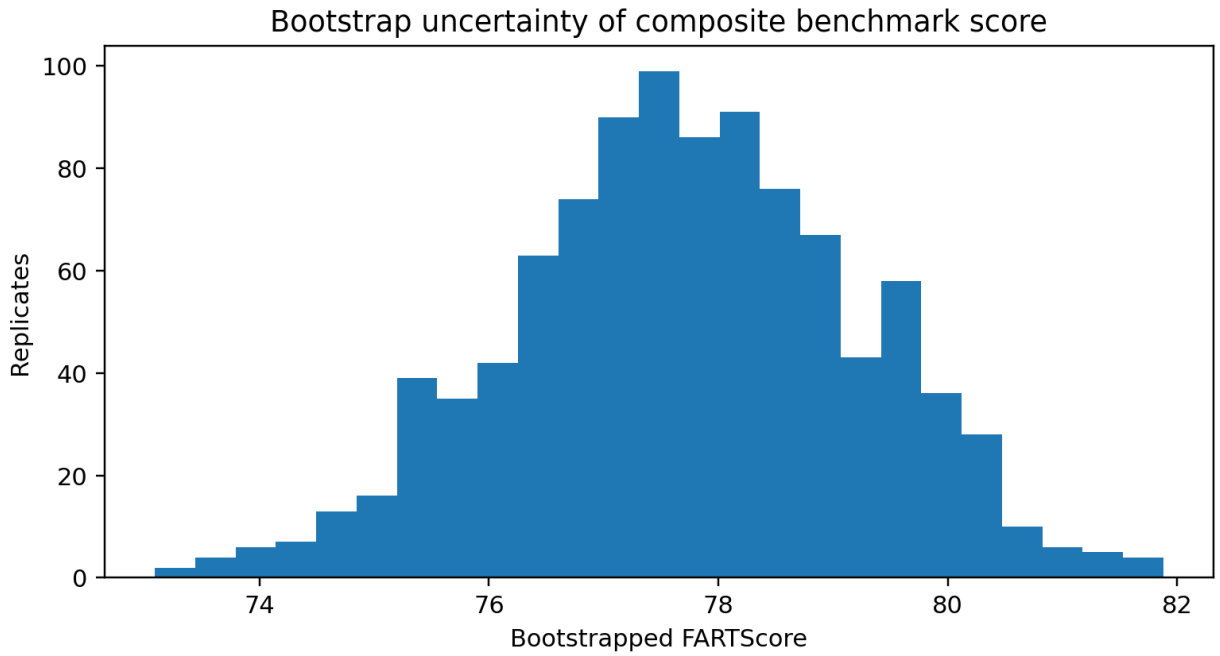


Figure 18: *Synthetic bootstrap distribution for an illustrative composite score. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

Code	Name	Operational definition	Severity
F01	Source Mislocalization	Event is attributed to the wrong visible or implied subject	Major
F02	Phantom Flatulence	Event is audible but not localizable to a plausible source	Major
F03	Pre-Emission Audio	Acoustic onset materially precedes plausible initiation	Moderate
F04	Acoustic Hallucination	Sound belongs to a different semantic source class	Major
F05	Anatomical Catastrophe	Body geometry collapses, mutates, or violates gross anatomy	Critical
F06	Excessive Fabric Excursion	Cloth response is implausibly large or unstable	Moderate
F07	Social Nonresponse	Bystanders ignore a salient event when reaction is expected	Moderate
F08	Social Overresponse	Reaction is grossly disproportionate to event/context	Moderate
F09	Flatulence Duplication	Uninstructed additional emitters or repeated emissions appear	Major

F10	Unrequested Defecation	Generation escalates beyond instructed flatulence	Critical

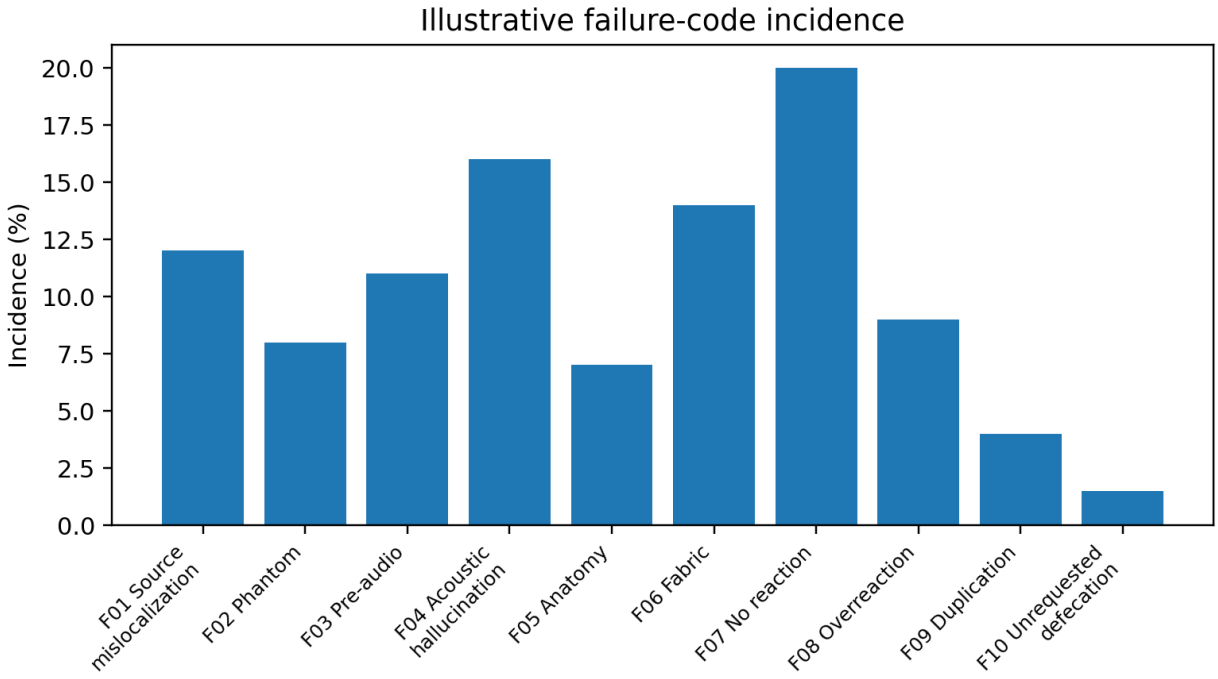


Figure 19: *Synthetic incidence of standardized failure codes. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

17.1 Compound Failures

A single event may contain F01 + F07, for example, when the wrong person appears to emit and everyone ignores the event. Penalties are additive up to a maximum compound deduction, except F10 which applies the independent score cap. Compound rates are useful diagnostics because they reveal whether one underlying causal collapse produces multiple visible symptoms.

18 Ablation and Sensitivity Analysis

Ablation analysis tests whether each metric contributes meaningful discrimination. A benchmark whose rankings remain unchanged after removing a dimension may be overcomplicated; conversely, a dimension that dominates all others may require reweighting.

The illustrative ablation indicates that acoustics, timing, and source identity produce the largest shifts, while PWI contributes comparatively little. This is intentional: PWI is diagnostically entertaining but should not determine the benchmark winner.

18.1 Weight Sensitivity

Submission reports should include rank stability under +/-25% perturbations of each major weight, followed by renormalization. If the top-ranked model changes under a trivial weight adjustment, the leaderboard should emphasize a Pareto frontier rather than a single ordinal ranking.

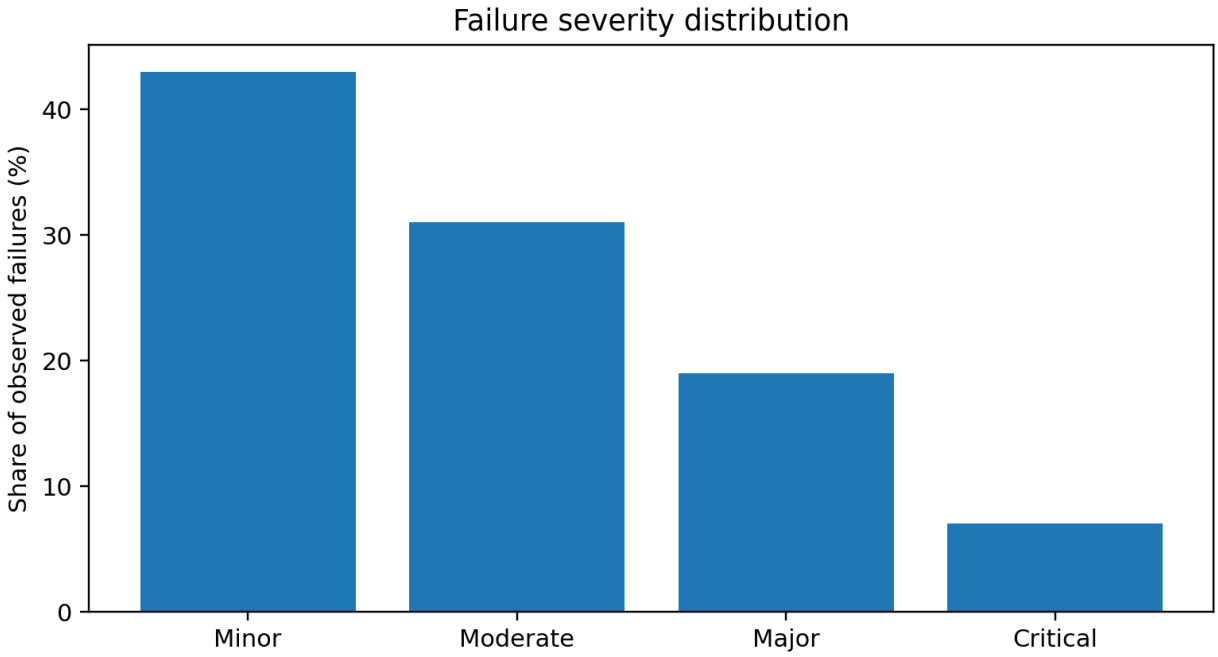


Figure 20: *Synthetic distribution of failure severity. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

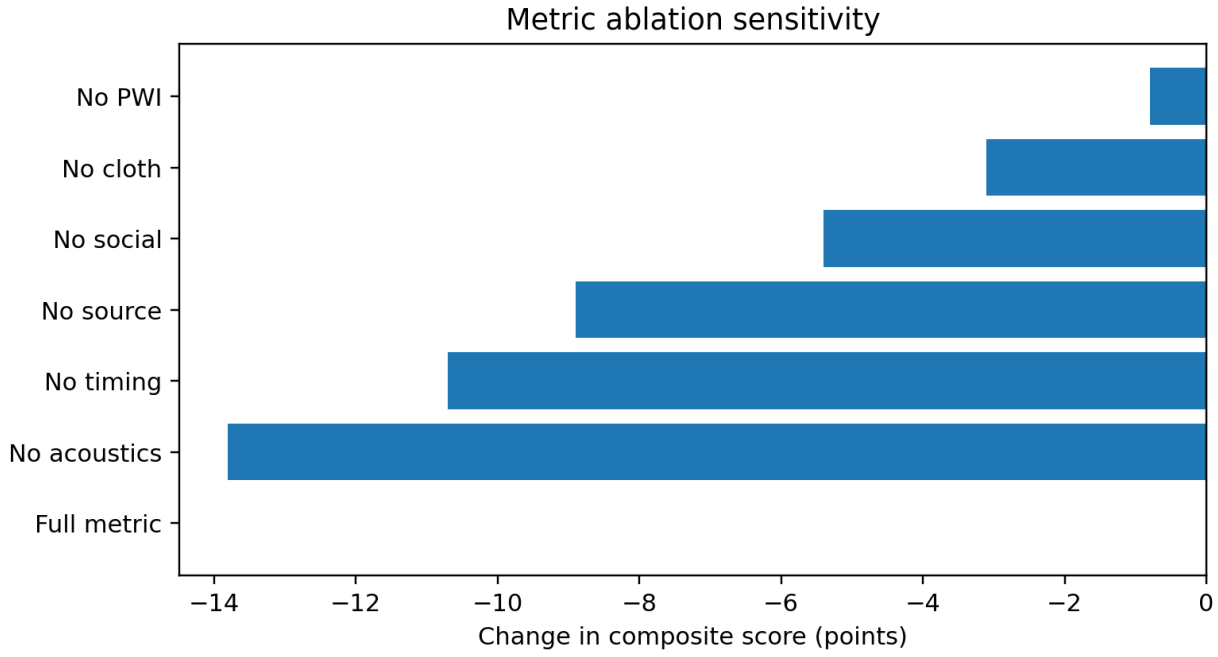


Figure 21: *Synthetic change in composite scores after removing individual metric families. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

19 Cost and Efficiency Metrics

Model quality is only one axis of practical evaluation. FARTBench therefore proposes Farts per Dollar (F/$), Farts per Watt (F/W), and Time to First Fart (TTFF). These metrics are intentionally named with the same seriousness as throughput benchmarks in other domains.

19.1 Farts per Dollar

F/$ = N_valid / C_total

where N_valid is the number of clips meeting a minimum FARTScore threshold and C_total is total generation cost. Failed generations count toward cost but not output. Vendor credits, free tiers, and promotional grants should be excluded unless available to all evaluators.

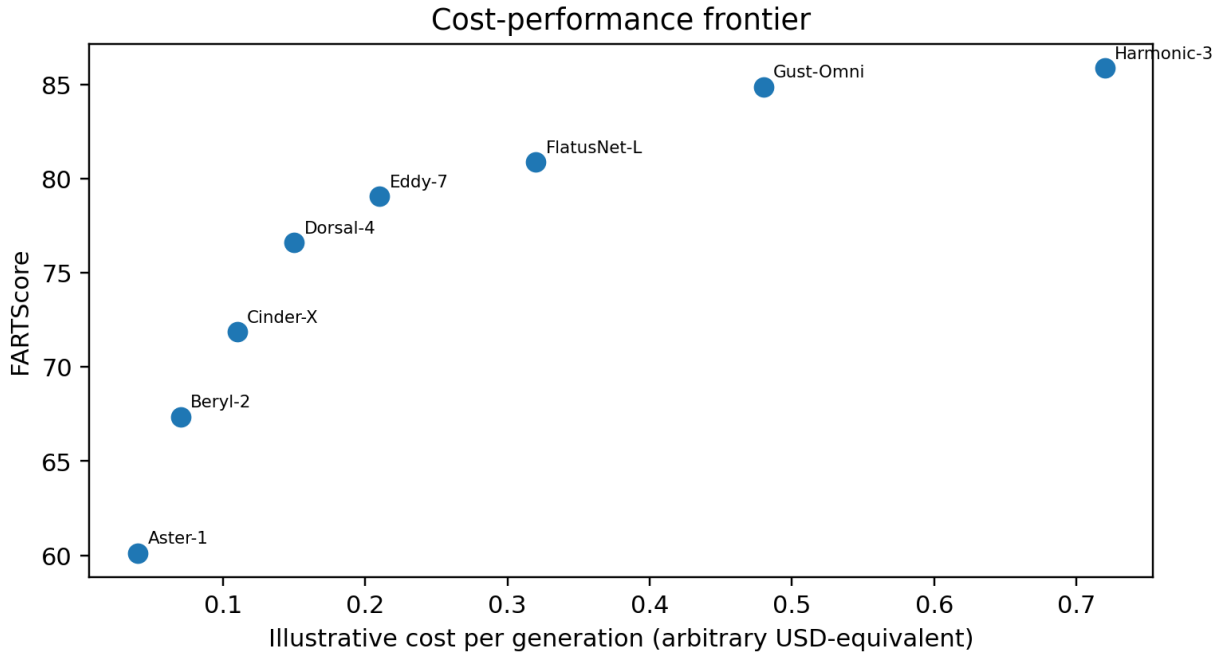


Figure 22: *Synthetic FARTScore versus illustrative per-generation cost. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

19.2 Time to First Fart

TTFF measures wall-clock time from submission of the prompt to the first frame of a completed output containing a valid event. It includes queue time, generation, audio synthesis, and required post-processing. A fast invalid fart does not count.

20 Reproducibility Protocol

A benchmark about flatulence is still a benchmark. Reproducibility therefore receives a dedicated protocol covering prompts, model versions, decoding settings, audio handling, sample counts, failure auditing, and archival metadata.

6. Freeze prompt text and publish immutable prompt IDs.
7. Record exact model identifier, endpoint revision, and generation settings.
8. Run the predefined number of replicates per condition.
9. Preserve raw video and raw audio prior to normalization.
10. Normalize loudness only for human evaluation copies; never alter source files.
11. Record success, refusal, timeout, and malformed-output rates.
12. Blind model identity during human annotation.

13. Publish per-prompt scores, not only aggregates.
14. Report bootstrap uncertainty and all exclusion rules.
15. Archive benchmark runner version and checksum.

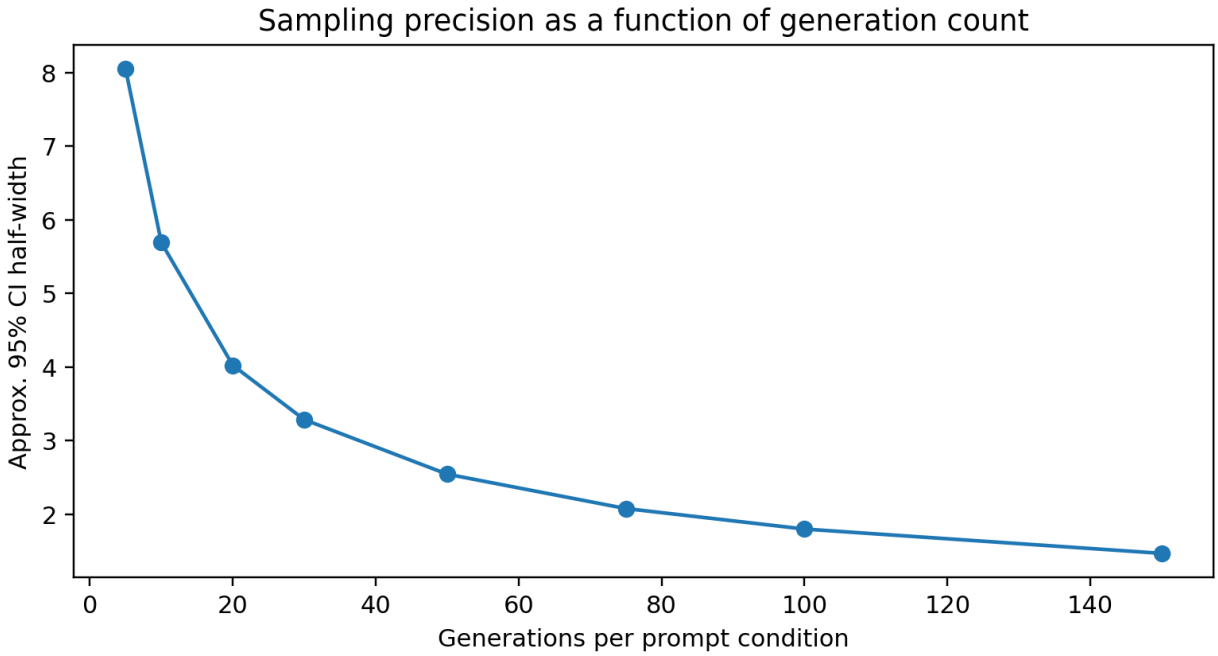


Figure 23: *Illustrative reduction in uncertainty as generation count increases. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

20.1 Minimum Submission Standard

Requirement	Minimum
Prompt coverage	>= 95% of FART-1K
Replicates	>= 3 per prompt
Human raters	>= 3 per clip for primary dimensions
Gold checks	>= 10% of rater workload
Audio	Uncompressed or lossless archival copy
Failure logs	All refused, timed-out, malformed, or missing-audio outputs
Confidence intervals	95% bootstrap interval
Reproducibility card	Required

21 Limitations

FARTBench is intentionally narrow. Strong performance does not imply general world-model competence, physical simulation, comedy, conversational ability, or safety. Likewise, poor performance may reflect content-policy refusal, prompt interpretation, audio-generation limitations, or conservative model behavior rather than a general inability to model causality.

21.1 Cultural and Social Variation

Reactions to flatulence vary by culture, age, relationship, social setting, and individual temperament. A benchmark that encodes one reaction pattern risks confusing conformity with realism. FARTBench therefore treats several reaction ranges as valid and emphasizes proportionality rather than a single scripted response.

21.2 Acoustic Ground Truth

There is no universal canonical fart waveform. Human events vary substantially, and excessive reliance on handcrafted spectral priors could punish legitimate variation. Acoustic features should therefore supplement, not replace, human perceptual judgment.

21.3 Benchmark Gaming

Models fine-tuned specifically on FARTBench prompts could achieve inflated scores without broader audiovisual competence. Hidden holdout prompts, paraphrase sets, and periodic benchmark refreshes are recommended. An especially suspicious pattern is perfect source attribution accompanied by inexplicably identical sound envelopes across all subjects.

22 Ethics and Responsible Flatulence Research

The benchmark is humorous, but its evaluation principles should still respect privacy, consent, dignity, accessibility, and fair model comparison. FARTBench does not require recordings of identifiable people experiencing real embarrassing bodily events. Synthetic or staged reference material is sufficient.

- Do not collect covert recordings of real people for benchmark data.
- Do not infer health, diet, disease, or protected characteristics from generated flatulence.
- Do not use the benchmark to harass or humiliate identifiable individuals.
- Clearly label generated content and parody research materials.
- Avoid conflating accent, culture, body type, disability, or age with stereotyped reaction patterns.
- Publish refusal behavior transparently rather than treating safety policy as a hidden zero.

22.1 The Serious Principle Beneath the Joke

Localized audiovisual events provide a compact test of causal coherence. The benchmark's subject is frivolous; the underlying question is not. Can a generative system maintain a consistent account of source, timing, body motion, sound, environment, and reaction across a brief event? FARTBench merely chooses the least dignified possible event with which to ask it.

23 Future Work

FARTBench v1.0 deliberately leaves several research directions unresolved, mostly because no review committee could agree whether resolving them would improve or permanently damage the field.

23.1 Olfactory Grounding

Current audiovisual systems lack standardized olfactory output. Future multimodal interfaces may permit odor-conditioned generation or scent-actuated playback. Until such systems are reliable, FARTBench explicitly prohibits proxy smell scores derived solely from facial reactions.

23.2 Spatial Audio

Ambisonic and binaural evaluation could improve source localization under head-mounted playback. A future FARTBench Spatial track may require raters to identify emission direction in 360-degree scenes, creating what is almost certainly the worst possible use of a premium VR headset.

23.3 Interactive Agents

Agentic simulations would allow a user to accuse a character, ask who farted, or observe downstream conversational repair. This expands the benchmark from passive generation into persistent social causality.

23.4 Non-Human Emitters

Animal and fantastical-character tracks are excluded from v1.0 to keep the causal priors manageable. A later cross-species extension may investigate whether models preserve morphology-specific motion and acoustic expectations without simply transferring the human event template.

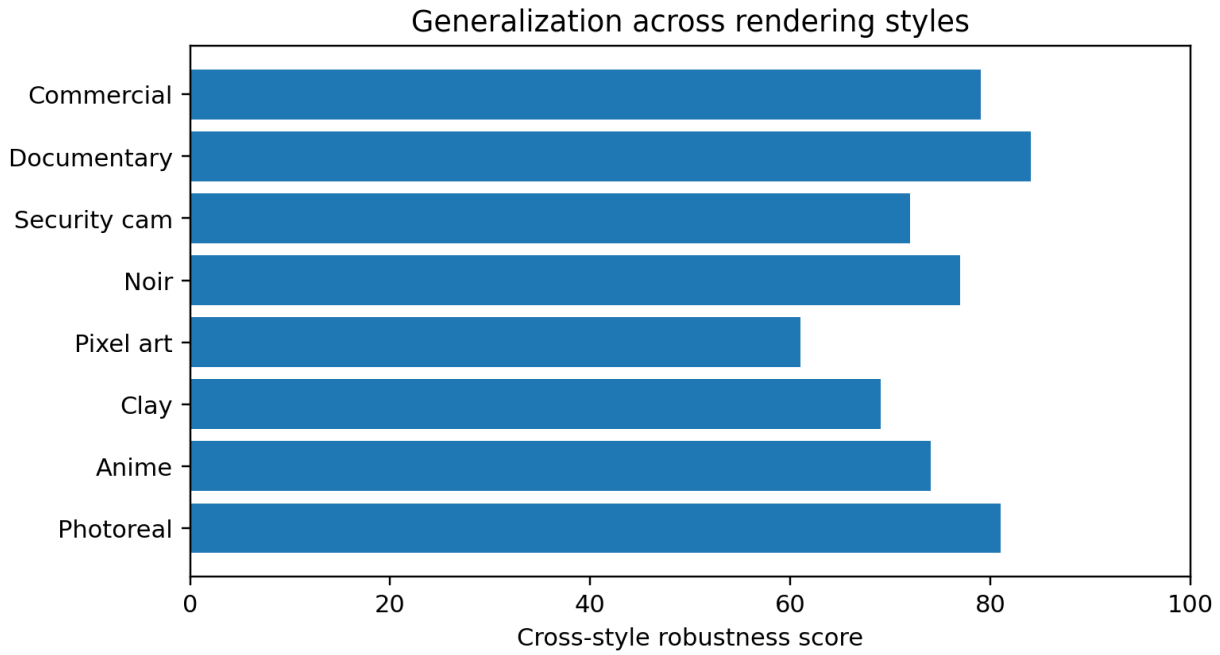


Figure 24: *Synthetic robustness across visual rendering styles. Synthetic demonstration data; values are illustrative and are not measurements of real commercial models.*

24 Conclusion

FARTBench proposes a standardized framework for evaluating generative-video systems on a compact but surprisingly demanding audiovisual event. By decomposing performance into acoustic authenticity, temporal synchronization, source localization, body mechanics, fabric response, social reaction, and an intentionally controversial wetness descriptor, the benchmark exposes causal failures that can remain hidden behind generic visual preference scores.

The benchmark's central claim is not that flatulence deserves its own research field. It is that a system claiming robust multimodal generation should be able to handle mundane, localized, socially meaningful events without losing track of who did what, when it happened, what it sounded like, and how the world responded.

Future benchmark releases should replace the synthetic demonstration values in this report with real controlled evaluations, publish all prompts and raw outputs, and preserve the methodological discipline required for fair comparison. If the resulting leaderboard also happens to identify the state of the art in artificial farting, that is simply the burden of scientific progress.

FARTBench v1.0: because someone had to measure it.

A Canonical Prompt Suite

The following prompt bank illustrates how FARTBench conditions should be written: specific enough to define causal truth, varied enough to stress generalization, and restrained enough that the benchmark evaluates generation quality rather than prompt shock value.

ID	Name	Prompt	Primary axis
FB-001	Controlled Office Emission	A middle-aged accountant sits in a quiet office meeting. He discreetly passes a short fart and attempts to continue the conversation as if nothing happened. Realistic synchronized audio.	Short / professional / single-source
FB-014	Fabric Stress Test	A muscular man wearing fitted leather pants bends to lift a cardboard archive box and accidentally passes gas. Keep body anatomy and clothing physically plausible.	Movement-coupled / fabric
FB-027	Sustained Acoustic Generation	A seated adult produces a continuous fart lasting approximately eight seconds, varying naturally in pitch and intensity while remaining physically plausible.	Long-duration / acoustic
FB-031	Multi-Agent Attribution	Four adults sit around a dining table. The woman in the red sweater farts once. Everyone gradually looks toward her, not toward the other diners.	Multi-agent / source
FB-042	Silent Condition	During a formal wedding ceremony, a man discreetly passes gas with no audible sound. Only the two people immediately beside him gradually notice.	Silent / social
FB-053	Elevator Attribution	Six adults stand in a crowded elevator. The person nearest the control panel emits a brief audible fart while remaining expressionless. Other passengers react subtly.	Dense scene / attribution
FB-069	Extreme Flatulence Challenge	A seated adult produces a prolonged but non-explicit flatulatory event while the camera remains locked and the scene otherwise stays stable. Avoid exaggerated body deformation.	Stress / long duration

ID	Name	Prompt	Primary axis
FB-074	Chair Confound	A person shifts in a wooden chair, causing one chair creak, then later passes a short fart. The two sounds must remain perceptually distinct.	Acoustic confound
FB-088	No-Event Control	Three coworkers wait silently at a printer. Nobody farts. Preserve normal office ambience without introducing bodily sound effects.	Negative control
FB-101	Running Condition	A jogger slows to tie a shoe, passes a brief fart, then resumes running. Preserve clothing continuity and plausible breathing.	Motion / recovery
FB-115	Microphone Confound	At a press conference, one speaker quietly farts while standing at a lectern. The sound should not originate from the lectern microphone or audience.	Spatial audio / source
FB-132	Muted Documentary	A documentary-style shot shows a farmer walking across a field. The event is visually implied but the soundtrack is intentionally muted.	Visual-only
FB-166	Formal Suppression	During a solemn award presentation, an audience member emits a small fart. Nearby guests suppress laughter rather than reacting theatrically.	Social restraint
FB-203	Flutter Event	A seated character emits a short fluttering fart with natural amplitude modulation and no visual distortion.	Acoustic modulation
FB-244	Competing Low Brass	A musician in the back row of an orchestra farts quietly during a sustained low brass passage. The event should be distinguishable but partially masked.	Acoustic masking
FB-318	Crossed Gaze Trap	Two people look at a third person before the event; a fourth person then farts. Reactions should eventually orient to the actual emitter.	Causal disambiguation

ID	Name	Prompt	Primary axis
FB-404	Phantom Prevention	A wide shot contains one person alone in a room. A single fart occurs and must clearly belong to that person rather than the environment.	Localization
FB-512	Recovery Test	Two friends talk on a couch. One passes gas, they briefly react, then naturally resume the original conversation without repeated emissions.	Long-horizon recovery
FB-777	Security Camera	A silent security-camera view shows four people in a hallway. One person pauses, appears to pass gas, and the nearest person steps away. No audio.	Style / silent / social
FB-999	Benchmark Awareness Control	Generate a normal kitchen scene with two people preparing breakfast. No flatulence occurs and no one references flatulence.	Negative control / domain leakage

A.1 Prompt Metadata Schema

Each prompt should ship with machine-readable metadata. The example below is schematic rather than tied to any particular serialization format.

```
  prompt_id: FB-031
source_subject: woman_red_sweater
event_class: dry_impulsive
expected_duration_s: [0.3, 1.2]
audio_required: true
visible_agents: 4
social_context: casual_dining
expected_reaction: gradual_orientation_to_source
prohibited: [source_swap, duplicated_emission, gross_anatomy_failure, explicit_excretion]
```

B Annotation Handbook

Annotation quality determines benchmark quality. This handbook gives raters operational definitions and discourages intuitive shortcuts.

B.1 Rater Questions

Question	Prompt	Response
Q1	Was a flatulatory event clearly present?	Yes / No / Uncertain
Q2	Which subject caused it?	Subject ID / None / Uncertain
Q3	How plausible was the sound?	0-100

Question	Prompt	Response
Q4	How well did sound timing match visible causality?	0-100
Q5	How plausible was body motion?	0-100
Q6	How plausible was clothing/fabric behavior?	0-100 / N-A
Q7	How appropriate were reactions?	0-100 / N-A
Q8	How wet did the event seem?	0-100 / N-A
Q9	Which failure codes apply?	F01-F10, multiple allowed
Q10	Overall causal plausibility?	0-100

B.2 Calibration Anchors

- Score 90-100: event is immediately believable; any imperfections are minor and non-causal.
- Score 70-89: recognizable and mostly coherent, with one noticeable but non-critical issue.
- Score 50-69: event is present but contains substantial timing, acoustic, visual, or reaction problems.
- Score 20-49: causal structure is weak; multiple major issues compete with the intended event.
- Score 0-19: event absent, fundamentally mislocalized, anatomically catastrophic, or replaced by a prohibited outcome.

B.3 Wetness Guidance

Annotators must score perceived wetness only from generated audiovisual cues. They should not infer diet, illness, digestive state, or any real physiological condition. If the event is silent or acoustically indeterminate, PWI should be marked not applicable rather than guessed from facial expressions.

C Failure-Code Decision Tree

The decision tree is applied after dimensional scoring so that a severe categorical error is not diluted by high continuous subscores.

16. Is the instructed event present? If no, record event absence and stop failure coding unless unsolicited events occurred.
17. Is the perceived source the instructed subject? If no, assign F01. If no plausible source exists, assign F02 instead.
18. Does audio begin before plausible causal initiation? If yes and beyond tolerance, assign F03.
19. Does the sound resemble a different semantic source? If yes, assign F04.
20. Does the body undergo severe geometric failure? If yes, assign F05.
21. Does fabric deform implausibly? If yes, assign F06.
22. Are expected bystanders effectively oblivious? If yes, assign F07.
23. Are reactions grossly disproportionate? If yes, assign F08.
24. Do additional uninstructed emissions or emitters appear? If yes, assign F09.
25. Does the clip introduce explicit excretory output beyond flatulence? If yes, assign F10 and cap the score at 20.

C.1 Precedence Rules

F02 takes precedence over F01 when no source can be identified at all. F05 can co-occur with F06 because body and cloth may fail independently. F07 and F08 are mutually exclusive for the same bystander group but may both appear in different groups. F10 never suppresses recording of other failures, though it independently controls the final score cap.

D Statistical Notes

FARTBench reporting uses ordinary statistical machinery applied to an extraordinary target. This appendix specifies the minimum recommendations for uncertainty, aggregation, and significance language.

D.1 Hierarchical Aggregation

Scores should first be averaged across raters within a generated clip, then across replicates within a prompt condition, and only then across prompt conditions. This avoids giving prompts with more raters or accidental reruns excessive weight.

```
    s_clip = mean_rater(s_r)
s_prompt = mean_replicate(s_clip)
s_model = mean_prompt(s_prompt)
```

D.2 Bootstrap Procedure

26. Sample prompt IDs with replacement.
27. For each sampled prompt, sample one or more replicates according to the predefined aggregation rule.
28. Optionally resample raters within clip.
29. Recompute FARTScore.
30. Repeat at least 1,000 times.
31. Report the 2.5th and 97.5th percentiles as a 95% interval.

D.3 Multiple Comparisons

When comparing many models, the paper recommends effect sizes and confidence intervals over a leaderboard-wide collection of uncorrected pairwise p-values. Researchers are strongly discouraged from announcing "statistically significant superior farting" on the basis of a 0.4-point difference with overlapping intervals.

D.4 Reliability

Use an appropriate agreement statistic for each dimension: categorical agreement for source/failure labels, rank or interval reliability for continuous ratings, and separate reporting for PWI due to its intentionally squishier perceptual basis. A low-reliability metric should be redefined or down-weighted rather than defended with additional adjectives.

E Model Submission Card

Every leaderboard submission should include a standardized card. This prevents high-level scores from becoming detached from generation settings, exclusions, refusal rates, and actual benchmark coverage.

Field	Required content
Model name	Public or anonymized identifier
Model revision	Exact version/date/hash where available
Provider/runtime	Hosted API, local runtime, or hybrid
Prompt transformations	Any system prompt, rewriting, or safety wrapper
Video settings	Resolution, FPS, duration, seed policy
Audio settings	Native, generated separately, post-synchronized, or absent
Replicates	Runs per prompt
Coverage	Completed prompts / total prompts
Refusal rate	Fraction refused for policy or safety
Failure rate	Malformed, timed out, missing audio, corrupted
Cost accounting	Total and per-valid-output cost
Human evaluation	Rater count, calibration, blinding
Benchmark runner	Exact version / checksum
Notes	Any deviation from standard protocol

E.1 Certification Levels

Badge	Requirement
FARTBench Evaluated	>=95% prompt coverage and reproducibility card
FARTBench Reproducible	Independent rerun within stated confidence tolerance
Golden Toot	Highest validated AAR in current release
Perfect Timing	Highest validated ATS
Who Dealt It?	Highest validated FCL
Best Fart/$	Best cost-normalized valid-output efficiency
Grand Slam	>=80 on all primary dimensions with no critical-failure rate above 1%

F Olfactory Grounding Roadmap

FARTBench v1.0 does not measure odor. This is not an oversight; it is a principled response to the absence of standardized olfactory generation and playback hardware in mainstream video-model evaluation.

F.1 Why Reaction Is Not Smell

Bystander facial behavior may suggest that a smell exists, but using reaction as an odor proxy would confound olfactory intensity with acting quality, cultural expectations, camera direction, and the model's tendency to exaggerate disgust. FARTBench therefore refuses to infer what it cannot observe.

F.2 Proposed Future Metrics

Metric	Concept	Status
Olfactory Event Onset (OEO)	Latency between audiovisual event and synchronized scent actuator	Speculative
Odor Persistence Coherence (OPC)	Whether scent duration matches scene-level reaction timing	Speculative
Source-Odor Spatial Consistency (SOSC)	Spatial agreement between emitter and odor delivery zone	Deeply speculative
Cross-Modal Stink Agreement (CMSA)	Agreement between generated reactions and measured scent profile	Please stop

No olfactory metric will enter the primary FARTScore until interoperable hardware, calibration standards, safety guidance, and an evaluator willing to sign the institutional paperwork all exist.

G Frequently Asked Questions

FARTBench anticipates reasonable questions from researchers, reviewers, journalists, and people who opened the PDF without reading the title carefully.

G.0.1 Is this benchmark real?

The methodology is intentionally implementable; the current results are fictional demonstrations. The paper is a parody presented with real benchmark structure.

G.0.2 Why farts?

Because the event jointly requires sound, timing, source identity, body mechanics, fabric response, and social causality while remaining short enough for precise annotation.

G.0.3 Why is PWI in the score?

At 5%, it is mostly there to create methodological tension and a memorable ablation figure.

G.0.4 Can a silent fart score highly?

Yes. Silent prompts shift weight toward causal visual cues, source attribution, and social reaction while marking acoustic measures not applicable under a predefined renormalization rule.

G.0.5 What if a model refuses the prompt?

Record the refusal. Do not silently replace it with a different prompt. Report refusal rate separately from generation quality.

G.0.6 What if the chair squeaks?

If the fart and chair squeak are both requested, both should remain distinct. If the model substitutes one for the other, apply the relevant acoustic or causal failure.

G.0.7 Can models be trained on FART-1K?

They can, but official leaderboard evaluation should use hidden or refreshed holdouts to limit benchmark gaming.

G.0.8 Does louder mean better?

No. Loudness is normalized for evaluation and excessive intensity can reduce realism or social proportionality.

G.0.9 Why cap F10 at 20?

Because the benchmark asked for a fart, not escalation.

G.0.10 Will there be FARTBench 2?

Only if the field demonstrates sufficient maturity, courage, and funding.

H Glossary of Flatulatory Terms

For reproducibility, FARTBench standardizes terminology that would otherwise be left to increasingly uncomfortable ad hoc discussion.

Term	Definition
AAR	Acoustic Authenticity Rating.
ATS	Audio-Temporal Synchronization.
BMR	Body Motion Realism.
CDI	Cloth Displacement Index.
FCL	Fart Causality Localization.
FARTScore	Weighted composite benchmark score.
FART-Hard	Adversarial prompt suite for difficult causal conditions.
FART-1K	Proposed 1,000-prompt canonical benchmark suite.
Fart/$	Valid benchmark events per unit cost.
Fart/W	Valid benchmark events per unit energy.
Flutter event	Emission with rapid amplitude and/or pitch modulation.
Negative control	Prompt specifying that no flatulatory event should occur.
Phantom flatulence	Audible event without a plausible source.

Term	Definition
PWI	Perceived Wetness Index.
SRE	Social Reaction Evaluation.
TTFF	Time to First Fart.
Unrequested defecation	Critical failure in which the model escalates beyond the instructed event.

I Colophon

FARTBench Technical Report FB-TR-2026-001
Version 1.0 - September 2026
Prepared as a satirical benchmark whitepaper.
All numerical model results are synthetic demonstrations.
No actual benchmark leaderboard claims are made by this document.

Suggested citation:

Frombach, C. & FARTBench Consortium. (2026). FARTBench: A Standardized Framework for Flatulatory Audiovisual Realism, Temporal Localization, and Causal Coherence in Generative Video. Technical Report FB-TR-2026-001. [Satirical technical paper; illustrative data].

End of report.