NJBDA 2026 · ROWAN UNIVERSITY · MAY 20, 2026

SceneTwin: a reference-free audit of audio description for blind and low vision users

A visual grounding and comprehension pipeline that scores whether an audio description actually conveys what a blind or low vision user cannot see, without needing a single human reference script.

Adarsha Mishra · Department of Computer Science · William Paterson University · Advisors: Dr. Nan Wang, Bohan Fan, Xiaoshan Wang

MAIN RESULT · RANKING AD BY HOW WELL IT SERVES BLV USERS

ρ = 0.929

- **54/54** professional AD pairwise wins
- **15/18** clips fully ordered
- **0.90-0.96** bootstrap 95% CI

What is SceneTwin?

An automated quality audit for audio description, the spoken narration blind and low vision users depend on to experience the same feeling as actually watching a video.

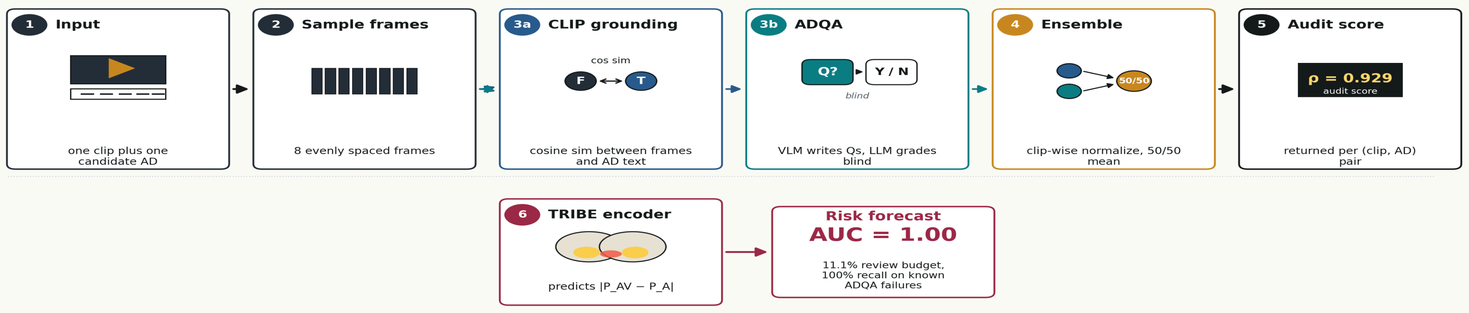
Audio description is how blind and low vision users experience video. SceneTwin takes one clip and one candidate description and returns a single audit score plus a flag for whether the score is likely to be reliable. The audit grounds the description against the actual video frames, not against another text, so no reference script is required and AD can be checked at the scale BLV users actually consume video.

Two reference free signals run in parallel. CLIP visual grounding measures cosine similarity between sampled frames and the description text and catches hallucinated or wrong scene descriptions. Frame grounded ADQA has a vision language model write comprehension questions from the frames, then a separate language model grades the description blind, catching vague or under specified text. The two signals are averaged with fixed equal weights.

A separate TRIBE fMRI encoder runs alongside the main pipeline and forecasts which clips are likely to be hard for the automated scorer, allowing a small fraction of clips to be flagged for human review with full recall on known scorer failures.

Methodology

Two reference-free signals run in parallel on each clip and candidate description. A separate fMRI encoder forecasts fragile evaluations from video and audio alone.



Methodology. One clip and one candidate description go in. Two reference free signals run in parallel and are averaged into a single audit score that tells whether the description preserves the visual content a blind or low vision user would otherwise miss. A separate fMRI encoder forecasts fragile evaluations from video and audio alone.

Why this matters

Audio description quality is becoming a scale problem, not just a captioning problem.

AI can generate descriptions quickly. The missing piece is an audit that checks whether those descriptions actually match the visual scene.

The problem

Bad AD can be present but still fail.

A description can exist, sound fluent, and still miss the action, identify the wrong scene, or omit the visual fact a blind viewer needs.

Availability is not quality.

Current QA options

Manual review
Human judgment is strongest, but does not scale to every generated clip.

BLEU / ROUGE
Reference text metrics need a gold script and miss visual grounding.

Visual-only proxies
They catch wrong scenes, but not vague or unhelpful descriptions.

What SceneTwin adds

Reference-free audit
Scores one candidate AD against the video itself, not against a single human script.

- **0** reference scripts
- **0.929** rho vs tier
- **54/54** pairwise wins
- **AUC 1.00** risk forecast

Problem statement. Roughly 285 million blind and low vision people worldwide rely on audio description to follow video. Availability alone does not tell whether an AD matches the visual scene a sighted user would see.

Problem: current AD QA does not scale to AI-generated descriptions

Accessibility standards require descriptions of visual information not available from audio alone. But practical QA still leans on availability checks, manual review, or narrow automated proxies.

Current / baseline	rho	wins	ordered	limitation
Manual / compliance QA	N/A	N/A	N/A	Accurate when expert-led, but slow; availability checks do not rank quality.
CLIP-only visual grounding	0.801	48/54	11/18	Misses vague but visually plausible descriptions.
Frame-grounded ADQA only	0.803	51/54	8/18	Strong comprehension proxy, but less stable than the ensemble.

SceneTwin

CLIP + frame-grounded ADQA + TRIBE triage

- **0.929** rho
- **54/54** pairwise wins
- **15/18** fully ordered

What it measures

Grounding: does the AD match the frames?
Answerability: can it answer visual questions?
Risk: should this clip be sent to stronger review?

Pilot evaluation on 18 complete clips; requires BLV-user validation.

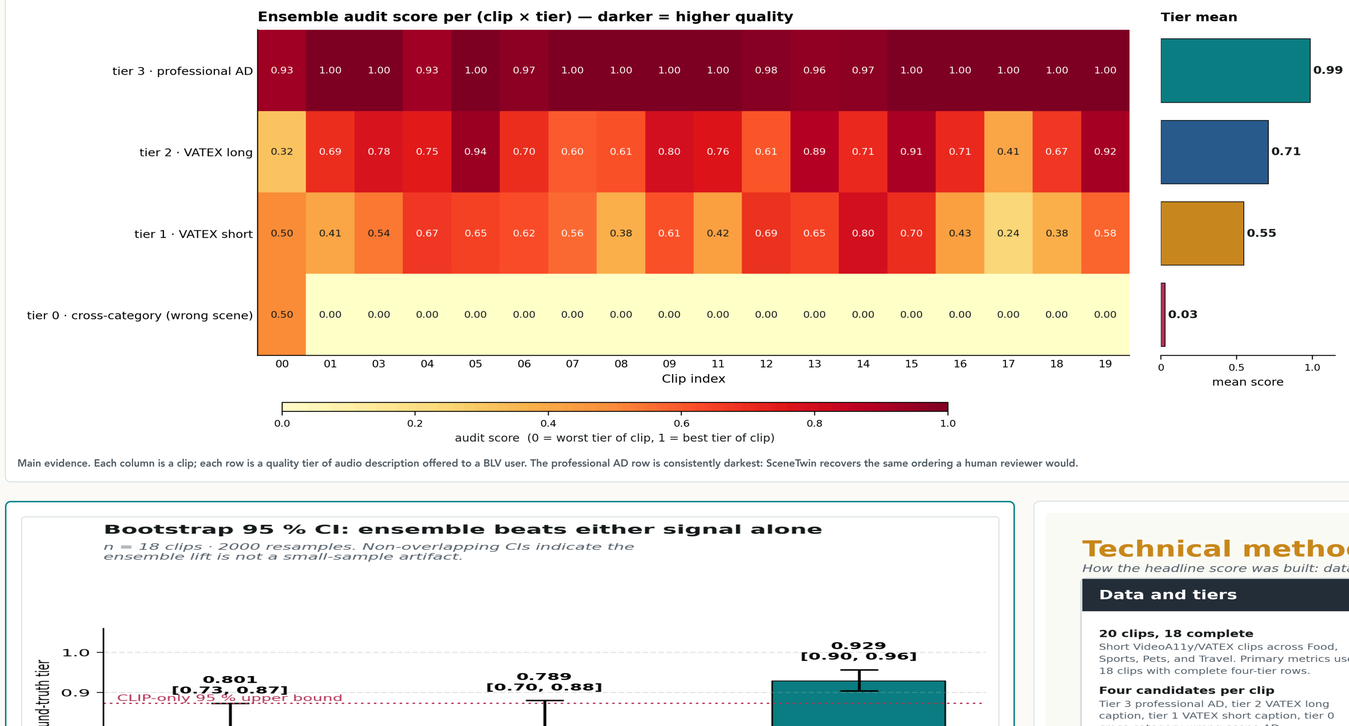
Automatic rows use SceneTwin's 18 complete clips x 4 quality tiers. Manual/compliance QA is included as the dominant practical baseline; it does not produce rho or pairwise-win metrics.

SceneTwin changes QA from 'does description exist?' to 'is it visually grounded, answerable, and likely reliable?'

Positioning. SceneTwin is not just checking whether AD exists; it tests grounding, answerability, and fragility at scale.

ρ = 0.929 [CI 0.90, 0.96] 54 / 54 pairwise wins · 15 / 18 fully ordered

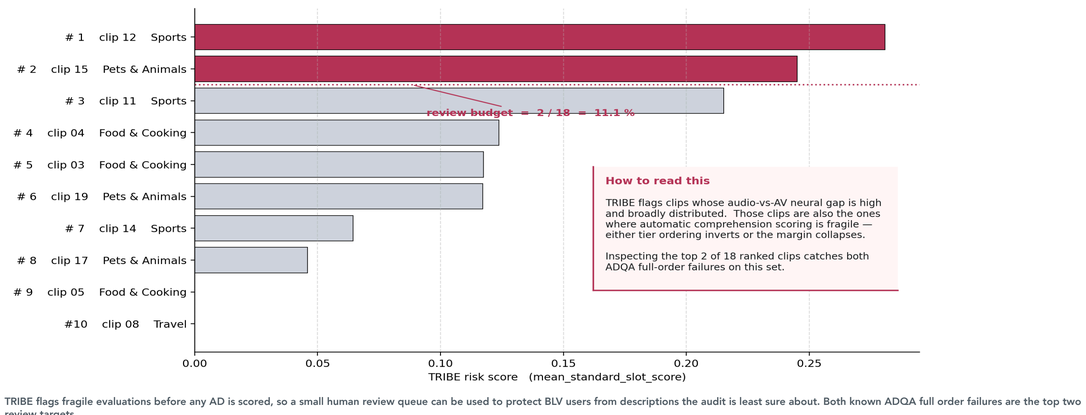
n = 18 clips × 4 quality tiers · permutation p < 0.0005



Main evidence. Each column is a clip; each row is a quality tier of audio description offered to a BLV user. The professional AD row is consistently darkest: SceneTwin recovers the same ordering a human reviewer would.

TRIBE flags fragile evaluations before any AD is scored

Top-10 clips by TRIBE risk forecast — red = clip the ADQA ensemble failed to fully order.
Computed from video + audio alone. Recall @ 11.1 % review budget = 100 % · ROC-AUC = 1.00 · uncorrected p = 0.0065 (Bonferroni p = 0.065). Pilot evidence.



TRIBE flags fragile evaluations before any AD is scored, so a small human review queue can be used to protect BLV users from descriptions the audit is least sure about. Both known ADQA full order failures are the top two review targets.

Bootstrap 95 % CI: ensemble beats either signal alone

n = 18 clips · 2000 resamples. Non-overlapping CIs indicate the ensemble lift is not a small-sample artifact.

Ensemble confidence interval sits above the CLIP-only upper bound.

Technical methodology

How the headline score was built: dataset, tiers, reference-free signals, validation, and TRIBE risk forecast.

Data and tiers

20 clips, 18 complete
Short VideoA11y/VATEX clips across Food, Sports, Pets, and Travel. Primary metrics use 18 clips with complete four-tier rows.

Four candidates per clip
Tier 3 professional AD, tier 2 VATEX long caption, tier 1 VATEX short caption, tier 0 cross-category wrong-scene AD.

Ground truth rank
Tier 0 = 0, tier 1 = 1, tier 2 = 2, tier 3 = 3. Evaluation is within-clip tier ordering.

Reference-free scoring

Signal 1: CLIP
Eight evenly spaced frames are embedded with CLIP ViT-L/14 and compared with the candidate AD text by cosine similarity.

Signal 2: frame-grounded ADQA
GPT-4o writes visual comprehension questions from frames. Claude grades the candidate AD blind, seeing no frames or tier labels.

Fixed ensemble
Both signals are min-max normalized within clip, then averaged 50/50. No learned weights or reference script are used.

Validation and risk

Primary statistics
Spearman rho = 0.929 across 72 clip-tier pairs; bootstrap 95 percent CI [0.904, 0.957]; permutation p < 0.0005.

Ordering checks
54/54 pairwise tier wins and 15/18 clips fully ordered. Length-only baseline reaches only rho = 0.318.

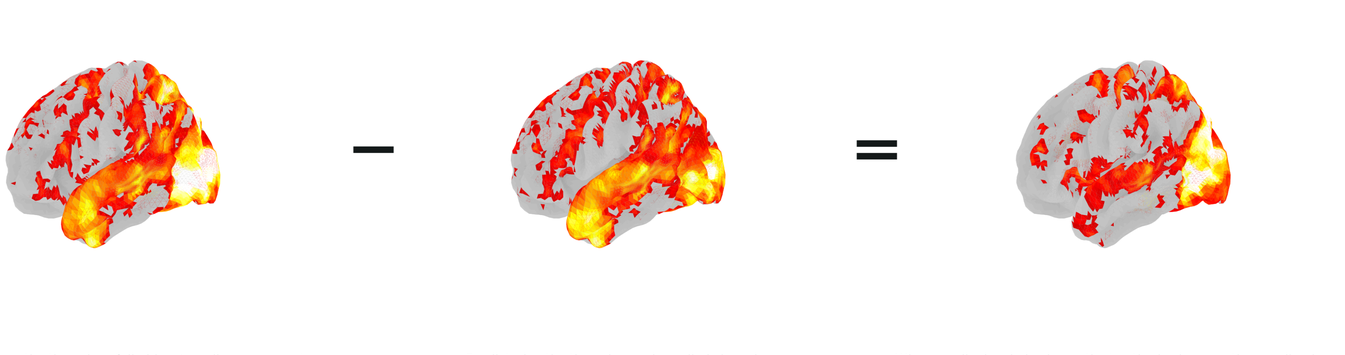
TRIBE side analysis
TRIBE v2 predicts |P_AV - P_A| from video/audio alone. Its risk feature gives ROC-AUC = 1.00 for known ADQA full-order failures.

Source: output/reports/scenetwin-technical-methodology.md

Limitations: n = 18 headline clips; LLM-judged ADQA; TRIBE risk is pilot evidence.

Technical methodology summary. Full details are in output/reports/scenetwin-technical-methodology.md.

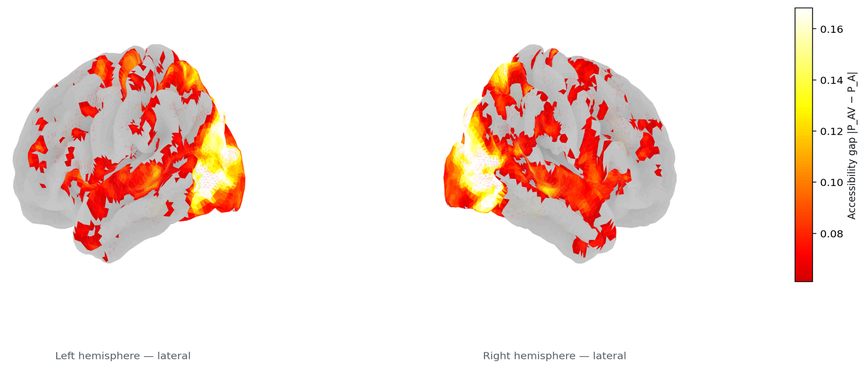
How the TRIBE accessibility gap is built



Brain-model visual: useful for timing and risk triage, not the final text-quality score.

Average accessibility gap across 20 clips

TRIBE average gap |P_AV − P_A| across audiovisual vs audio-only viewing



Average TRIBE accessibility gap across 20 clips. The bright regions are where audiovisual viewing drives brain activity that the audio track alone cannot recreate. This is exactly the information BLV users depend on audio description to recover.

What did not survive validation

Each of these branches looked plausible on a whiteboard. Each was measured against ground-truth quality tiers or professional audio description. Each was dropped before being claimed in the final pipeline.

1. Description Gain (MVRR) DROPPED

Hypothesis.
A good AD should fill the neural gap between audio-only and audiovisual viewing.

Why it failed.
Unstable on the two-clip TRIBE smoke test; unrelated AD sometimes beat matching AD.

2. ROI content typing DROPPED

Hypothesis.
High-gap visual ROIs should reveal what type of content an AD needs to cover.

Why it failed.
Glasser ROI typing reached 19.0% agreement versus 16.7% chance against pro AD.

3. Neural closure DROPPED

Hypothesis.
Adding AD text to audio should move predicted brain response back toward full video.

Why it failed.
Closure stayed negative; shorter captions could beat professional AD because language volume dominated.

4. TRIBE-weighted ADQA DROPPED

Hypothesis.
High TRIBE-gap moments should receive more weight in frame-grounded ADQA.

Why it failed.
Null result: Delta rho < 0.002 versus uniform ADQA weighting.

Takeaway: reference-free visual/comprehension scoring survived; direct neural text scoring did not.

Negative results kept the final claim narrow: reference-free scoring survived; direct neural text scoring did not.

References

1. Radford, A., et al. (2021). *Learning Transferable Visual Models From Natural Language Supervision.* ICML.
2. d'Ascoli, S., et al. (2026). *A foundation model of vision, audition, and language for in-silico neuroscience.* arXiv:2605.04326.
3. Li, C., et al. (2025). *VideoA11y: Method and Dataset for Accessible Video Description.* arXiv:2502.20480.
4. Wang, X., et al. (2019). *VATEX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research.* ICCV.

Acknowledgements

Supported by NSF Grant #2028011. Thank you to Dr. Nan Wang, Bohan Fan, and Xiaoshan Wang for their guidance, and to the Department of Computer Science and the College of Science and Health at William Paterson University for their support.