Problem Definition

Problem Visual media such as comics are largely inaccessible for people who are blind or have vision impairments. Existing tools either (a) read plain text aloud or (b) describe images; few provide a unified, language-flexible experience tailored to the narrative structure of comics. Ensuring equitable access to this content is an ethical imperative.

Solution Enable a user to upload one or more comic pages and receive an accurate, engaging audio narration in their preferred language. This narration must faithfully convey the story, dialogue, and scene transitions of the comic.

Why it matters This unlocks independent access to cultural and informational content and supports inclusion for visually impaired readers.

Inputs Comic page(s) as images, user language preference, and possibly pacing/voice settings

Outputs Synthesized speech audio description of panels and dialogue in the user's language

Subproblems

1. Creating a simple interface for users to upload pictures of comics and specify their preferred output language
2. Creating an accurate and engaging text description of the comic without leaving out parts or confusing the order of cells. (Might be achieved by using a multimodal large language model)
3. Translating the text description into the user-preferred language. (Via a translation model, trained by our team members)
4. Creating speech based on the translated text description
5. Presenting the created speech in an intuitive way to the user

Objectives

Hierarchy (4 levels)

Level 1 - North-star outcome

Any user can understand a comic's full narrative and dialogue in their preferred language via a simple "upload -> play" flow.

Level 2 - Strategic goals

- **G1. Accessibility & Experience** - Functional parity with sighted reading for comprehension and enjoyment
- **G2. Quality & Fidelity** - Faithful, coherent story rendering with minimal errors
- **G3. Efficiency & Reliability** - Predictable processing time for multi-page comics and stable operation

Level 3 - Objectives (aligned to goals)

- **O1.1 (G1)**: Support multi-page uploads and preserve correct panel order
- **O1.2 (G1)**: Distinguish dialogue, narration, and SFX; maintain speaker turns
- **O2.1 (G2)**: Produce high-adequacy English descriptions with low omission/hallucination rates
- **O2.2 (G2)**: Translate into at least one other language
- **O2.3 (G2)**: Generate clear, natural speech with appropriate prosody for dialogue
- **O3.1 (G3)**: Keep end-to-end latency within a defined target per page and for full books
- **O3.2 (G3)**: Handle peak loads without degradation; provide basic health monitoring

Level 4 - Indicative milestones

- **M1**: Demonstrate faithful meaning reproduction and correct ordering on a diverse sample
- **M2**: Validate translation accuracy on test corpus
- **M3**: TTS prosody patterns for dialogue vs. narration validated by pilot listeners
- **M4**: End-to-end pilot ("upload -> play") on 10 comics; collect feedback on quality and latency

Measurements / KPIs

Comprehension & accessibility (G1).

- Narrative comprehension score from user tests (>= 90 %).
- Reading-order accuracy across pages (>= 95 %).
- Dialogue attribution accuracy (>= 95 %).

Quality & fidelity (G2).

- Description adequacy/fluency (subjective rating).
- Translation adequacy & consistency (measured on test-set).
- Prosody naturalness & intelligibility (listening tests).
- Hallucination/omission rate (< 5 %).

Efficiency & reliability (G3).

- End-to-end latency per page (< 10 seconds).
- Failure rate/timeouts (<= 5 %).

- Cost/energy per processed page (tracked trend; target reductions over time).