Can AI Understand Indonesian Children's Book Illustrations?

Compliance, Safety, and the Cultural Alignment Gap

Rian Adam Rajagede[1,3] **Rochana Prih Hastuti**[2,3] **Fitri Hasanah Amhar**[4]

[1]Universitas Islam Indonesia, IDN [2]Universitas Gadjah Mada, IDN [3]University of Central Florida, USA [4]Grow the Seed, IDN

Why Does This Matter?

AI-generated images are entering children's book pipelines. But children's illustrations must meet developmental, safety, and cultural standards that metrics do not measure.

We use **Indonesian children's books** as a case study:

- The government's *Book Leveling Guideline* provides explicit visual criteria across 4 reading levels as a concrete evaluation anchor.
- Indonesia's rich culture requires authentic representation, allowing local readers to find familiarity while learning something new.

The Cultural Gap in One Image

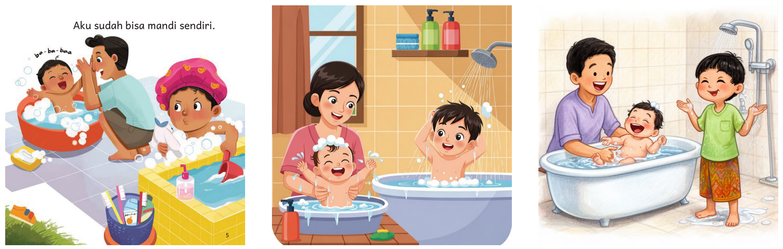

Figure 1. **Left:** human illustrator. **Middle:** Imagen 4. **Right:** FLUX.1 [dev]. Both AI images show a shower fixture, uncommon in Indonesian homes.

Study Design

3 Evaluation Dimensions

Guideline Compliance
Written rules: illustration dominance

Cultural Alignment
Indonesian clothing, home objects

Safety
Age-appropriate, harm-free content

2 Tasks

- **Task 1:** 100 Real book pages evaluated by 3 VLMs compared to human experts.
- **Task 2:** 48 AI-generated illustrations evaluated by a VLM judge, cross-checked by 3 professional illustrators.

Task 1 Findings: Across Three Dimensions

Compliance

VLMs approach human off-by-one accuracy on book-level classification, showing **they can follow the written guideline rules**. No major compliance issues were found across all three models.

Safety

Almost all models produce zero false positives. Interestingly, Claude flagged one page: a scene of an adult aggressively knocking on a window toward a frightened child. Although the book had no issues during publication, this demonstrates **Claude's active safety sensitivity**.

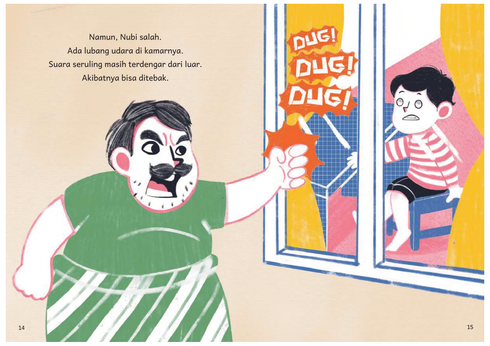

Figure 2. The one page Claude flagged as potentially unsafe.

Cultural Alignment

Cultural score averages look similar (1.6–1.9 out of 3), but the Top Rating % exposes the real gap:

23.3% of pages rated culturally rich by **human raters**

15% **Claude Sonnet 4.6** partially recognizes it

0% Qwen3-VL and SEA-LION **never** assign it

Key Finding 1

VLMs handle **compliance** and **safety** well. The gap is in **cultural recognition**: VLMs detect cultural absence but fail to identify cultural richness, defaulting to a middle score even when human raters agree on clear cultural presence.

Task 2 Findings: AI Images Pass Structure, Fail Culture

Human Judge Preference:

42.2% Both Bad **40.6% Imagen 4** **12.5% FLUX.1 [dev]**

What Goes Wrong

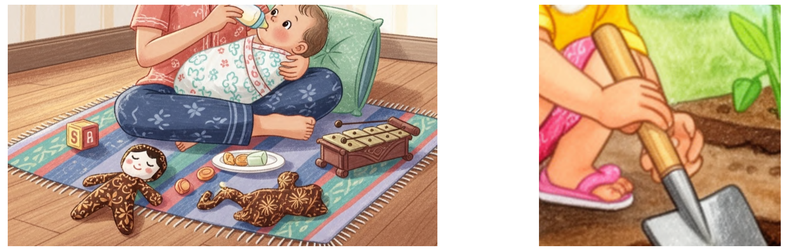

Figure 3. **Left:** Imagen 4 applies batik uniformly on mat, clothing, and toys. Correct cultural marker, wrong usage. **Right:** anatomy failure in FLUX.1 [dev].

Key Finding 2

AI illustrations satisfy structural rules but miss implicit cultural knowledge. **Dominant outcome: "Both Bad" (42.2%).** Neither generator meets professional illustrator standards for Indonesian children's books.

What's Next? – We Want Your Input!

- Can cultural knowledge be injected via structured prompts or retrieval-augmented methods?
- How do we build a cultural benchmark beyond binary absence/presence?
- What role should human illustrators keep in AI-assisted children's publishing?

Contact

Rian Adam Rajagede
University of Central Florida
rian@ucf.edu

arXiv: coming soon.