

Not Every Word: Measuring and Redesigning Burned-in Captions for Short-form Vertical Video

Hocky Yudhiono

Independent researcher, Hong Kong SAR

`its@hocky.id`

8 October 2026

Abstract

Short-form vertical video is watched in feeds, often with the sound off, and its captions are burned into the picture, where they serve both as the accessible record of speech and as graphic design. This paper measures the burned-in captions of a working channel of 25–45-second Indonesian explainer Shorts presented by a VTuber, tests a fix in the channel’s own caption pipeline, and derives a design from the results. Across 533 caption pages in 11 projects, pages last a median of 0.62 s and 74% are shorter than Netflix’s 5/6-s minimum, although the speaking rate (134–197 words a minute) is within broadcast norms: the cause is pagination, not speed. A 0.8-s minimum page time, implemented as the pipeline’s default and run on 12 projects at 1080×1920, moved the share of short pages only from 95% to 94%, because the large one-line layout leaves no room to merge pages; a simulated two-line budget brings it to 17%. Grounded in subtitle standards (Netflix, BBC, DCMP, FCC, WCAG), reading and emoji research, and a structured analysis of Japanese *telop*, Korean entertainment captions and Chinese decorated text, I propose two layers: an exported subtitle track that never changes, and a burned-in page that may step aside, add an emoji or a picture, or change style, within limits graded by the strength of their evidence.

Keywords: captions; subtitles; *telop*; short-form video; accessibility; kinetic typography; VTuber

1 Introduction

A Short is decided in its first seconds. Viewers scroll a feed, many with the sound off: in a 2019 survey of 5,616 US adults, 69% said they watch video with the sound off in public and 25% even in private, and 80% of caption users had no hearing impairment (StreamingMedia, 2019). On the channel studied here the captions are drawn into the frame by the editor, so every viewer sees them and none can switch them off.

That makes the burned-in caption a design object as much as an accessibility feature. Subtitle standards were written for switchable subtitles on 16:9 screens, and they ask for verbatim text, phrase-shaped line breaks and a minimum display time. Short-form fashion asks for big, bouncing, colourful words. East Asian variety television has spent forty years on a third answer: an editor’s second voice that sometimes repeats what is said, sometimes adds what is not said, and sometimes leaves the words alone.

This paper studies one working case, the author’s channel of 25–45-second Indonesian explainer Shorts presented by a VTuber, whose captions are produced by the author’s own editing pipeline. It asks three questions:

RQ1

How long do the channel’s caption pages stay on screen, and why?

RQ2

Does a minimum page time, built into the caption pipeline, fix the problem?

RQ3

When should burned-in text repeat speech, add to it, or replace it, and under what limits?

Contributions.

1. **A measurement** of 533 caption pages from 11 projects (RQ1, Section 5), showing that pages are far shorter than every subtitle standard’s minimum while the speaking rate is normal.
2. **An implementation study** of a 0.8-s minimum page time in a production caption pipeline, run on 12 projects (RQ2, Section 6), showing that the rule barely changes the result because the one-line layout is the binding constraint, and that a second line would.
3. **A two-layer caption design** (RQ3, Section 7) with evidence-graded guidelines (Section 8), grounded in subtitle standards, reading research and a structured analysis of Japanese, Korean and Chinese variety practice, condensed into five relations between on-screen text and speech.

Sections 2 and 3 give the background, Section 4 the method, Sections 5 and 6 the two studies, Sections 7 and 8 the design and guidelines, and Sections 9–11 future work, limitations and conclusions.

2 Background: standards and reading research

Claims in the background carry one label each: [STANDARD] a published guideline; [STUDY] research with a method and a sample, whose size is given because several are small; [PRACTICE] what editors and shows do, documented by a researcher, broadcaster or the show’s own staff; [OPINION] a vendor’s, a blogger’s or the author’s judgement, never presented as evidence. *Second-hand* marks a claim read in a source that cites it rather than in the original.

2.1 Subtitle standards

Table 1 condenses the subtitle guidelines of Netflix (Netflix, 2022, 2025a,b), the BBC (v1.2.5, the only one with explicit 9:16 numbers; BBC, 2026), DCMP (Described and Captioned Media Program, 2026), the FCC (US Code of Federal Regulations, 2026) and WCAG 2.2 (W3C, 2023a,b). They were written for switchable subtitles, so they are the floor for readability, not a style guide for Shorts. The platforms publish safe zones only for ads (Google Ads Help, 2026; TikTok Ads Help, 2026); the pipeline studied here uses its own measured zones (the bottom fifth for the player’s controls, a right rail from 86% of the width), stricter than anything the platforms publish.

2.2 Reading research

- **Speed is less of a problem than chopping.** In eye tracking, 74 viewers kept up with subtitles at 12, 16 and 20 characters a second without losing comprehension (Szarkowska and Gerber-Morón, 2018) [STUDY]; another study already saw a cost at 20 cps and a clear one at 28 (Kruger et al., 2022) [STUDY]. Breaks that cut through phrases (Figure 1) raised effort and frustration without lowering comprehension (Gerber-Morón et al., 2018) [STUDY].
- **Building a page word by word pulls the eye.** With live subtitles shown word for word, viewers switched between picture and text more often than with phrase chunks

Table 1: Subtitle standards, condensed [STANDARD].

Topic	What they say
Minimum time	Netflix 5/6 s (20 frames at 24 fps); DCMP 40 frames; BBC about 0.3 s per word, never leading speech or lingering over 1.5 s after it.
Reading speed	Netflix Indonesian 17 characters a second for adults, 20 for SDH; BBC 160–180 words a minute; DCMP 130–160 wpm by school level.
9:16	BBC: up to 3 lines, about 25 characters a line at a line height of 3.9–4.5% of the frame, 90% of the width; “a little higher” than in 16:9 but still in the lower third.
Line breaks	Break after punctuation and before conjunctions and prepositions; never split article and noun, adjective and noun, first and last name, or the parts of a verb (Netflix; BBC 3.4; DCMP).
Placement	Never over faces, mouths or featured on-screen text (FCC; BBC 10.1; Netflix; DCMP); open captions must not hide relevant information (WCAG).
Text already on screen	Netflix deletes forced narratives identical to on-screen text and never mixes on-screen text with dialogue in one subtitle; BBC 2.10 requires subtitles to carry on-screen text that is illegible at streaming size.
Emphasis and speakers	BBC: capitals for stress, sparingly; one colour per speaker, kept for the whole programme. Netflix Japanese: side dots (bōten) for emphasis.
Verbatim and sound	FCC: the spoken words in order, no substitution or paraphrase except for time; BBC: prefer verbatim, never simplify; WCAG 1.2.2: captions include sound effects, music and speaker identification.

(Rajendran et al., 2013) [STUDY]. Many short-form looks, including two of the studied pipeline’s, build their pages this way (Figure 2).

- **Muted viewers depend on every page.** Without sound, 168 viewers had higher cognitive load and lower comprehension and leaned on the subtitles (Szarkowska et al., 2024) [STUDY].
- **Captions help most people.** A review of more than 100 studies found captions improve comprehension, memory and attention, especially for second-language viewers and deaf and hard-of-hearing viewers (Gernsbacher, 2015) [STUDY].
- **Placing captions near the action** gave hearing-impaired viewers gaze patterns closer to those of viewers without subtitles (Brown et al., 2015) [STUDY].
- **Redundancy cuts both ways.** On-screen text identical to narration impaired learning, slightly different text helped, and learners still preferred identical text (Yue et al., 2013) [STUDY]. Partial captions synchronised to speech matched full captions for second-language learners (Mirzaei et al., 2014) [STUDY], while irregular keyword captions distracted (Montero Perez et al., 2014, second-hand via Mirzaei et al. 2014).

2.3 Emphasis and motion

Every source that discusses frequency says to emphasise rarely. BBC 11.2 warns that text sprinkled with capitals is hard to read [STANDARD]; a highlighted word is the one learners remember (17 children; Finger Bou and Muñoz Lahoz, 2024) [STUDY]; and in Japanese variety shows the telops that print whole utterances stay white while colour is kept for the cut-out keyword (Morikawa, 2023) [PRACTICE].

For how to emphasise, the best evidence comes from deaf and hard-of-hearing viewers: across three studies with 39 participants, font colour combined with weight, or colour combined with size, was preferred over seven other typographic encodings and best let viewers recognise the speaker’s emotion (de Lacerda Pataca et al., 2024) [STUDY]. Loudness mapped to weight, pitch

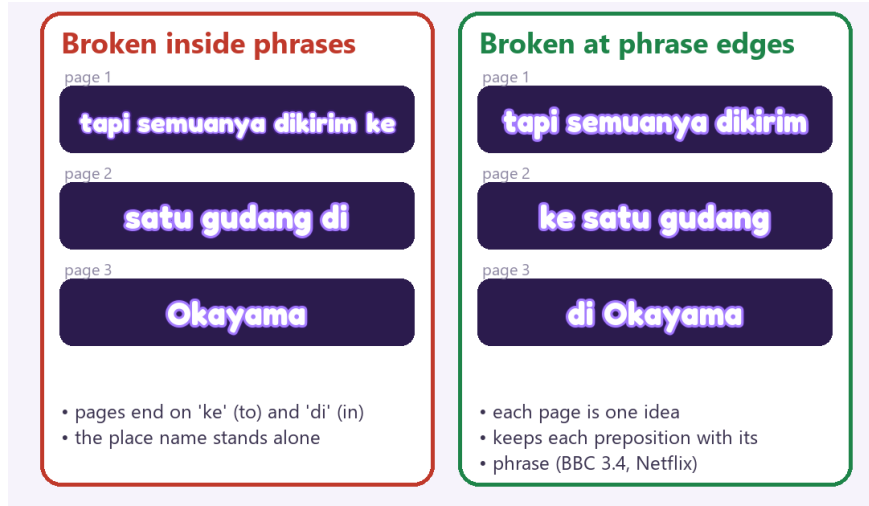


Figure 1: Paging by phrase. One sentence from the channel (“tapi semuanya dikirim ke satu gudang di Okayama”, *but all of it is shipped to one warehouse in Okayama*) paged inside phrases (left) and at phrase edges (right). Mock-up.

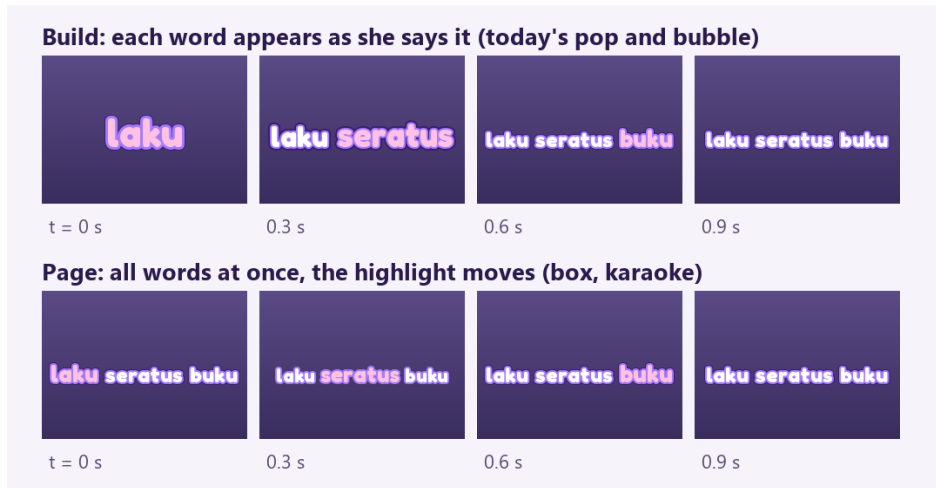


Figure 2: One three-word page (“laku seratus buku”, *sells a hundred books*) built word by word (top) or shown whole with a moving highlight (bottom). Mock-up.

to baseline and duration to spacing let 117 people match text to its audio 65% of the time, whether the text moved or not (de Lacerda Pataca and Costa, 2023) [STUDY]. In a small eye-tracking test, font changes and animated effects scattered viewers’ gaze (preliminary; Sasamoto and O’Hagan, 2020) [STUDY]. So motion should mark time (a page appears when it is spoken, the highlight follows the voice) and only one thing should move at a time.

2.4 Emoji and pictures

Replacing a word with a picture has a measured cost. In self-paced reading, an emoji standing in for a noun took about 800 ms against about 450 ms for the word, and over 900 ms when it depicted a homophone (Scheffler et al., 2021, as reported by Ruhr-Universität Bochum 2021) [STUDY]. In eye tracking, replacement emoji drew longer total viewing than words and ambiguous emoji were worse on every measure, while an emoji placed *after* its word drew less viewing than the word itself (Paggio and Tse, 2022) [STUDY]. Very frequent emoji behave almost like words in Chinese text, at about 50 ms extra (Lu et al., 2025) [STUDY]; the authors suggest a logographic script makes emoji cheaper, so the English figures are the better guide for Indonesian.

Deaf and hard-of-hearing users did not prefer emoji captions in a preliminary study ($n = 4$; Oomori, 2022) [STUDY]. On a page that lasts 0.6 s, a third of a second is a large share.

3 Background: how variety television decides

Japanese telop (*teroppu*), Korean entertainment captions (*yeneung jamak*) and Chinese “flower text” (*huazi*) are open captions added in post-production by the producers. They change colour, font and size with the tone of the words, show fragments rather than whole statements, and are explicitly not an aid for deaf viewers (Higashino, n.d.; Dublin City University, 2016; Sasamoto and O’Hagan, 2020). They are the closest professional practice to a designed Short, and they have decided, show by show, when text should repeat, add or step back. Figure 4 summarises their devices.

3.1 Japan: Getsuyō kara Yofukashi

Morikawa (2023) surveyed 7 news and 15 prime-time variety programmes broadcast between 31 August and 7 September 2022 and graded their telops from ◎ (the whole utterance) through ○ (a summary), △ (a keyword or fragment) to × (nothing) [PRACTICE]. Every one of the 15 variety programmes put more telop on filmed segments than on studio talk. Studio talk repeats what the film already showed, filmed speakers are mostly members of the public whose speech is less clear, and full telops on studio talk would remind viewers that it is recorded.

Getsuyō kara Yofukashi (Nippon TV) is the textbook case. Its street interviews are ◎: almost every word printed, in white with an edge colour per speaker. Its studio segments with Matsuko Deluxe are △: keywords only, in colours chosen by what is said. Across the survey, full-utterance telops stay in one colour family, mostly white, because long colourful telops are hard to read; cut-out keywords get bright stock colours by emotion (Figure 3). Nakano (2018) names the show among those whose telop acts as the comedy partner that delivers the *tsukkomī*, pointing out what interviewees themselves have not noticed is odd [PRACTICE].

The show also supplies the cautionary tale. In March 2025, Nippon TV admitted that a 24 March broadcast had edited a street interview so that a Chinese woman appeared to say something she had not said; the network’s president called it beyond the bounds of production, and street interviews were halted pending new rules (Daily Sports, 2025; Shūkan Josei PRIME, 2025). On-screen text must never put words in anyone’s mouth.

3.2 Korea: Running Man

Comparing Japanese and Korean variety, Ohe (2021) found Japanese telops mainly aid understanding, while Korean shows use far more captions for entertainment: lines rewritten for comedy, the speaker’s feeling set before the line like a stage direction, word play, written onomatopoeia and expression marks, and inner monologue in parentheses [STUDY]. Korean shows also give animals and objects speech and labels throughout, where Japanese shows do so only in comic set-ups (Ohe, 2022) [STUDY].

On SBS’s *Running Man*, a frame from 2 June 2019 shows the devices together: a small mood label above a member’s line, the line in large white type with a dark edge, a red onomatopoeia for coughing placed beside the cougher, and an exclamation mark for surprise (Twig24 (Seoul Shinmun), 2019) [PRACTICE]. The captions are also the production team’s voice: in December 2016 the team apologised to viewers on air in a caption (Tenasia, 2016), and Park (2015) shows how *Infinite Challenge*, *2 Days & 1 Night* and *Running Man* use captions to state members’ characters and to comment from an omniscient position [STUDY]. That voice carries risk: viewers read the wording of a caption in the same 2019 episode as mocking a 1987 torture death, and SBS promised more care (Twig24 (Seoul Shinmun), 2019).

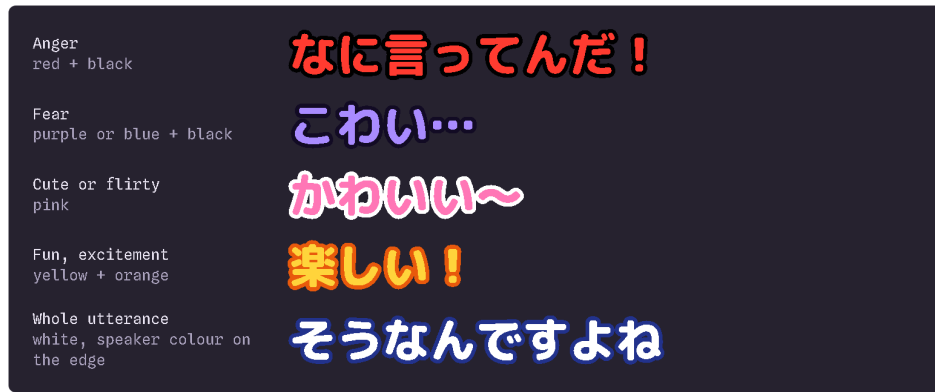


Figure 3: Stock telop colours found by Morikawa (2023) in 22 Japanese programmes: anger red with black, fear purple or blue with black, cute pink, fun yellow with orange, and whole utterances white with the speaker’s colour on the edge. The sample lines are invented and the lettering redrawn; the colour rules are his.

3.3 China: huazi

The Chinese industry name *huazi* covers decorated post-production text with images and sound effects (Nan and Chen, 2017) [PRACTICE]; Zeng (2017) calls it a “third text” over the script and the picture, which activates background knowledge, marks the focus and guides interpretation [STUDY]. Hunan TV’s *Dad, Where Are We Going?* (2013) made it famous: tags explained who people were for viewers who do not follow celebrities, and a Zhihu user counted explanatory huazi at over 60% of one episode (second-hand, via Nan and Chen 2017). *Keep Running* (the licensed Running Man) repeated game rules on screen so viewers could follow, and recalled a running gag in new forms rather than labelling it the same way each time (Zeng, 2017).

An editing tutorial analysing *Qipa Shuo* describes three kinds of huazi, one per point of view: emphasis (the speaker’s key phrase repeated), description (a bystander’s reaction, exaggerated) and creative (the editor’s commentary and imagined inner thoughts) (Banpie Sishu, 2017) [PRACTICE]. And rather than write “inwardly helpless” over a guest, editors use a sweat-drop sticker, an ellipsis and a sound effect (Nan and Chen, 2017) [PRACTICE]: the picture adds a feeling nobody put into words.

3.4 VTuber clip channels

The evidence here is thin. I found no editor interview or analysis of hololive or Nijisanji clip-channel (*kirinuki*) telop conventions that could be opened; one Japanese article on the subject sits behind a members-only login and was left out. Clip channels work like variety shows (a recording of people talking, text added in post), so the sections above apply. The one VTuber-specific borrowing is a fixed colour per talent, which also matches BBC’s speaker-colour rule [OPINION].

3.5 Five relations between text and speech

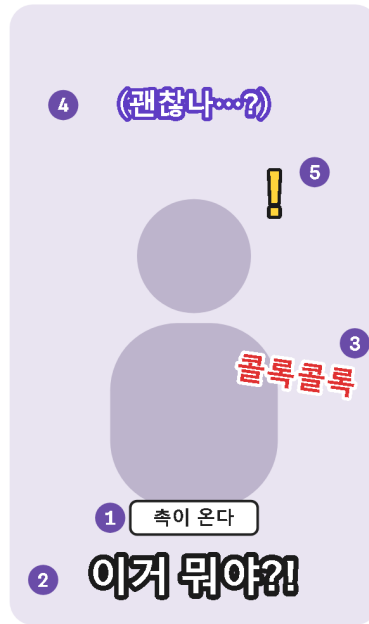
Table 2 condenses the three traditions. Two things hold across all of them. First, the plain layer and the designed layer are kept apart: in another style, often in another place, and never pretending to be the subtitle. Second, the designed layer uses a small, consistent vocabulary (stock colours, fixed member colours, recurring tags) that regular viewers learn.

Japan · テロップ



- 1 whole line: white, speaker edge
- 2 keyword: stock colour
- 3 sound word (onomatopoeia)
- 4 a picture for a word

Korea · 예능 자막



- 1 mood label before the line
- 2 the line (often rewritten)
- 3 sound word 4 inner voice ()
- 5 reaction mark

China · 花字



- 1 the plain subtitle
- 2 the same line, decorated
- 3 who they are 4 the moment
- 5 a sticker for a feeling

Figure 4: Anatomy of variety-show text, redrawn as a schematic. The devices are those documented in Section 3 (Higashino, n.d.; Sasamoto, 2021; Morikawa, 2023; Ohe, 2021, 2022; Twig24 (Seoul Shinmun), 2019; Nan and Chen, 2017); the sample lines and the name are invented, and the figure is a grey placeholder.

Table 2: What variety editors do with on-screen text, and where the design of Section 7 sits.

Relation	Examples	Design
Repeat all	white full-utterance telop on filmed segments; the channel's captions today	the base layer
Repeat the key part, decorated	keyword telop; emphasis huazi; a host's catchphrase shown big	(d) style; keyword emphasis
Add what is not said	onomatopoeia, inner voice, mood labels, role tags, the <i>tsukkomi</i>	a separate lane
Replace a word with a picture	an egg picture inside a telop in place of the word for egg (Higashino, n.d.)	(b) emoji, (c) picture
Leave the words to something else	no telop on studio talk; hosts who refuse telop	(a) step aside

4 Method

4.1 Review of standards, research and practice

The background was built from sources read on 7–8 October 2026: published subtitle guidelines; eye-tracking and reading studies on subtitles, captions and emoji; and research papers, trade press and practitioners’ accounts of Japanese, Korean and Chinese variety editing. Every cited source was opened. Where a page could not be opened (a paywall or a bot check), the claim was dropped or marked as second-hand, and each claim was labelled as described in Section 2.

4.2 Study 1: measuring the channel’s captions

The channel’s captions are produced by its own editing pipeline, Poiesis, which the author develops. Word timings come from speech recognition (ElevenLabs Scribe v2, or Whisper large-v3), are corrected by hand for recognition errors only, and are grouped into pages of at most one line of 16 characters and 4 words. I read every project’s caption file in the channel’s projects folder (11 projects, 533 pages, measured 8 October 2026). A page lasts from its first word’s onset to the next page’s onset when the gap is under 0.5s, otherwise to its last word’s end. I report page duration, words per page, characters per second per page and speaking rate, against the standards’ minimum durations and reading speeds.

4.3 Study 2: testing a minimum page time in the pipeline

After Study 1, a minimum page time of 0.8s and phrase-aware breaks were built into the pipeline as the default (Section 6). To measure their effect, I ran the pipeline’s own caption code, read-only, on all 12 projects that had captions by then, at 1080×1920: the default look (`bubble`) and the older house look, each with and without the minimum. I also simulated a page budget of two lines (32 characters, at most 6 words) with the same joining rule, ignoring font metrics. Each condition reports page count, median duration, the shares under 5/6s and 0.5s, and words per page.

4.4 Grading the guidelines

Section 8 grades each guideline: **A**, standards and at least one study agree; **B**, one or two studies, or consistent documented practice; **C**, practice or opinion only, to be tested.

5 Study 1: the channel’s caption pages

The speaking rate is normal for subtitling [STANDARD]; the pages are not (Figure 5, Table 3). Pauses over 0.5s, the one-line rule and the 16-character limit split the speech into flashes of under two words. The high per-page reading rate comes mostly from the short pages: merging two neighbours averages out the spikes. With the audio on, a hearing viewer can follow by ear, so the cost falls on muted viewers, deaf and hard-of-hearing viewers and anyone reading along, the very viewers captions exist for.

A second problem is placement (Figure 6): captions drawn over the text of on-screen news boards. There the board text is Japanese and the caption Indonesian, so the board cannot stand in for the spoken words.

Answer to RQ1. The pages last a median of 0.62s and three in four are shorter than Netflix’s minimum, while the speaking rate is normal. The cause is pagination: pauses, the one-line rule and the character limit.

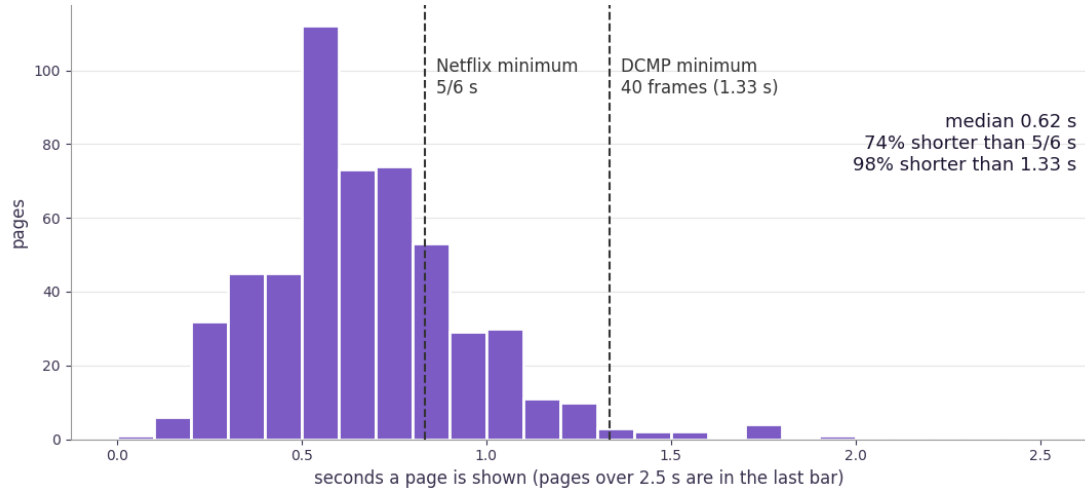


Figure 5: How long the channel’s caption pages stay on screen: 533 pages from 11 projects, measured on 8 October 2026. Dashed lines: Netflix’s minimum event duration (5/6 s) and DCMP’s (40 frames, about 1.33 s).

Table 3: The channel’s caption pages (11 projects).

Measure	Value	Reference
Pages	533	
Median page duration	0.62 s	Netflix minimum 0.83 s; DCMP 1.33 s; BBC about 0.3 s a word
Pages shorter than 5/6 s	74%	
Pages shorter than 0.5 s	24%	
Pages shorter than 40 frames (1.33 s)	98%	
Words per page (mean)	1.86	
Pages over 17 characters a second	45%	Netflix Indonesian: 17 cps adults (20 for SDH)
Speaking rate (per project)	134–197 wpm	BBC target 160–180 wpm



Figure 6: Captions over on-screen text. Frames from the channel's video of 7 October 2026 at 0:05.7 and 0:20.9; the article screenshots (NTV News NNN, 22 September 2026; Yomiuri TV via Yahoo! News Japan, 1 October 2026) appear with the video's own source credits.

6 Study 2: a minimum page time in the pipeline

The pipeline burns captions into the video from the corrected word timings and writes the subtitle track separately, as two lines of up to 42 characters and 6 s, close to Netflix’s numbers. Following Study 1, it gained phrase-aware breaks and a minimum page time, on by default. No page ends on a word that leads into the next (Indonesian function words plus *sama*, *buat*, *jadi*), between a number and what it counts, or inside a name. A page shown for less than 0.8 s joins a neighbour when the two still fit the look’s line, never across a sentence’s end or a pause over 1 s. Only page boundaries move; the subtitle track never sees it.

The result was not what the design intended (Figure 7, Table 4).

The looks first split the recognised pages further: at 1080 pixels wide, with type about 6–8% of the frame’s height and the spoken word drawn bigger still, a line holds little more than one word. That split shortens pages more than the joining rule can lengthen them, because a short page can only join a neighbour that still fits on the same line, and it rarely does. The rule joined about 5% of pages; the share under the minimum fell by one point.

The simulation shows where the room is. Giving a joined page a second line (BBC already allows three lines on 9:16) brings the median page to 1.16 s and the share under 5/6 s to 17%.

Answer to RQ2. Not on its own. A minimum page time only works where the layout leaves space to merge: in a large one-line look it changes almost nothing, and a second line for joined pages, or a leaner look of about 5–6% of the frame height for dense stretches, is the change to test next.

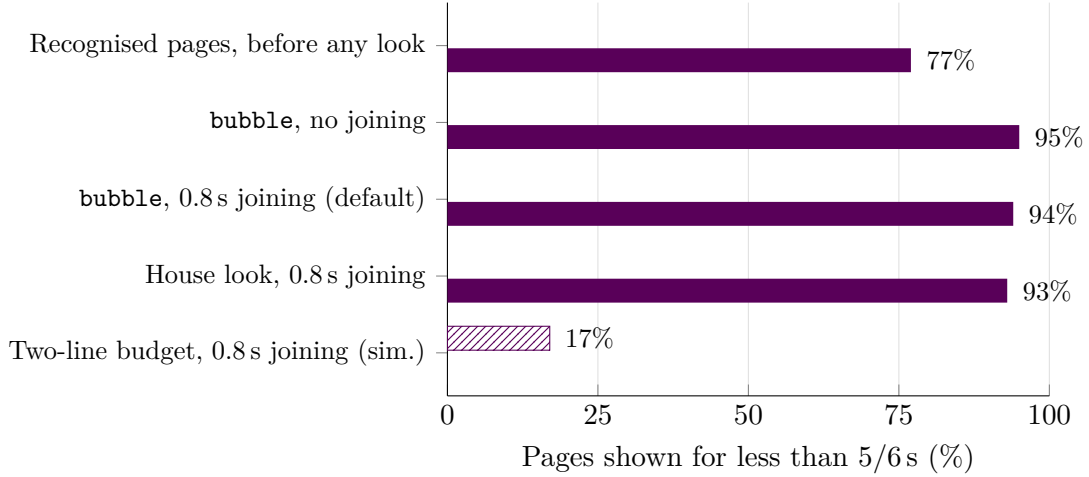


Figure 7: Share of caption pages shown for less than Netflix’s 5/6-s minimum, by condition, on 12 projects at 1080×1920. The hatched bar is a simulation (32 characters and at most 6 words a page, no font metrics). Table 4 has every number.

Table 4: Caption pages by condition (12 projects, 1080×1920).

Condition	Pages	Median	<5/6 s	<0.5 s	Words a page
Recognised pages, before any look	594	0.62 s	77%	24%	1.94
bubble , no joining	948	0.38 s	95%	69%	1.22
bubble , 0.8 s joining (default)	895	0.40 s	94%	65%	1.29
House look, no joining	907	0.40 s	94%	65%	1.27
House look, 0.8 s joining	849	0.43 s	93%	60%	1.36
Two-line budget, 0.8 s joining (simulated)	338	1.16 s	17%	1%	3.41

7 A two-layer design

The studies show what the pages need (time, and room to merge); the background shows how professional editors decide what to draw. Together they give a design in two layers.

Layer 1, the record. The exported subtitle track (SRT and VTT) carries every word, in order, verbatim. It is what a deaf viewer switches on, what search and automatic translation read, and what the channel’s translations are built from. FCC, BBC and WCAG all require it [STANDARD], and any swap or step-aside that touched it would lose words in all three places.

Layer 2, the page. The burned-in page is a design layer over the record. It may make one of four moves (Figure 8), each with limits. Designed captions also take control away from the viewer, who cannot switch them off (Zdenek, 2018) [OPINION], which is one more reason to keep the moves few.

7.1 (a) Step aside

The caption hides while the same words are on screen. *For:* Netflix deletes forced narratives that duplicate on-screen text, and every standard forbids covering featured text [STANDARD]; identical on-screen text and narration impaired learning (Yue et al., 2013) [STUDY]; Japanese studio talk often carries no telop at all [PRACTICE]. *Against:* on this channel the boards are mostly Japanese or English under Indonesian narration (Figure 6), so they rarely carry the spoken words; a phone-sized board may be less legible than the caption, which BBC 2.10 forbids [STANDARD]; and a muted viewer must find the words elsewhere on screen. *Limits:* only when



Figure 8: The design, one moment five ways, on the channel’s opening line from 7 October 2026 (“Toko buku bekas di Jepang lagi diborong sama pembeli misterius”, *Japan’s used bookshops are being bought out by a mysterious buyer*) over an invented headline card: today; (a) step aside; (b) emoji; (c) picture; (d) style. The last panel is the subtitle track every version exports, unchanged. The performer is a grey placeholder. Mock-up.

the board’s highlighted span equals the spoken words (same language, same order), with the highlighter sweeping them in sync, at least as large as the caption, a small marker left in the caption band, and the caption back as soon as a word is spoken that is not on the board.

7.2 (b) Emoji

For: frequent emoji are read almost like words, and an emoji after its word is read faster than the word (Paggio and Tse, 2022; Lu et al., 2025) [STUDY]; huazi use stickers to add feelings [PRACTICE]. *Against:* a replacing emoji costs about 300–350 ms, ambiguous ones more (Scheffler et al., 2021, via Ruhr-Universität Bochum 2021; Paggio and Tse 2022) [STUDY]; deaf and hard-of-hearing users in two small studies found emoji captions ambiguous or unwanted (Oomori, 2022; Zhang et al., 2025) [STUDY]. *Limits:* prefer the emoji beside the word; replace only a concrete, common noun with one obvious emoji, never a verb, name, number or a word whose sound carries the joke; at most one per sentence, and never on a page that already has a keyword.

Table 5: Verdict on the four moves.

Move	Verdict	Use when
(a) Step aside	Yes, narrowly	the board shows the spoken words, in the spoken language, readable, highlighted in sync
(b) Emoji	Beside the word, yes; instead of it, rarely	a concrete common noun, one per sentence
(c) Picture	Rarely	the story’s main object, after its first mention, one per sentence
(d) Style	Yes	a small fixed set of voice and emotion styles, used consistently
Verbatim export	Always	every version

7.3 (c) Picture

For: Japanese telop does it (Higashino, n.d.) [PRACTICE]; labels and pictures together help learning (Mayer and Johnson, 2008, second-hand via Yue et al. 2013) [STUDY]; a picture of the story’s own object says what a word cannot. *Against:* everything that slows emoji applies more strongly to an unfamiliar picture, and a sticker the size of a word is small on a phone. *Limits:* only the story’s main object or person, and only after its word has been heard and seen once, at least 1.3–1.5 times the caption’s cap height, one per sentence, never stacked with an emoji.

7.4 (d) Style

The best supported of the four. BBC gives each speaker a colour for the whole programme [STANDARD]; colour with weight or size helped deaf and hard-of-hearing viewers recognise emotion (de Lacerda Patata et al., 2024) [STUDY]; Japanese stock colours and Korean voice styles do it daily [PRACTICE]. Unexplained colour confuses (Zhang et al., 2025) [STUDY] and stereotyped mappings draw criticism (Morikawa, 2023), so the vocabulary stays small and fixed: the presenter’s own voice, one colour per quoted person kept across videos, an aside or inner voice, two or three emotions, and a sound-word style, at most one change per sentence.

Implementation status. Besides the minimum page time of Study 2, the pipeline has a sound-word lane (a sound effect’s cue may carry text, drawn telop-style beside its source and outside the caption band) and an experimental display layer that can step a page aside or swap a word for an emoji, in the burned-in page only. Voice and mood styles, emoji beside the word drawn with Google’s Noto Color Emoji, and step-aside checked against a board’s recognised words are planned; picture swap is deferred.

Answer to RQ3. Repeat by default, as the verbatim record; add in a separate lane (sound words, asides, labels); replace rarely and only under the limits above; and step aside only when the screen shows the very words spoken.

8 Guidelines and the strength of their evidence

Table 6 lists the guidelines with their grades (Section 4.4).

Table 6: Guidelines for burned-in captions on 9:16 explainers. Grades: A, standards and at least one study agree; B, one or two studies or consistent documented practice; C, practice or opinion only.

#	Guideline	Grade	Main sources
1	The exported subtitle track stays verbatim, whatever the page does.	A	FCC; WCAG 1.2.2; BBC 2.1–2.2; Gernsbacher (2015)
2	Page by phrase: never end a page on a preposition, conjunction or the relative <i>yang</i> ; never split a name or a number from its noun.	A	BBC 3.4; Netflix; DCMP; Gerber-Morón et al. (2018)
3	Give each page time to be read: about 0.8 s at least, by merging neighbours, which needs room on the page (guideline 7).	A	Netflix; DCMP; BBC; Szarkowska et al. (2024); Study 2
4	Keep captions off faces, mouths and on-screen text; on 9:16, lower third but above the player’s controls.	A	FCC; BBC 10.1; Netflix; DCMP; Brown et al. (2015)
5	Outline or box all text: at least 3:1 contrast for large text, 4.5:1 for small.	A	WCAG 1.4.3
6	Show the page whole and move the highlight, rather than building it word by word.	B	Rajendran et al. (2013) (live subtitles; test it)
7	Big type needs fewer words a line: pair large one-line looks with a second line or a leaner look for dense stretches.	B	BBC 9.2.1; Study 2
8	One keyword per sentence at most: the number, the name or the word the joke hangs on.	B	BBC 11.2; Finger Bou and Muñoz Lahoz (2024); Morikawa (2023)
9	Colour means one thing, always: one colour per voice, two to four fixed emotion styles, encoded with colour plus weight or size.	B	BBC; de Lacerda Pataca et al. (2024); Morikawa (2023)
10	Sound words are a separate lane near their source, never in the caption band and never replacing speech.	B	WCAG 1.2.2; BBC 17; Sasamoto (2021)
11	Emoji beside the word, not instead of it; a replacing emoji only for a concrete common noun.	B	Scheffler et al. (2021); Paggio and Tse (2022); Lu et al. (2025)
12	One moving thing at a time; motion follows the voice; entrances of about 0.1–0.25 s.	C	Sasamoto and O’Hagan (2020) (preliminary); opinion
13	Step aside only for the exact spoken words, in sync.	C	Netflix; Mirzaei et al. (2014) (learners)
14	A picture stands in for a word only for the story’s main object, after its first mention.	C	Higashino (n.d.); Zeng (2017)
15	Test style claims on the channel itself; most “this caption style doubles retention” claims are vendor opinion.	C	opinion

9 Future evaluation

Most short-form caption claims are vendor opinion [OPINION], so the next step is a test on the channel itself, with YouTube Studio’s per-video retention as a first read:

1. **Pagination:** the same Short with today’s pages against pages of at least 0.8s on two lines. Measure average view duration, share viewed and rewatches.
2. **Build against page:** bubble building word by word against whole pages with a moving highlight.
3. **Designed exceptions:** plain captions against captions plus a sound-word lane and one style change every ten seconds or so.
4. **A viewer check:** a few deaf or hard-of-hearing Indonesian viewers, and someone watching muted, watch each version and say what they missed.

Variants go out a week apart on similar topics, or as a Shorts A/B test where the platform offers one.

10 Limitations

- **One channel.** The measurement covers 11–12 projects of one creator, one language and one format; the numbers describe this channel, not short-form video in general.
- **Durations, not outcomes.** The studies measure how long pages stay on screen, not what viewers understood or how long they watched. Section 9 is the plan for that.
- **Small and indirect studies.** Several studies in the background are small (4, 11, 17 or 39 participants), the word-for-word finding comes from live subtitles, the emoji studies used running text in English and Chinese rather than captions, and none tested Indonesian.
- **Second-hand and missing sources.** Some claims are taken from the papers that cite them and are marked so; no credible source on VTuber clip-channel telop could be opened.
- **The simulation** counts characters and ignores the look’s font, size and the grown spoken word, so it overstates what a real two-line layout would hold.
- **The author designs the tool and runs the channel.** The defaults evaluated here are the author’s own, and the review and measurement were carried out with an AI assistant (see the acknowledgements). The data and code paths can be re-run.

11 Conclusion

The channel’s caption pages are too short to read, and the reason is pagination, not speech. The obvious fix, a minimum page time, does almost nothing in a large one-line layout and works once a page has room for a second line: the oldest rule in subtitling, give each page time to be read, needs space as well as time. Beyond pagination, the evidence supports a verbatim record that never changes and a designed page on top of it, which borrows from variety television its separation of layers and its small, consistent vocabulary: one colour per voice, a few emotion styles, a lane for sound words. Of the four moves on the page, a style change is well supported, an emoji beside the word is safe, stepping aside fits only when the screen shows the very words spoken, and a picture in place of a word should be rare. Not every word needs to be drawn, but every word needs to be kept.

Acknowledgements

The literature review, the measurements and the draft were prepared with Claude (Anthropic) under the author’s direction, using the author’s editing pipeline, Poiesis. The author is responsible for the content.

Figure credits

Figures 5 and 7 and Tables 3 and 4: the channel’s own caption data. Figure 6: frames from the channel’s video of 7 October 2026; the article screenshots in them are shown with the video’s own source credits. Figures 1, 2 and 8: mock-ups made with the channel’s own words; the sticker in Figure 8 is the project’s own and the headline card is invented. Figures 3 and 4: drawn from the documented rules and devices; no broadcast frame is reproduced.

References

- Banpie Sishu. Qipa Shuo zongyi jianji de san zhong huazi taolu [Three patterns of decorated captions in the editing of Qipa Shuo]. Toutiao, <https://www.toutiao.com/zixun/7531232531161843738>, 2017. 14 October 2017; in Chinese.
- BBC. Subtitle Guidelines, version 1.2.5. <https://www.bbc.co.uk/accessibility/forproducts/guides/subtitles/>, 2026. March 2026.
- A. Brown, R. Jones, M. Crabb, J. Sandford, M. Brooks, M. Armstrong, and C. Jay. Dynamic Subtitles: The User Experience. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video (TVX)*, 2015. URL <https://research.manchester.ac.uk/en/publications/dynamic-subtitles-the-user-experience/>.
- Daily Sports. “Getsuyō kara Yofukashi” netsuzō mondai: Nittere shachō ga shazai [The “Getsuyō kara Yofukashi” fabrication: Nippon TV’s president apologises]. <https://www.daily.co.jp/gossip/2025/03/31/0018816522.shtml>, 2025. 31 March 2025; in Japanese.
- C. de Lacerda Pataca, S. Hassan, N. Tinker, R. L. Peiris, and M. Huenerfauth. Caption Royale: Exploring the design space of affective captions from the perspective of deaf and hard-of-hearing individuals. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024.
- Caluã de Lacerda Pataca and Paula Dornhofer Paro Costa. Hidden bawls, whispers, and yelps: Can text be made to sound more than just its words? *IEEE Transactions on Affective Computing*, 2023. arXiv:2202.10631.
- Described and Captioned Media Program. Captioning Key. <https://dcmp.org/learn/captioningkey>, 2026.
- Dublin City University. Research on Japanese impact captions which shake, pulse and capture your attention on TV screen. <https://www.dcu.ie/research/news/2016/jun/research-japanese-impact-captions-which-shake-pulse-and-capture-your>, 2016. 9 June 2016.
- R. Finger Bou and C. Muñoz Lahoz. The effects of textual enhancement on young learners’ attention and vocabulary acquisition through captioned cartoons. <https://diposit.ub.edu/items/5f957ae3-2963-4715-b177-e6d7d7c845c3/full>, 2024.
- O. Gerber-Morón, A. Szarkowska, and B. Woll. The impact of text segmentation on subtitle reading. *Journal of Eye Movement Research*, 11(4), 2018. URL <https://bop.unibe.ch/JEMR/article/view/4267>.

- M. A. Gernsbacher. Video captions benefit everyone. *Policy Insights from the Behavioral and Brain Sciences*, 2(1):195–202, 2015.
- Google Ads Help. Your guide to YouTube Shorts ads. <https://support.google.com/google-ads/answer/16041697>, 2026.
- Yuka Higashino. Memes in Japanese language learning: Unintended consequences of the modern evolution of telop. In *Proceedings of the Central Association of Teachers of Japanese (CATJ)*, volume 30, n.d. URL https://www.cla.purdue.edu/academic/slc/1/japanese/documents/catj30/yuka_higashino.pdf.
- J.-L. Kruger, N. Wisniewska, and S. Liao. Why subtitle speed matters: Evidence from word skipping and rereading. *Applied Psycholinguistics*, 43(1):211–236, 2022.
- W. Lu, H. Du, F. Gu, and J. Han. The lexicalization of emojis. *Frontiers in Psychology*, 2025. URL <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2025.1631967/full>.
- M. S. Mirzaei, Y. Akita, and T. Kawahara. Partial and synchronized caption generation to develop second language listening skill. In *Proceedings of the International Conference on Computers in Education (ICCE)*, 2014. URL <https://library.apsce.net/index.php/ICCE/article/download/3065/2941/4029>.
- Toshio Morikawa. “Moji darake no gamen” wa nani o utsushite iru ka: Chijōha terebi bangumi ni okeru komento forō teroppu no genjō bunseki [What does a “screen full of text” show? An analysis of comment-follow telops in terrestrial television programmes]. *Edogawa Daigaku Kiyō [Bulletin of Edogawa University]*, 33:267–280, 2023. URL <https://edo.repo.nii.ac.jp/record/1223/files/DK2023-16.pdf>. In Japanese.
- Naga Nakano. Baraeti bangumi ni okeru teroppu no yakuwari ni henka: “Jōhō o oginau” kara “tsukkomi” ni [The changing role of telop in variety programmes: from “supplementing information” to “tsukkomi”]. ORICON NEWS, <https://www.oricon.co.jp/special/51622/>, 2018. 19 August 2018; in Japanese.
- Qidao Nan and Yi Chen. Weishenme shipin huazi ruci liuxing? [Why are decorated video captions so popular?]. 36Kr, <https://www.36kr.com/p/1721681674241>, 2017. 11 July 2017; in Chinese.
- Netflix. Timed Text Style Guide: General Requirements. <https://partnerhelp.netflixstudios.com/hc/en-us/articles/215758617>, 2022. Updated 7 October 2022.
- Netflix. Indonesian Timed Text Style Guide. <https://partnerhelp.netflixstudios.com/hc/en-us/articles/216009727>, 2025a. Updated 17 October 2025.
- Netflix. Japanese Timed Text Style Guide. <https://partnerhelp.netflixstudios.com/hc/en-us/articles/215767517-Japanese-Timed-Text-Style-Guide-v4-0>, 2025b. Updated 19 December 2025.
- Hye-Gyeong Ohe. Geinō baraeti bangumi ni okeru moji teroppu no Nikkan hikaku [A Japan–Korea comparison of text telops in entertainment variety programmes]. *Ōsaka Daigaku Gengo Bunkagaku [Osaka University Studies in Language and Culture]*, 30:37–53, 2021. In Japanese.
- Hye-Gyeong Ohe. Nikkan no geinō bangumi ni okeru dōbutsu no gijinka [The personification of animals in Japanese and Korean entertainment programmes]. *Gogaku Kenkyū [Language Research Bulletin, ICU]*, 37:107–119, 2022. In Japanese.

- Kotaro Oomori. A preliminary study on understanding online meetings using emoji-based captioning for deaf or hard of hearing users. https://digitalnature.slis.tsukuba.ac.jp/wp-content/uploads/2022/03/Kotaro-Oomori_Thesis_Abstract.pdf, 2022. Thesis abstract, University of Tsukuba.
- P. Paggio and A. P. P. Tse. Are emoji processed like words? An eye-tracking study. *Cognitive Science*, 46:e13099, 2022.
- Sang-wan Park. TV yeneung peurogeuraem-ui yeongeukjeok gihoek jeollyak-e gwanhan yeongu [A study of the theatrical planning strategies of TV entertainment programmes]. *Daejung Seosa Yeongu [Journal of Popular Narrative]*, 21(3), 2015. In Korean.
- D. J. Rajendran, A. T. Duchowski, P. Orero, J. Martínez, and P. Romero-Fresco. Effects of text chunking on subtitling: A quantitative and qualitative examination. *Perspectives*, 21(1): 5–21, 2013.
- Ruhr-Universität Bochum. Emojis can replace words in text. Does its comprehensibility suffer as a result? <https://news.rub.de/english/press-releases/2021-11-11-linguistics-how-we-understand-emoji>, 2021. On Scheffler et al., Computers in Human Behavior, 2021.
- R. Sasamoto. Onomatopoeia, impressions and text on screen. Preprint, Dublin City University, <https://doras.dcu.ie/26625/>, 2021.
- R. Sasamoto and M. O’Hagan. Relevance, style and multimodality: Typographical features as stylistic devices. Preprint, Dublin City University, <https://doras.dcu.ie/26624/>, 2020.
- Shūkan Josei PRIME. Article on the fabricated street interview in “Getsuyō kara Yofukashi”. <https://www.jprime.jp/articles/-/36130>, 2025. 29 March 2025; in Japanese.
- StreamingMedia. 80% of Video Caption Users Aren’t Hearing Impaired, Finds Verizon. <https://www.streamingmedia.com/Articles/News/Online-Video-News/80-of-Video-Caption-Users-Arent-Hearing-Impaired-Finds-Verizon-131860.aspx>, 2019. 20 May 2019; on a Verizon Media and Publicis Media survey.
- A. Szarkowska and O. Gerber-Morón. Viewers can keep up with fast subtitles: Evidence from eye movements. *PLoS ONE*, 13(6):e0199331, 2018.
- A. Szarkowska, V. Ragni, S. Szkriba, S. Black, D. Orrego-Carmona, and J.-L. Kruger. Watching subtitled videos with the sound off affects viewers’ comprehension, cognitive load, immersion, enjoyment, and gaze patterns. *PLoS ONE*, 19(10):e0306251, 2024.
- Tenasia. Reonningmaen jejakjin gongsik sagwa [The Running Man production team’s official apology]. <https://www.tenasia.co.kr/article/2016122523384>, 2016. 25 December 2016; in Korean.
- TikTok Ads Help. TikTok Reservation TopView. <https://ads.tiktok.com/help/article/tiktok-reservation-topview>, 2026.
- Twig24 (Seoul Shinmun). Reonningmaen jamak nollan [The Running Man caption controversy]. <https://twig24.com/news/entertainments/broadcasting/2019/06/03/20190603500072>, 2019. 3 June 2019; in Korean.
- US Code of Federal Regulations. 47 CFR § 79.1(j)(2): Caption quality standards. <https://www.law.cornell.edu/cfr/text/47/79.1>, 2026.

- W3C. Understanding SC 1.2.2: Captions (Prerecorded), WCAG 2.2. <https://www.w3.org/WAI/WCAG22/Understanding/captions-prerecorded.html>, 2023a.
- W3C. Understanding SC 1.4.3: Contrast (Minimum), WCAG 2.2. <https://www.w3.org/WAI/WCAG22/Understanding/contrast-minimum.html>, 2023b.
- C. L. Yue, E. L. Bjork, and R. A. Bjork. Reducing verbal redundancy in multimedia learning: An undesired desirable difficulty? *Journal of Educational Psychology*, 2013.
- S. Zdenek. Designing captions: Disruptive experiments with typography, color, icons, and effects. *Kairos: A Journal of Rhetoric, Technology, and Pedagogy*, 23(1), 2018. URL <https://kairos.technorhetoric.net/23.1/topoi/zdenek/introduction.html>.
- G. Zeng. The communication value of multi-style subtitles. In *Proceedings of the International Conference on Education, Sports, Arts and Management Engineering (ICESAME)*. Atlantis Press, 2017. URL <https://www.atlantis-press.com/article/25877328.pdf>.
- Yu Zhang, Yi Wen, Siying Hu, and Zhicong Lu. SpeechCap: Leveraging playful impact captions to facilitate interpersonal communication in social virtual reality. arXiv:2502.10736, 2025.