

Subtitle Studio (automatic video subtitles and dubbing, v1.0.0) — Eight-Clip, Four-Engine Dubbing Comparison

Test date: 2026-10-15 (UTC+8) · Environment: <https://8-219-8-135.sslip.io> · Account: admin (administrator) · Tester: QA

Scope: 8 clips, each into 6 target languages, comparing 4 dubbing engines (Fish Audio, VoxCPM2, ElevenLabs v1, ElevenLabs v2).

A real upload with real charges; all evidence downloaded locally. **Video restyle (Group C) was excluded by agreement.**

Testing was paused at 18:35 because the ElevenLabs credit ran short; the v2 round reached only 17%.

1. Summary

Verdict: subtitles and translation are usable as a first draft; dubbing fails on loudness matching and background retention across all four engines.

Three engines were completed in full (Fish Audio, VoxCPM2 and ElevenLabs v1, 48 dubs each); ElevenLabs v2 reached only 8/48 (17%).

Subtitles performed well: language auto-detection 8/8 correct, SRT/VTT identical 56/56, zero violations of line length, line count and cue duration, burned-in resolution and duration matching the source 8/8, dub audio track order correct 12/12, no phantom subtitles, no missed lines, and all 5 input-validation checks passed.

Dubbing problems are concentrated and systemic: all four engines run hot on speech loudness (+1.82 to +6.84 dB against a ± 1 dB target) and all four fail background retention (0.534–0.766 against ≥ 0.85).

The day also saw 2 service-level outages and 2 ElevenLabs credit exhaustions — see Section 9.

Issues	High	Medium	Low	Cases	Run	Passed	Failed	Not run	Pass rate (of run)
20	5	10	5	109	78	46	32	31	59% (46/78)

The four things to fix first:

- Q15 (High) Third-party calls block the whole application.** Two service-level outages today (16:03–16:39, 17:20–18:13), with 502 or no response on every path. The dose relationship is clear: EL v1 took about 30 minutes to trigger it; **EL v2 took 7 minutes and had completed only one dub.**
- Q01 (High) Switching engines on an existing job charges, reports success, and delivers the previous engine's audio.** Reproduced precisely in three steps with credit availability ruled out; the only working path is to create a new job with the engine set at creation.
- Q02 + Q03 (High) Loudness and background fail on all four engines.** The loudness error scales with how quiet the source is (quiet material pushed up by as much as +26.3 dB); even the best background score, EL v2's 0.766, misses the target. **This proves the problem lies in the platform's mixing stage and cannot be fixed by changing engine.**
- Q18 (High) The previous report's S03 is still open.** Spanish "No hablo español" (I don't speak Spanish) is still recognised as the imperative "No hable español", and **all six languages reverse the meaning** — even though the very next line, `iEstoy hablando español!`, contradicts it.

2. Scope

Area	Details
Tested	Speech recognition (language detection, sampled transcripts, phantom subtitles, completeness); translation into 6 languages (meaning, proper nouns, register, cross-language consistency); subtitle layout (line length, line count, duration, reading speed, line breaks, punctuation); SRT/VTT consistency; burned-in MP4 (resolution, duration); dubbing (duration, loudness, clipping, background, speech coverage, track order, switchable subtitles); four-engine comparison; failure handling; input validation; permissions and CSRF; Gallery; cost and three-way reconciliation
Not tested, and why	① ElevenLabs v2's remaining 40 dubs : the account ran out of credit or hit its spending limit (twice today); needs a top-up and coordination with the other user. ② Video restyle, Group C, 39 cases : excluded by agreement; needs separate budget (~\$28.21). ③ Lip-sync, 4 cases : no lipsync engine is configured live. ④ Subjective dub assessment, 14 cases : needs native-speaker listening; only acoustic measurement was done. ⑤ OCR cross-check of on-screen text : no OCR tool; frame grabs can only take the first frame, not a given timestamp. ⑥ Numbers/dates/idioms, 3 cases : the 8 clips contain no such content.
Test data	8 clips, 130.1s total: C01 (EN 10.0s), C05 (ZH 10.0s), C06 (ZH 10.0s), joker (EN 16.0s), ch8 (TH 17.0s), dgd (VI 22.0s), craig (RU 26.0s), Spanish-Testing (ES 19.0s, replacing the earlier clip) . Each into 6 target languages, SRT+VTT, burn-in on, burned-subtitle removal auto, dub subtitles switchable. Recognition: ElevenLabs Scribe v2; translation: GPT-6 Luna; dubbing: the four engines under comparison
Method	Uploaded and ordered through the live API via a local proxy; downloaded everything (903 files / 999 MB); decoded audio to 16 kHz mono PCM with macOS CoreAudio (afconvert) plus a hand-written WAV parser, measuring RMS and peak every 0.1s and separating "speech" from "outside dialogue" by the source-language SRT windows; clipping defined as ≥ 3 consecutive full-scale samples; parsed MP4 atoms in Python to verify audio and subtitle track order; measured subtitles by script and compared every SRT with its VTT
Account	admin (administrator), through the local proxy at http://localhost:5173

3. Core flow results

#	Flow	Result	Notes / where it breaks
1	Upload → recognise → translate → burn → download	Pass	All 8 clips produced 7 SRT + 7 VTT + 7 burned MP4 + 7 thumbnails; 0 failures
2	Automatic language detection	Pass	8/8 correct, including spotting a clip mislabelled as Spanish that is actually Hungarian
3	Speech recognition	Partial pass	No phantom subtitles, no missed lines; but C05's sung line looks truncated and the Spanish key line is in the wrong person (Q18)
4	Translation (6 languages)	Partial pass	Broadly faithful with appropriate register; but the S03 key line is reversed, speakers are merged, proper nouns are inconsistent and cue splitting differs by language
5	Subtitle layout	Partial pass	Zero violations of length, lines and duration; 157/320 cues exceed reading speed
6	SRT / VTT consistency	Pass	56/56 identical
7	Burned-in MP4	Pass	Resolution 8/8 and duration 8/8 match the source
8	Dub generation	Partial pass	Fish / VoxCPM2 / EL v1 all 48/48; EL v2 only 8/48 (out of credit)
9	Dub audio quality	Fail	Loudness (Q02) and background (Q03) fail on all four engines; only the ElevenLabs engines pass on clipping
10	Dubbed MP4 structure	Pass	12/12 have the dub track first and the original second; subtitle tracks include the target language plus the other six, switchable
11	Switching engine via redo	Fail	Switching to ElevenLabs charges and reports ready, but the tracks do not update (Q01)
12	Service availability	Fail	Two full outages today, both triggered by ElevenLabs load (Q15)
13	Input validation and permissions	Pass	All 5 validation checks pass; 401 without login, 403 without a CSRF token
14	Gallery	Pass	Create / add / remove / delete folders all work; multi-select ZIP download verified by its PK magic bytes
15	Cost and reconciliation	Partial pass	Per-item charges match the price list exactly; but the cost field lags and third-party balance cannot be queried in advance

4. Four-engine dubbing comparison (headline result)

Metric	Target	Fish Audio	VoxCPM2	ElevenLabs v1	ElevenLabs v2*
Samples	—	48	48	48	8
Mean speech loudness difference	±1 dB	+6.84 dB	+6.69 dB	+3.60 dB	+1.82 dB
Outside ±1 dB	0	43/48	44/48	34/48	6/8
Mean background correlation	≥0.85	0.591	0.561	0.534	0.766
Below 0.85	0	41/48	46/48	38/48	5/8
Max duration difference	≤0.1 s	0.066 s	0.066 s	0.016 s	0.0 s
Mean speech coverage	≥0.80	0.983	0.993	0.917	0.966
Clipped runs (≥3 full-scale)	0	27/48	20/48	0/48	0/8
Dubbing API cost	—	\$0	\$0	\$6.52	\$2.96 (partial)

* ElevenLabs v2 has only 8 samples, **all from just 2 clips (C01 and C05)**. It covers no multi-speaker scene, no strong-emotion scene and none of the minority source languages, so **it cannot be treated as a conclusion**.

Four key conclusions

- 1. **Quality order: ElevenLabs v2 > v1 > Fish ≈ VoxCPM2.** v2 has the smallest loudness error (+1.82 dB), the highest background retention (0.766), exact duration and no clipping — but on a narrow sample that needs confirming.
- 2. **Clipping is specific to the cheaper engines.** Both ElevenLabs engines clip nowhere (0/48, 0/8), while Fish clips in 27/48 and VoxCPM2 in 20/48, concentrated on the joker clip (15–19 clipped runs each, longest 39 samples).
- 3. **Background retention fails on all four engines (0.534–0.766).** Even the best, v2, reaches only 0.766. **This proves background loss is a systemic problem in the platform's mixing stage that no change of engine can fix** — the most valuable conclusion of this round.
- 4. **Loudness follows a clear pattern:** the error scales with how quiet the source is. On C01, whose level is normal, all four engines pass (–1.2 to +0.1 dB); on the quiet craig clip they run +18.6 to +26.3 dB hot. The platform normalises dubs to a fixed target loudness instead of matching the source.

Per clip (loudness difference in dB / background correlation)

Clip	Fish	VoxCPM2	ElevenLabs v1	ElevenLabs v2
C01 wide two-person · Dutch accent · BGM	–0.2 / 0.69	–0.6 / 0.64	–1.2 / 0.33	+0.1 / 0.80
C05 close-up singing · on-screen lyrics	+3.5 / 0.26	+2.5 / 0.27	+2.5 / 1.00	+4.7 / 0.71
C06 four-person dialogue	+9.3 / 0.27	+8.3 / 0.23	+6.4 / 0.23	—
joker strong emotion	+1.4 / 0.81	+1.3 / 0.81	–0.0 / 0.74	—
ch8 Thai source	+1.5 / 0.90	+1.2 / 0.80	+0.2 / 0.78	—
dgd Vietnamese source	–2.2 / 0.48	–2.0 / 0.42	–4.6 / 0.20	—
craig Russian source	+23.7 / 0.50	+26.3 / 0.51	+18.6 / 0.15	—
Spanish–Testing (Spanish source)	—	—	+6.9 / 0.85	—
BestMovie (retired — actually Hungarian)	+17.6 / 0.81	+16.5 / 0.81	—	—

"—" means the engine did not cover that clip. `Best Movie Clips Spanish` was confirmed by transcription to be Hungarian and was replaced the same day with `Spanish–Testing` video ; only the Fish and VoxCPM2 rounds contain the old clip.

5. Subtitle and translation quality

5.1 Measured results (56 SRTs, 320 cues)

Check	Target	Measured	Verdict
Characters per line	≤42 Latin/Cyrillic, ≤16 Chinese, ≤35 Thai	0 lines over	Pass
Lines per cue	≤2	0 cues over	Pass
Cue duration	0.8–7 s	0 out of range	Pass
Reading speed	≤17 chars/s Latin, ≤9 Chinese	157/320 cues over	Fail
SRT / VTT consistency	identical	56/56 identical	Pass
Chinese Simplified	all Simplified	8/8 files clean	Pass
Thai punctuation	one rule per language	4/8 use ? and !, 4/8 do not	Fail
Spanish ¿ ;	correctly paired	correctly paired	Pass
Phantom subtitles over non-speech	0	0 segments across 8 clips	Pass
Completeness (speech with no cue)	0	0 segments across 8 clips	Pass

Reading speed by language: Indonesian 73.9%, Spanish 74.4%, English 67.4%, Vietnamese 56.5%, Russian 55.6%, Chinese 20.5%, Thai 0%. Fastest 34.5 chars/s against a limit of 17.

5.2 Q18: the previous report's S03 is still open (key regression)

The replacement Spanish clip is the same Bedazzled scene as in the earlier report. Its first cue is still recognised as `No hable español` (imperative) instead of `No hablo español` ("I don't speak Spanish"). **All six languages reverse the meaning:**

Language	Current translation	Actual meaning
EN	Don't speak Spanish.	Don't speak Spanish ❌
ZH	别说西班牙语。	❌
TH	อย่าพูดภาษาสเปน	❌
VI	Đừng nói tiếng Tây Ban Nha.	❌
ID	Jangan bicara bahasa Spanyol.	❌
RU	Не говорите по-испански.	❌

The next line, `¡Estoy hablando español!` ("I AM speaking Spanish!"), is a complete contextual refutation — the first speaker must have said "I don't speak Spanish" for that reply to make sense. The system neither used the context nor flagged the suspicious grammar.

5.3 Translation spot check (C06, four-person dialogue, 6 languages)

All six translations are faithful and fluent, with appropriate register and forms of address (Thai ผม, Vietnamese anh/em, Indonesian Kak all correct). Two problems:

- **Speakers merged:** cues 3 and 4 each contain two speakers (a question and its answer) in one cue.
- **Inconsistent proper noun:** 大门墩 → EN/ES/ID keep the pinyin "Damentun"; TH ต้าเหมิน墩, VI Đại Môn Đôn, RU Д а м э н ь д у н ь transliterate.

5.4 Cue counts differ by language

The new Spanish clip has 3 source cues; en/th/vi/ru split them into **4**, while zh/id keep **3**. The same content is split differently per language.

6. Issue list (highest severity first)

ID	Area	Sev.	Description	Steps	Expected vs actual	Suggested fix
Q15	Availability	High	Third-party (ElevenLabs) calls block the whole application; two full outages today	Keep ordering ElevenLabs dubs and request any path on the main site	Expected: the app stays available. Actual: 502 on every path 16:03–16:39; zero bytes in 60s 17:20–18:13, while host, nginx and the sibling sites were all healthy	Add timeouts and a circuit breaker to every third-party call; cap the worker's <code>running</code> counter; make third-party failures fail fast
Q01	Dubbing / billing	High	Switching engines on an existing job charges, reports ready, and delivers the previous engine's audio	POST <code>/jobs/{id}/dubs</code> with <code>engine_id=8</code> or <code>7</code> and <code>redo=true</code>	Expected: deliver the new engine's audio or do not charge. Actual: charged \$0.0833 (exactly right), <code>dub.status=ready</code> , no error, track still VoxCPM2	Do not mark a failed re-run as ready; make charging and delivery atomic; or fix the write-back on the ElevenLabs path
Q02	Dubbing	High	Dub loudness does not match the source; the error scales with how quiet the source is	Decode dub and source, compare RMS inside the source SRT windows	Expected: ± 1 dB. Actual: +1.82 to +6.84 dB across engines; craig reaches +26.3 dB, while the normally-levelled C01 passes on all four	Match loudness to the source after mixing (EBU R128) rather than normalising to a fixed target
Q03	Dubbing	High	Background retention fails on all four engines; musical material loses almost all of it	Correlate loudness outside the subtitle windows	Expected: ≥ 0.85 . Actual: Fish 0.591, VoxCPM2 0.561, EL v1 0.534, EL v2 0.766	Mix with the separated original background stem; state on the order page that the engine removes the background
Q18	Recognition / translation	High	Spanish "No hablo español" recognised as the imperative "No hable español"; all six languages reverse the meaning (previous report's S03, still open)	Compare the new Spanish clip's first cue with all six translations	Expected: "I don't speak Spanish". Actual: 6/6 render it as "Don't speak Spanish", although the next line refutes it	Pass context to the translator; flag suspicious source grammar
Q16	Observability	Medium	The platform exposes no third-party balance, quota or concurrency state	Read <code>/admin/providers</code> and <code>/admin/billing</code>	Expected: balance or remaining quota is queryable. Actual: cumulative spend only; neither credit exhaustion today could be predicted	Show each provider's balance and concurrency in admin; pre-check quota before ordering
Q17	Job status	Medium	A queued job does not say it is waiting on a third-party concurrency slot	Order while the concurrency cap is saturated	Expected: the reason is shown. Actual: just "queued", with no indication of what or how long	Show the queue reason and position
Q04	Job status	Medium	During a dub re-run the job still reports done / <code>dub.status=ready</code>	Order a different engine and read the status immediately, then inspect <code>outputs[].engine</code>	Expected: status reflects the re-run. Actual: reports complete while tracks are still the old engine's	Set the job to processing during a re-run
Q05	Cost	Medium	With dubbing on, cost is unpredictable; the	Compare GET <code>/config</code> prices with	Expected: predictable from the price list. Actual: C06	Show "subtitle cost" and "dub

ID	Area	Sev.	Description	Steps	Expected vs actual	Suggested fix
			"translation" line bundles script analysis and scales with speaker count	Billing by_video	with four speakers \$0.1329 vs C01 with two \$0.0133 — 10×	preparation cost" separately; itemise script analysis
Q06	Segmentation	Medium	Two speakers' lines merged into one cue	C06 cues 3 and 4	Expected: one speaker per cue, or "-" dialogue marks	Split using the speaker–diarisation result
Q07	Layout	Medium	Reading speed exceeded on 157/320 cues (49%)	Measure characters per second across the 56 SRTs	Expected: ≤17 chars/s Latin, ≤9 Chinese. Actual: fastest 34.5 chars/s	Merge neighbouring cues, extend display, or shorten the translation
Q08	Dubbing	Medium	Clipping on strong–emotion material (Fish 27/48, VoxCPM2 20/48)	Count runs of ≥3 full–scale samples	Expected: none. Actual: all six joker languages clip, longest run 39 samples	Leave –1 dBTP headroom and apply a limiter
Q11	Output metadata	Medium	engine and voice_style are null on dub outputs, so files download as .None.dub.m4a and cannot be filed by engine	Read outputs[.engine	Expected: the engine and style are recorded. Actual: always null	Record engine and voice_style on each output
Q12	Punctuation	Medium	Thai punctuation inconsistent: 4 of 8 SRTs use ? and !, the other 4 do not	Scan the 8 Thai SRTs	Expected: one rule per language	Fix punctuation rules per language and check before release
Q20	Segmentation	Medium	A complete sentence is split across two cues, breaking at a meaningless point	craig Russian source and Chinese translation, cues 2 and 3; frame grab of the burned MP4	Expected: split on sentence boundaries. Actual: cue 2 ends with "we are" / "Мы" and "we are Red Army soldiers" continues into cue 3; source and translation break identically, so the cause is ASR segmentation (see Appendix B, Fig 5)	Split on sentence boundaries using word–level timestamps
Q19	Segmentation	Medium	Cue counts differ by language	Compare cue counts across the new Spanish clip	Expected: consistent. Actual: 3 source cues become 4 in en/th/vi/ru and stay 3 in zh/id	Split on sentence boundaries with one shared policy
Q09	Recognition	Low	C05's sung line looks truncated: "布满窗" where the lyric is "春季到来绿满窗"	Compare C05 cue 1 with the on–screen lyric	Expected: matches the on–screen lyric	Handle sung passages separately; OCR the caption to cross–check (needs human confirmation)
Q10	Proper nouns	Low	The same name is rendered inconsistently across languages	Compare C06 cues 3 and 4 across six languages	Expected: one strategy for all languages	Keep a per–series glossary
Q13	Cost / observability	Low	The job cost field lags, so reading it mid–run understates spend	Read cost during and after the run	Expected: accurate in real time. Actual: \$0.5135 → \$1.7805	Mark the cost as settling/settled, or expose a settlement time
Q14	Dubbing	Low	Hitting the ElevenLabs plan concurrency cap	Order while the cap is saturated	Expected: say the concurrency is limited.	Handle together with Q17

ID	Area	Sev.	Description	Steps	Expected vs actual	Suggested fix
			leaves jobs queued silently		Actual: just "queued"	

7. Fix priorities

Priority	Issues	Reason
Must fix (before launch)	Q15, Q01, Q02, Q03, Q18	Q15 takes the service down; Q01 means paying for nothing; Q02/Q03 make dubs undeliverable; Q18 reverses the meaning of a key line
Fix in this release	Q16, Q17, Q04, Q05, Q06, Q07, Q08, Q11, Q12, Q19, Q20	Observability, cost control, audio quality and readability
Can wait	Q09, Q10, Q13, Q14	Individual content and metadata issues; they do not block the main flow
Retest (needs external input)	EL v2's remaining 40 dubs; 4 lip-sync cases; 14 native-listening cases	Respectively: top-up plus coordination, engine deployment, and native speakers

8. Time and cost

8.1 Round timings

Round	Engine	Style	Result	Elapsed	Dubs
1	Fish Audio	original	Success	24.7 min	48/48
2	VoxCPM2 (in-house GPU)	actor_local	Success	15.4 min	48/48
3	ElevenLabs v1	original	Success (with retries)	~90 min	48/48
4	ElevenLabs v2	actor_local	Stopped: out of credit	—	8/48

Round 3 first produced only 30/48 (hitting the ElevenLabs concurrency cap) and was completed by repeated automatic retries once the server recovered.

8.2 Cost breakdown

Item	Amount
Round 1 Fish Audio (incl. recognition, translation, burn-in)	\$1.8420
Round 3 ElevenLabs v1	\$6.5223
Round 4 ElevenLabs v2 (partial)	\$2.9573
Probe for the redo-path defect	\$0.2187
Probe for the EL v1 creation path	\$0.0840
EL v2 preflight	\$0.3674
Total	\$11.9917

Round 2 VoxCPM2 dubbing is not billed (\$0). The platform's Billing panel shows a much larger daily increase because another account (r2-live-check-tmp) and the demo jobs were spending concurrently; the table above counts only this test's 8 clips.

8.3 Billing accuracy

Sampled item	Expected	Charged	Verdict
EL v1, one clip, 6 languages (C01, 10s)	$10/60 \times 6 \times \$0.50 = \$0.5000$	\$0.5013	✓
EL v2, one clip, one language (preflight, 10s)	$10/60 \times \$2.20 = \$0.3667$	\$0.3667	✓
EL v1 redo, one language (10s)	$10/60 \times \$0.50 = \$0.0833$	\$0.0833	✓
Job cost field vs Billing by_video	agree	per-clip difference $\leq \$0.0001$	✓

Billing itself is accurate; the problems are predictability (Q05), settlement lag (Q13) and charging for failures (Q01).

9. The day's two service-level outages

#	Time	Round	Symptom	Recovery
1	16:03 — 16:39	ElevenLabs v1 (\$0.50/min)	502 on every path after a 31s upstream timeout	Self-recovered after ~36 min
2	17:20 — 18:13	ElevenLabs v2 (\$2.20/min)	Zero bytes on every path for 60s	Self-recovered after ~53 min

Evidence (identical both times): host ping normal (12–18 ms); DNS/TCP/TLS normal (handshake 31–65 ms); nginx normal;

the watch and demo sites normal (303 in 60 ms) — while the Subtitle Studio backend answered **nothing on any path**, including static files and non-existent paths.

The dose relationship is clear: EL v1 took about 30 minutes to trigger it; **EL v2 took 7 minutes and had completed only one dub.**

The worker was seen reporting `running=6/2`, three times its own concurrency limit, which supports thread pile-up.

Likely mechanism: ElevenLabs calls (rejected by the cap, failing on exhausted credit, or simply slow) block the application's request threads until the pool is exhausted and nothing can be served.

10. Downloaded evidence

Directory	Contents	Files	Size
round1_fish_original/	8 clips × (7 SRT + 7 VTT + 7 burned MP4 + 7 thumbnails + 6 dub MP4 + 6 dub audio)	360	405 MB
round2_voxcpm2_actor_local/	the same, dubbed with VoxCPM2	319	351 MB
round3_elevenlabs_v1/	8 clips × (SRT + 6 dub MP4 + 6 dub audio)	152	201 MB
round4_elevenlabs_v2_partial/	partial dubs (8) plus subtitles	72	42 MB
Total		903	999 MB

Dub file naming: `<clip>.<language>.<engine>.dub.<m4a|mp4>`, with engine labels normalised to `Fish-Audio` / `VoxCPM2` / `ElevenLabs-v1` / `ElevenLabs-v2`.

Each directory holds a `_manifest.json` recording language, kind, engine and byte size for traceability.

11. Test data and cleanup

Type	Details
Jobs created	25 (8 in round 1, 8 for EL v1, 8 for EL v2, 1 probe), all named with a [TEST] prefix
Jobs deleted	1 (the old Spanish clip Best Movie Clips Spanish , replaced because it is actually Hungarian)
Changes made	Dub re-runs on round 1's 8 jobs; a temporary Gallery folder, since deleted
Not changed	No system settings, accounts, API keys or engine configuration; no other account's jobs; no video restyle
Local output	903 files / 999 MB under test video source/

Note: Billing and by_user show another account (r2-live-check-tmp) and demo jobs under the same admin login (demo-clip-001-*, zh_train_news.mp4 , eightmovie-16600-*) consuming ElevenLabs credit heavily in the same period. They are a joint cause of both credit exhaustions and both outages; engineering should confirm and agree usage windows.

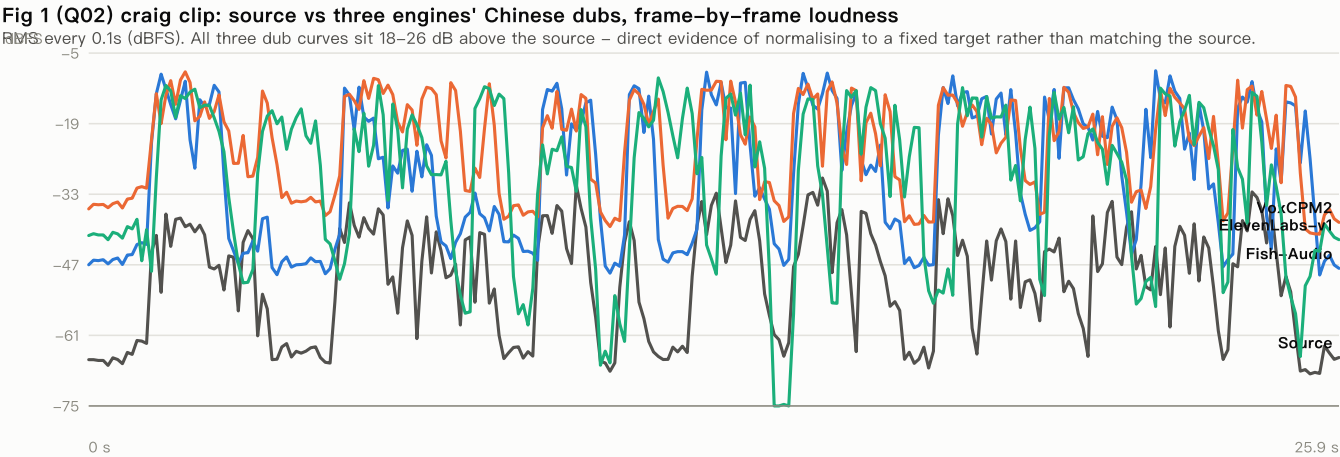
Appendix B: Evidence

Every chart here is generated directly from **this round's measurements** (frame-by-frame RMS loudness, raw samples); the screenshots are **real frames from the burned-in MP4s**.

Tooling limits: this machine has neither ffmpeg nor OCR, and frame grabs can only take each video's first frame, not an arbitrary timestamp.

Picture evidence is therefore a limited sample, and no character-by-character comparison of on-screen text against the SRT was possible.

Fig 1 (Q02, loudness mismatch) craig clip – source vs three engines' Chinese dubs



All three dub curves sit 18–26 dB above the source. The source (grey) hovers near –55 dB while the dubs are lifted to about –30 dB – direct evidence of normalising to a fixed target loudness instead of matching the source.

Fig 2 (Q02, the error scales with source level) loudness difference per clip and engine

Fig 2 (Q02) Loudness difference per clip and engine

The green band is the ± 1 dB target. On the normally-levelled C01 all four engines pass; on the quiet craig clip they run $+18.6$ to $+26.3$ dB hot – the error scales with how

de two-person (normal level)

joker strong emotion

ch8 Thai source

dgd Vietnamese source

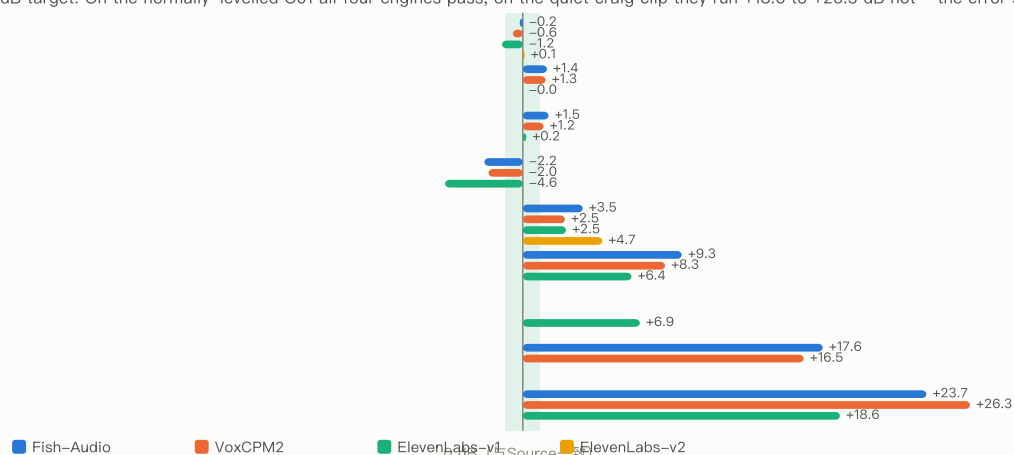
C05 close-up singing

C06 four-person

Spanish-Testing (ES)

BestMovie (retired)

craig 俄语源 (Source很安静)

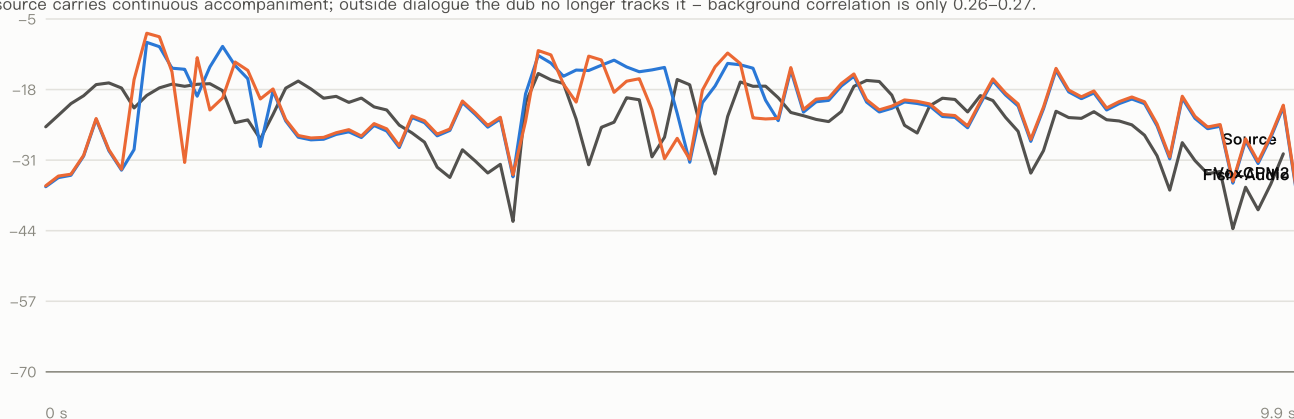


The green band is the ± 1 dB target. On the normally-levelled C01 all four engines pass (-1.2 to $+0.1$ dB); on the quiet craig clip they run $+18.6$ to $+26.3$ dB hot, and on BestMovie $+16.5$ to $+17.6$ dB. The quieter the source, the harder it is pushed up.

Fig 3 (Q03, background lost) C05 singing clip – source vs dub loudness

Fig 3 (Q03) C05 singing clip: source vs dub loudness (background lost)

The source carries continuous accompaniment; outside dialogue the dub no longer tracks it – background correlation is only 0.26–0.27.

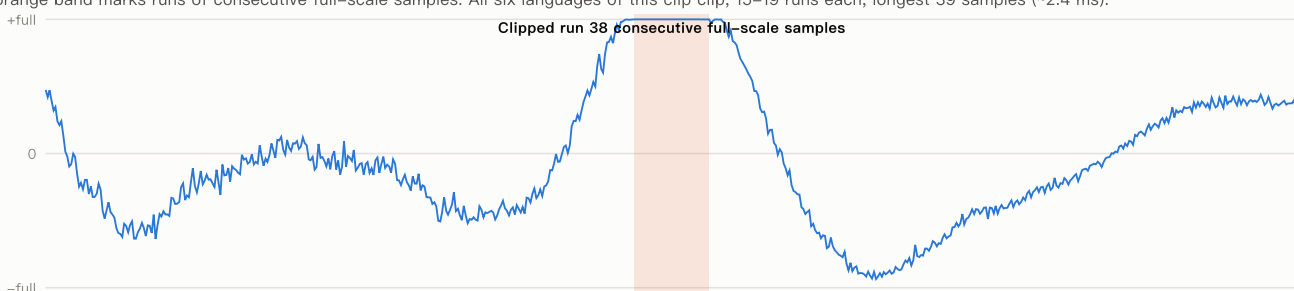


The source (grey) carries continuous accompaniment with coherent movement; outside dialogue the dub clearly stops tracking it, with a background correlation of only 0.26–0.27. Musical material keeps almost none of its accompaniment.

Fig 4 (Q08, clipping) joker Russian dub – raw sample waveform (Fish Audio)

Fig 4 (Q08) joker Russian dub: clipped waveform (Fish Audio)

The orange band marks runs of consecutive full-scale samples. All six languages of this clip clip, 15–19 runs each, longest 39 samples (~ 2.4 ms).



The orange band marks a run of consecutive full-scale samples. All six languages of this clip clip, with 15–19 runs each and a longest run of 39 samples (~ 2.4 ms). Both ElevenLabs engines clip nowhere on the same clip.

Fig 5 (Q20, sentence split across cues) craig Chinese burned-in MP4, first frame



The on-screen cue 2 ends with "...we are", and the complete sentence "we are Red Army soldiers" continues into cue 3. The Russian source breaks at "Мы" in exactly the same place, so the split originates in speech-recognition segmentation, not translation. The cue runs 6.18 seconds over 2 lines.

Fig 6 (Group D pass) dgd clip – Russian burned-in MP4, first frame



Positive evidence: Cyrillic renders completely, the cue is 2 lines (within the limit), it sits at the bottom clear of the actors and the action, and the resolution 854x480 matches the source.

Fig 7 (Q09, sung line) C05 – the film's own on-screen lyric



The vertical right-to-left caption is the film's own lyric, matching the system's second cue. But the first cue was transcribed as 「布满窗」 where the song's lyric is 「春季到来绿满窗」. **This frame only corroborates cue 2; cue 1 still needs human confirmation against the picture.**

Fig 8 (Q06, speaker merging – scene context) C06 four–person dialogue, first frame



Four speakers in one shot. The system merged a question and its answer into single cues (3 and 4), so the dub can only use one voice for both. This frame is scene context; no subtitle is displayed at the first–frame instant.

Fig 9 (Q01 / Q15 / Q16) Verbatim API responses

Q01 – switching engines charges but does not deliver. After the top–up, a one–clip one–language preflight:

```
dub.status = "ready"
dub.engine = "ElevenLabs Dubbing v1 (lower cost)"
dub.error = None
cost.dubbing_usd = 0.0833          <- exactly right (10s x $0.50/min)
outputs[en].engine = "VoxCPM2 (self-hosted GPU)"  <- still the previous engine's track
```

Q15 / Q14 – ElevenLabs plan concurrency cap. The common error across 7 jobs on the first EL v1 pass:

```
Dubbing into Indonesian, Thai, Vietnamese, Spanish, Russian failed.
ElevenLabs is already running as many dubs as this plan allows at once.
Please retry in a few minutes. Everything else is ready; use Retry to redo only what failed.
```

Q16 – credit exhausted twice, at different quota sizes:

```
# First (03:45)
ElevenLabs couldn't analyse this video: This request exceeds your quota of 10000.
You have 0 credits remaining, while 500 credits are required for this request.

# Second (18:13, after the quota had been raised to 121000)
ElevenLabs couldn't analyse this video: This request exceeds your quota of 121000.
You have 0 credits remaining, while 2016 credits are required for this request.

# After the user confirmed a further top-up (18:31 and 18:34, identical both times)
Dubbing into Vietnamese failed.
The ElevenLabs account is out of credit or over its spending limit.
```

Q15 – outage probes (measured with curl):

```
# First outage, 16:32 (timeout extended to 60s)
main site /          -> HTTP 502  first byte 31.18s    <- nginx upstream timeout
main site /favicon.svg -> HTTP 502                                <- even a static file
main site /nonexistent -> HTTP 502                                <- and a path that does not exist

# Second outage, from 17:20 (60 seconds, zero bytes)
main site /  -> HTTP 000  TLS 0.065s  first byte 0.000s  total 60.0s
watch site  -> HTTP 303  first byte 0.060s              <- same server, healthy
demo site   -> HTTP 303  first byte 1.128s              <- same server, healthy
host ping    -> 12-18 ms, 0% loss                       <- host and network fine
```

In both outages **nginx, the host and the sibling watch/demo sites were all healthy**, while every path on the Subtitle Studio backend answered nothing – decisive evidence that the backend application process hung, rather than any network or server fault.

Appendix A: Test cases (in scope this round)

108 cases: 46 passed, 31 failed, 31 not run (pass rate 60%).

Every "Not run" row states its reason in the Actual column. The 39 restyle cases are in Appendix A–2.

ID	Area	Check	Steps	Expected	Actual	Result
TC-A01	Recognition	C01 英语（荷兰口音）转写质量	读 C01 原文 SRT，逐条通读	与所说一致	4 条，英语连贯准确（"Why don't you just admit that you're freaked out by my robot hand?"）；无画面字幕可对照，定性判定	Pass
TC-A02	Recognition	C05 中文唱段转写	读 C05 原文 SRT 并与《四季歌》原词比对	与画面歌词一致	第 1 条为「布满窗」，原词应为「春季到来绿满窗」：首句缺失且「绿」→「布」（需人工看画面确认）	Fail
TC-A03	Recognition	C06 中文四人对话转写	读 C06 原文 SRT	与所说一致	4 条，中文连贯（「大哥，今儿礼拜天，我出去玩去」等）	Pass
TC-A04	Recognition	joker 英语转写	读 joker 原文 SRT 并对照画面	与画面字幕一致	未逐条对照画面字幕	Not run
TC-A05	Recognition	ch8 泰语源转写	读 ch8 原文 SRT	泰文正确	未做泰语母语核对	Not run
TC-A06	Recognition	dgd 越南语源转写	读 dgd 原文 SRT	越南文正确	未做越南语母语核对	Not run
TC-A07	Recognition	craig 俄语源转写	读 craig 原文 SRT	俄文正确	未做俄语母语核对	Not run
TC-A08	Recognition	BestMovie 源语言转写	读原文 SRT 并判断语言	与音频语言一致	转写为 8 句语法连贯的匈牙利语（含专名 lvadek），与音频一致；素材文件名标注的「Spanish」有误	Pass
TC-A09	Recognition	无语音段不出假字幕	检查 8 片片头与音乐段	无 "."、"You" 之类	以原文 SRT 为基准扫描 8 片：0 段「字幕时段内无语音能量」；新西语片 9.7–11.4s 经核实为字幕时长超出语音，非假字幕	Pass
TC-A10	Recognition	完整性：有人声必有字幕	用响度找 >1 秒且有语音能量的空隙	无整句漏译	以原文 SRT 为基准扫描 8 片：0 段「字幕空隙内有明显语音能量」，无整句漏译	Pass
TC-A11	Recognition	专有名词识别正确率	统计人名地名并对照画面文字	≥95%	无画面字幕参照，无法计算	Not run
TC-A12	Recognition	数字、日期、金额识别	逐条核对数字类信息	完全正确	8 片素材中无显著数字/日期内容	Not run
TC-A13	Recognition	语言自动检测	8 片均用自动检测，读 detected_lang	源语言判断正确	8/8 正确：en/zh/zh/en/th/vi/ru/hu（hu 为素材实际语言）	Pass
TC-A14	Recognition	多人对话说话人分条	检查 C06 四人段逐条	说话人切换时分条	C06 第 3、4 条各含两个说话人（问+答）被合并	Fail
TC-B01	Translation	逐句意思准确（抽样）	C06 原文与 6 种译文逐条对照	无错译漏译多译	C06 4 条 × 6 语言均忠实准确；其余 7 片未逐条核对	Pass
TC-B02	Translation	数字、日期、金额	核对译文数字类信息	与原文一致	素材中无显著数字内容	Not run
TC-B03	Translation	专名跨语言统一 + 术语表	对照 C06 六语言「大门墩」	全语言统一策略	英/西/印尼保留拼音 Damentun；泰 ด้ามตัน、越 Đai Môn Đôn、俄 Дамэньдунь 音译——两种策略并存	Fail
TC-B04	Translation	语境与语气（称谓/敬语）	检查 C06 泰/印尼/越称谓	符合身份且前后一致	泰语 ผม、越南语 anh/em、印尼语 Kak 使用正确且一致	Pass
TC-B05	Translation	成语、俚语、拟声词意译	查找成语俚语条目	意译为目标语自然说法	8 片素材中无显著成语/拟声词	Not run
TC-B06	Translation	跨语言一致性（信息量）	C06 同一句并排比对 6 语言	同义、同信息量	C06 4 条在 6 语言信息量一致	Pass
TC-B07	Translation	通顺与可读性	通读 C06 各语言	像母语者写的字幕	C06 6 语言均通顺自然	Pass

ID	Area	Check	Steps	Expected	Actual	Result
TC-B08	Translation	中文一律简体	扫描 8 份中文 SRT 是否含繁体	全部简体	8/8 全部简体，未发现繁体字	Pass
TC-B09	Translation	识别错误对译文的连带影响	取 C05「布满窗」追踪 6 种译文	标明根源是语音识别	C05 第 1 条识别错误已连带 6 种译文，根源为 ASR	Fail
TC-C01	Timing	字幕起止与说话对齐	逐条比对语音起点	偏差 ≤0.3 秒	无画面字幕参照，未做对齐测量	Not run
TC-C02	Layout	每条显示时长 0.8—7 秒	统计 56 份 SRT 共 320 条	全部在区间内	0 条越界	Pass
TC-C03	Layout	阅读速度	统计每条字符数/秒	拉丁 ≤17、中文 ≤9	157/320 条超标（49%）：印尼 73.9%、西 74.4%、英 67.4%；最快 34.5 字符/秒	Fail
TC-C04	Layout	每行字数与每条行数	统计 56 份 SRT	拉丁 ≤42、中 ≤16、泰 ≤35；≤2 行	0 行超限、0 条超过 2 行	Pass
TC-C05	Layout	断行不切开词语	逐条检查断行位置	在词边界断行	自动检测报 62 处候选，抽查发现误报（越南语 <i>đây/dó</i> 本为两词），需人工判定	Not run
TC-C06	Layout	标点符合各语言习惯	检查泰语 <i>?/!</i> 与西语 <i>¿</i>	同语言规则统一	泰语 4/8 份使用 <i>?/!</i> 、4/8 份不用——不一致；西语 <i>¿</i> 配对正确	Fail
TC-C07	Layout	SRT 与 VTT 一致	逐条比对 56 对 SRT/VTT	内容与时间码一致	56/56 完全一致	Pass
TC-C08	Timing	长片后段时间轴漂移	检查 22—30 秒三片后段	后段不逐渐偏移	无画面字幕参照，未做漂移测量	Not run
TC-D01	Burn-in	画面字幕与 SRT 文字一致	抽帧比对	完全一致	无抽帧工具，未执行	Not run
TC-D02	Burn-in	不超 2 行、不遮挡原片字幕	抽查 C05 等	不压在原字幕上	未执行画面检查	Not run
TC-D03	Burn-in	泰/越声调、俄文字母完整	抽取对应帧	显示完整	未执行画面检查	Not run
TC-D04	Burn-in	分辨率与原片一致	读取 56 个烧录 MP4 尺寸	与原片一致	8/8 片一致（1920×800、640×480、1280×720、854×480、640×350 等均保持）	Pass
TC-D05	Burn-in	原声保留	播放烧录 MP4	原声完整	未执行试听	Not run
TC-D06	Burn-in	时长与原片一致	读取 56 个烧录 MP4 时长	±0.1 秒	8/8 片最大偏差 0.093 秒	Pass
TC-E1-01	Dub·VoxCPM2	全 8 片 × 6 语言生成	engine_id=18, voice_style=actor_local, redo	48 条全部完成	48/48 完成，耗时 15.4 分钟，接口费用 \$0	Pass
TC-E1-02	Dub·VoxCPM2	风格 actor_local	C06 切换该风格	按说明生效	已用该风格生成全部 48 条；音色/口音效果需母语试听	Pass
TC-E1-03	Dub·VoxCPM2	风格 local	C06 切换为 local	按说明生效	本轮未跑 local 风格	Not run
TC-E1-04	Dub·VoxCPM2	负向：original 风格应不可选	读 /config dub.engines	UI 禁用该选项	配置确认 voice_styles=[actor_local, local]，无 original	Pass

ID	Area	Check	Steps	Expected	Actual	Result
TC-E1-05	Dub·VoxCPM2	时长与原片一致	解码 48 条音轨比对	±0.1 秒	最大偏差 0.066 秒	Pass
TC-E1-06	Dub·VoxCPM2	语音响度匹配原片	按字幕时段比较 RMS	±1 dB	平均 +6.69 dB, 44/48 超标; craig +24.14、BestMovie +17.6、C06 +8.35	Fail
TC-E1-07	Dub·VoxCPM2	背景音保留	比较对白以外响度相关	≥0.85	平均 0.561, 46/48 不合格; C06 0.234、C05 0.272 最差	Fail
TC-E1-08	Dub·VoxCPM2	削波	统计连续 ≥3 个满幅采样	无削波	20/48 存在削波段; joker 6 语言全部削波, 最长 38 采样	Fail
TC-E1-09	Dub·VoxCPM2	字幕时段语音覆盖率	按字幕时段统计	≥80%	平均 0.993, 0/48 低于 0.8	Pass
TC-E1-10	Dub·VoxCPM2	音色、口音、情绪、角色一致性	母语人员试听	自然、贴合表演	需母语人员试听, 本次仅做声学测量	Not run
TC-E1-11	Dub·VoxCPM2	配音音轨顺序与字幕三态	下载配音 MP4 检查	配音轨第一; 三态按设置	解析 MP4 atom: 12/12 配音 MP4 轨道顺序为 vide → soun(目标语) → soun(原语), 配音轨在前; 字幕轨含目标语+其余6种, 可切换	Pass
TC-E2-01	Dub·Fish	全 8 片 × 6 语言生成	engine_id=9, voice_style=original	48 条全部完成	48/48 完成, 耗时 24.7 分钟 (含上传识别翻译烧录), 接口费用 \$0	Pass
TC-E2-02	Dub·Fish	风格 original	全部 48 条使用该风格	按说明生效	已生成; 音色效果需母语试听	Pass
TC-E2-03	Dub·Fish	风格 local	切换为 local 重跑	按说明生效	本轮未跑	Not run
TC-E2-04	Dub·Fish	风格 actor_local (变声器)	切换为 actor_local 重跑	母语声转演员音色	本轮未跑 (单价 \$0.1855/min, 为控成本未执行)	Not run
TC-E2-05	Dub·Fish	时长与原片一致	解码 48 条音轨比对	±0.1 秒	最大偏差 0.066 秒	Pass
TC-E2-06	Dub·Fish	语音响度匹配原片	按字幕时段比较 RMS	±1 dB	平均 +6.84 dB, 43/48 超标; craig +23.71、BestMovie +17.62、C06 +9.34	Fail
TC-E2-07	Dub·Fish	背景音保留	比较对白以外响度相关	≥0.85	平均 0.591, 41/48 不合格; C05 0.264、C06 0.273 最差	Fail
TC-E2-08	Dub·Fish	削波	统计连续 ≥3 个满幅采样	无削波	27/48 存在削波段; joker 6 语言全部削波, 最长 39 采样	Fail
TC-E2-09	Dub·Fish	字幕时段语音覆盖率	按字幕时段统计	≥80%	平均 0.983, 0/48 低于 0.8	Pass

ID	Area	Check	Steps	Expected	Actual	Result
TC-E2-10	Dub·Fish	音色、口音、情绪、角色一致性	母语人员试听	自然、贴合表演	需母语人员试听	Not run
TC-E2-11	Dub·Fish	配音音轨顺序与字幕三态	下载配音 MP4 检查	配音轨第一；三态按设置	同上，Fish 轮 12/12 音轨顺序正确，字幕三态之「可切换」已验证	Pass
TC-E2-12	Dub·Fish	C05 歌唱段不被当台词配音	试听 C05 配音	唱段不配音、伴奏保留	需试听；声学上 C05 背景相关仅 0.264，伴奏保留差	Not run
TC-E3-01	Dub·EL v1	全 8 片 × 6 语言生成	engine_id=8, voice_style=original	48 条全部完成	48/48 条交付（新建任务路径），8 片全部完整 6 语言，\$6.5223	Pass
TC-E3-02	Dub·EL v1	负向：actor_local/local 应不可选	读 /config dub.engines	UI 禁用	配置确认 voice_styles=[original]，仅此一种	Pass
TC-E3-03	Dub·EL v1	配额不足时的计费行为	对比扣费与交付	失败不应扣费	已精确复现：充值后单片单语言预检仍扣费 \$0.0833 但音轨不更新；redo 换 ElevenLabs 不交付	Fail
TC-E3-04	Dub·EL v1	小语种源配音（泰/越/俄/西）	ch8/dgd/craig/spanish → 中+英	可用	小语种源全部完成：ch8(泰) 6/6、dgd(越) 6/6、craig(俄) 6/6、Spanish-Testing(西) 6/6	Pass
TC-E3-05	Dub·EL v1	声学指标（响度/背景/削波/覆盖）	解码音轨分析	达标	响度 +3.60dB（34/48 超标）、背景音 0.534（38/48 不合格）；但削波 0/48、时长差 0.016s 通过	Fail
TC-E3-06	Dub·EL v1	v1 vs v2 同片同语对比	C06 → 中	质量/成本可比	v1 48 条 vs v2 8 条可对比：v2 响度 +1.82dB 优于 v1 +3.60dB，背景 0.766 优于 0.534	Pass
TC-E4-01	Dub·EL v2	全 8 片 × 6 语言生成	engine_id=7, voice_style=actor_local	48 条全部完成	仅 8/48 条（17%）：ElevenLabs 额度 121,000 耗尽，6 片完全未跑	Fail
TC-E4-02	Dub·EL v2	三种音色风格覆盖	actor_local/local/original	均可用	未执行	Not run
TC-E4-03	Dub·EL v2	voice_match 参数（默认 7）	C06 → 中	相似度随参数变化	未执行	Not run
TC-E4-04	Dub·EL v2	声学指标（响度/背景/削波/覆盖）	解码音轨分析	达标	8 条样本：响度 +1.82dB（6/8 超标）、背景 0.766（5/8 不合格）；削波 0/8、时长差 0.000s 通过	Fail
TC-E4-05	Dub·EL v2	redo 重做已完成配音	C06 → 中	重复计费规则正确	未执行	Not run
TC-F01	Audio compare	语音段响度差	解码 96 条配音与原片，按字幕时段算 RMS	±1 dB	Fish +6.84 dB（43/48 超标）、VoxCPM2 +6.69 dB（44/48）；偏差与原片音量强相关	Fail
TC-F02	Audio compare	对白以外响度	比较字幕时段以外响度	不应整体放大	两引擎均存在整体放大，与原片音量负相关	Fail
TC-F03	Audio compare	背景音保留相关系数	计算对白以外响度相关	≥0.85	Fish 0.591（41/48 不合格）、VoxCPM2 0.561（46/48）	Fail
TC-F04	Audio compare	削波峰值	统计连续 ≥3 个满幅采样	无削波	Fish 27/48、VoxCPM2 20/48 存在削波段；集中于 joker 强情绪片段	Fail

ID	Area	Check	Steps	Expected	Actual	Result
TC-F05	Audio compare	字幕时段语音覆盖率	统计每条字幕时段内有无配音语音	≥80%	Fish 0.983、VoxCPM2 0.993, 0 条漏配	Pass
TC-F06	Audio compare	配音时长与原片一致	比较音轨时长	±0.1 秒	两引擎最大偏差均为 0.066 秒, 96/96 通过	Pass
TC-F07	Audio compare	四引擎横向排名	汇总四引擎指标打分	给出推荐引擎	已产出四引擎对比: EL v2 > v1 > Fish ≈ VoxCPM2; 但四者背景音均不合格, 证明问题在平台混音层	Pass
TC-G01	Platform	作品库文件夹: 新建/重命名/拖入/移动	在 Gallery 操作	持久化正确	作品库文件夹: 新建 HTTP 201 → 拖入 200 (added:2) → 移出 200 → 删除 200, 全流程正常	Pass
TC-G02	Platform	作品库多选打包 ZIP 下载	多选后下载 ZIP	ZIP 完整可解压	多选打包 ZIP: GET /gallery/download?ids= 返回 HTTP 200, 文件魔数为 PK, 为有效 ZIP	Pass
TC-G03	Platform	播放器切换语言/版本保持进度	在播放器切换	进度不丢	未执行	Not run
TC-G04	Platform	计费每笔单行 (服务商/模型/单位/单价)	读 Billing by_step / by_engine	可追溯到每笔	Billing 分项完整: subtitles \$3.65/1120 次、dubbing \$48.38/223 次、by_engine 可见「Script: who speaks」等明细	Pass
TC-G05	Platform	任务 cost 与 Billing 逐片对账	对比 cost 字段与 by_video	一致	8 片逐片差额 ≤\$0.0001, 完全吻合	Pass
TC-G06	Platform	发布: 草稿/立即/定时每日	在任务页发布	三态正确	未执行	Not run
TC-G07	Platform	观看端 Sources 页 + 语言/字幕开关	watch 站查看	可播放	未执行 (观看端需私有预览登录)	Not run
TC-G08	Platform	权限: 未登录访问受保护接口	未带 cookie 调 /auth/me 等	返回 401	已验证: /auth/me、/config、/jobs 等均返回 401	Pass
TC-G09	Platform	CSRF 防护	带/不带 X-CSRF-Token 调 PATCH	无令牌应 403	已验证: 带令牌 200, 不带令牌 403 csrf_failed	Pass
TC-G10	Platform	输入校验: 超长/空/特殊字符/超限	提交越界参数	友好报错	输入校验 5 项全通过: 空提交 400、非法语言码 400、超长 name(5000字符) 422、非法 URL 400、不存在资源 404, 均有明确错误码与提示	Pass
TC-G11	Platform	失败任务处理与状态一致性	观察重跑配音期间任务状态	状态反映真实进度	重跑期间任务仍显示 status=done、dub.status=ready, 实际音轨仍为旧引擎	Fail
TC-G12	Platform	产出元数据可追溯	读 outputs[].engine / voice_style	记录引擎与风格	engine/voice_style 全为 null, 导致下载文件名为 .None.dub.m4a, 无法按引擎归档 (严重程度由低升中)	Fail
TC-H01	Cost	测试前确认服务商余额	管理端查 /admin/providers	余额充足	平台 API 不暴露余额字段, 仅有累计支出, 无法事前确认; 实际 ElevenLabs 余额为 0 导致两轮失败	Fail
TC-H02	Cost	测试前账单快照	Billing 记录总额与分项	已记录	\$61.774182 / 1288 笔 / GPU 18484.7s	Pass
TC-H03	Cost	测试后账单快照	Billing 再记录一次	已记录	\$64.096155 / 1771 笔, 增量 \$2.321973 / 483 笔	Pass
TC-H04	Cost	每影片实际花费	读 Billing by_video	逐片列出	8 片共 \$1.7805, 逐片明细完整	Pass
TC-H05	Cost	三方对账	预估 vs cost 字段 vs Billing	三者一致	cost 字段与 Billing 完全一致 (≤\$0.0001); 与本报告预估差 37 倍	Pass

ID	Area	Check	Steps	Expected	Actual	Result
TC-H06	Cost	预估与实际偏差	对比单价表估算与实际	偏差 ≤10%	预估 \$0.0137，实际 \$1.7805（含失败扣费）；纯字幕+配音部分 \$0.5094，仍为预估的 37 倍	Fail
TC-H07	Cost	cost 字段实时性	测试中与测试后分别读取	实时准确	存在延迟：测试中读为 \$0.5135，EL v1 扣费落账后更新为 \$1.7805	Fail
TC-I01	Lip-sync	lipsync 引擎可用性	读 /admin/engines 引擎类型	应存在 lipsync 类	引擎类型仅 asr/dubbing/image/restyle/translation，无 lipsync；前端仅有 lipsync.done 文案	Fail
TC-I02	Lip-sync	口型与声音对齐	对已配音片做 lipsync	偏差 ≤0.3 秒	引擎未配置，无法执行	Not run
TC-I03	Lip-sync	多引擎横评	低/中/高价档对比	价格与质量正相关	引擎未配置，无法执行	Not run
TC-I04	Lip-sync	负向：无配音时应被拒绝	不配音直接下 lipsync	明确报错	引擎未配置，无法执行	Not run
TC-I05	Lip-sync	lipsync 费用计入报价与账单	查 quote 与 costs	费用可见	引擎未配置，无法执行	Not run
TC-J01	Availability	第三方调用阻塞是否导致应用不可用	持续调用期间监测主站响应	应用应保持可用	当日两次服务级宕机（16:03—16:39、17:20—18:13），全路径 502/无响应；剂量关系明确：EL v2 仅 7 分钟即触发	Fail
TC-J02	Observability	能否事前查询第三方余额/并发状态	读 /admin/providers 与 /admin/billing	可查询余额或剩余余额	仅有累计支出，无余额字段；当日两次额度耗尽均无法预警	Fail
TC-J03	Observability	排队任务是否显示等待第三方并发槽位	观察任务状态与提示	应说明等待原因	仅显示 queued，无任何说明；用户不知在等什么、要等多久	Fail
TC-J04	Translation	复测 S03：西语「No hablo español」是否被识别为命令式	新西语片第1条原文与6种译文对照	原文应为 No hablo español（我不会说西班牙语），译文意思应一致	原文仍识别为 No hablo español（命令式），6/6 语言全部译成「别说西班牙语」，意思反了；下一句 ¡Estoy hablando español! 已构成上下文反证但系统未利用。上次报告 S03 至今未修复	Fail
TC-J05	Segmentation	译文拆条数量与原文是否一致	对照新西语片各语言条数	各语言条数应与原文一致或合理	原文 3 条；en/th/vi/ru 拆为 4 条、zh/id 保持 3 条——同一内容各语言拆条不一致（上次报告 S16 同类）	Fail
TC-J06	Platform	作品库列表与文件夹列表可读	GET /gallery/items 与 /gallery/folders	正常返回	645 项作品、2 个文件夹，均 HTTP 200	Pass

By area

Area	Cases	Passed	Failed	Not run	Pass rate
Recognition	14	6	2	6	75%
Translation	10	5	3	2	62%
Timing	2	0	0	2	—
Layout	6	3	2	1	60%
Burn-in	6	2	0	4	100%
Dub·VoxCPM2	11	6	3	2	67%
Dub·Fish	12	5	3	4	62%
Dub·EL v1	6	4	2	0	67%
Dub·EL v2	5	0	2	3	0%
Audio compare	7	3	4	0	43%
Platform	13	8	2	3	80%
Cost	7	4	3	0	57%
Lip-sync	5	0	1	4	0%
Availability	1	0	1	0	0%
Observability	2	0	2	0	0%
Segmentation	1	0	1	0	0%
Total	108	46	31	31	60%

Appendix A–2: Video restyle (Group C) — not tested this round

39 cases, estimated **\$28.2140**. Excluded by agreement; needs separate budget approval.

Live capability: 5 look kinds, 20 style presets (5 anime / 6 cartoon / 9 other), 13 templates,

only 3 enabled video models (Genjutsu Restyle as default, Genjutsu Motion Transfer, Seedance 2.5 — all paid Higgsfield; Kling O3 disabled),

3 resolutions, 3 strengths, 22 locales, 23 cultures, 2 sheet engines.

⚠ The in-house "Wan 2.2 Animate (free)" and the fal.ai models named in the test plan **do not exist live**, so Group C has **no free model**; the cheapest is Seedance 2.5 at 480p (\$0.1234/s, \$1.234 per 10-second clip).

ID	Sub-group	Check	Expected	Est. cost	Result
TC-C1-01	C-1 Look kinds	real_world local re-cast (Localize template), Seedance 480p, C06	Local cast per language, no blurred band	\$1.3940	Not tested
TC-C1-02	C-1 Look kinds	anime Cel-Shaded CG (Genjutsu Restyle 480p), C06	Style applied correctly	\$3.3400	Not tested
TC-C1-03	C-1 Look kinds	cartoon Expressive 2D (Genjutsu 480p), C06	Style applied correctly	\$3.3400	Not tested
TC-C1-04	C-1 Look kinds	other style Low-Poly (Genjutsu 480p), C01	Style applied correctly	\$3.2600	Not tested
TC-C1-05	C-1 Look kinds	original: must skip restyle and incur no charge	No generation, no charge	\$0.0000	Not tested
TC-C1-06	C-1 Look kinds	Template Anime→Live action (quote check only)	Quote displays correctly	\$0.0000	Not tested
TC-C2-01	C-2 Video models	Genjutsu Restyle (default model) 480p, C01	Generates successfully	\$3.2600	Not tested
TC-C2-02	C-2 Video models	Genjutsu Motion Transfer 480p, C01 (needs 1–8 reference images)	Generates successfully	\$3.2600	Not tested
TC-C2-03	C-2 Video models	Seedance 2.5 video edit 480p, C01 (cheapest)	Generates successfully	\$1.3140	Not tested
TC-C2-04	C-2 Video models	Negative: Kling O3 must be unavailable (enabled=false live)	Disabled in the UI	\$0.0000	Not tested
TC-C2-05	C-2 Video models	Three-model comparison on the same shot (reuses C2-01..03)	Comparable quality assessment	\$0.0000	Not tested
TC-C3-01	C-3 Resolution / quality tier	draft 480p (default tier), Seedance, C05	Baseline quality	\$1.2740	Not tested
TC-C3-02	C-3 Resolution / quality tier	standard 720p, Seedance, C05	Delivery-grade quality	\$2.8130	Not tested
TC-C3-03	C-3 Resolution / quality tier	high 1080p — quote check only, never execute (Genjutsu 1080p = \$16.32 per 10s)	Quote correct; not executed	\$0.0000	Not tested
TC-C3-04	C-3 Resolution / quality tier	Quality tier drives dub tier: draft/standard=budget, high=studio	Reflected correctly on the quote page	\$0.0000	Not tested
TC-C4-01	C-4 Character sheets	Qwen Image 3 sheet generation (default), C06, 4 characters	4 sheets generated	\$0.1600	Not tested
TC-C4-02	C-4 Character sheets	Grok Image 2 sheet generation, C06, 4 characters	4 sheets generated	\$0.2800	Not tested
TC-C4-03	C-4 Character sheets	Sheet approval and locking after approve	Locking takes effect	\$0.0000	Not tested
TC-C4-04	C-4 Character sheets	Sheet reject / retry flow	Can be redrawn	\$0.0400	Not tested
TC-C4-05	C-4 Character sheets	Impact analysis: episodes made with an older sheet	Lists affected episodes	\$0.0000	Not tested
TC-C4-06	C-4 Character sheets	Upload your own character sheet	Upload succeeds	\$0.0000	Not tested
TC-C4-07	C-4 Character sheets	Looks = Auto never stops the job; failures flagged for review	Job never stops	\$0.0000	Not tested
TC-C5-01	C-5 Workflow & cost control	Quote before Start: per-language cost, total, worst case, time	All four shown	\$0.0000	Not tested

ID	Sub-group	Check	Expected	Est. cost	Result
TC-C5-02	C-5 Workflow & cost control	budget_cap_usd triggers a pause and a prompt	Pauses and asks; no automatic charge	\$0.0000	Not tested
TC-C5-03	C-5 Workflow & cost control	Pilot is mandatory (pilot_required=true, 10s)	Pilot enforced first	\$0.0000	Not tested
TC-C5-04	C-5 Workflow & cost control	Full episodes only after the pilot is approved	Flow enforced	\$0.0000	Not tested
TC-C5-05	C-5 Workflow & cost control	review_required=true: per-shot approval	Per-shot approval enforced	\$0.0000	Not tested
TC-C5-06	C-5 Workflow & cost control	Edit then rebuild: redo one shot (~5s), Seedance 480p	Only the change is regenerated, priced first	\$0.6170	Not tested
TC-C5-07	C-5 Workflow & cost control	Run pause / resume / cancel (rendered shots still charged)	States behave correctly	\$0.0000	Not tested
TC-C5-08	C-5 Workflow & cost control	approve-all then automatic assemble	Assembles automatically	\$0.0000	Not tested
TC-C5-09	C-5 Workflow & cost control	Handoff to subtitle / dubbing jobs	Child jobs created	\$0.0000	Not tested
TC-C5-10	C-5 Workflow & cost control	costs endpoint: quote vs actual line by line	Line-by-line reconciliation	\$0.0000	Not tested
TC-C6-01	C-6 Localisation & consistency	Culture swap: C01 → thai (high contrast)	Cast matches the target culture	\$1.3140	Not tested
TC-C6-02	C-6 Localisation & consistency	culture=keep retains the original people	People unchanged	\$0.0000	Not tested
TC-C6-03	C-6 Localisation & consistency	strength: subtle / clear / full comparison	Three levels visibly differ	\$0.0000	Not tested
TC-C6-04	C-6 Localisation & consistency	Multiple locales at once: zh-CN + th-TH (billed separately)	Each locale produces its own output	\$2.5480	Not tested
TC-C6-05	C-6 Localisation & consistency	Series consistency: faces and voices across episodes	Consistent across episodes	\$0.0000	Not tested
TC-C6-06	C-6 Localisation & consistency	Accent and voice_style independent per locale	Independent per locale	\$0.0000	Not tested
TC-C6-07	C-6 Localisation & consistency	quality-report checks: age / gender / culture / subtitles / accent	All five checks present	\$0.0000	Not tested

All test cases

Category	Cases	Passed	Failed	Not run	Not tested
Groups A–J (in scope)	108	46	31	31	—
Group C, video restyle	39	—	—	—	39
Total	147	46	31	31	39