

OCR cleanup prompt for video posts

You are analyzing on-screen text (OCR) extracted from a short social media video. You are ALSO given the spoken-audio transcript. Your job is to return **ONLY** the on-screen text that adds **NEW** information beyond what is already spoken in the transcript. Exclude any on-screen text that merely echoes the narration, and exclude incidental/background text that is part of the filmed scene (product packaging, clothing, signage, clocks) rather than a caption intentionally overlaid by the creator.

Each array element must be **ONE FINAL, DISTINCT** on-screen message -- never one detected fragment, OCR read, or re-appearance. The same message is very often split into pieces: across multiple lines or on-screen positions, re-detected across many frames, or re-read with OCR typos/jitter (e.g. "pov: your next leg day" / "Ipov: your next leg day" / "p0v: your next leg day" are the **SAME** message, and a multi-line title split across separate detections is **ONE** message, not one per line). When that happens, **MERGE** all of it into a **SINGLE** array entry containing the complete, corrected text -- do not return one entry per fragment, per repeat, or per OCR variant. Only return multiple entries for messages that are genuinely different pieces of information. If several fragments, read together in time order, form one coherent sentence or flowing multi-line caption, **MERGE** them into that one sentence even if they were detected at different on-screen positions or in a different order -- multi-line captions commonly wrap across positions.

Many detected fragments are isolated words, numbers, or garbled character strings scattered around the shot (e.g. gym equipment labels, product numbers, clock digits, background clutter) -- these are **NOT** captions and must be excluded, even though they were technically detected as on-screen text. A genuine overlaid caption reads as a coherent phrase or sentence. Treat a short (1-2 word), disconnected, or nonsensical fragment as background noise and drop it, **UNLESS** it combines with nearby fragments (similar time/position) into one coherent readable phrase -- in which case merge them per the rule above instead of dropping either piece.

TRANSCRIPT (spoken audio):

[raw audio transcript is inserted here; if no speech was detected, "(no speech detected)" is inserted here]

ON-SCREEN TEXT (OCR, one line per sampled frame with its timestamp):

[the raw OCR text is inserted here, formatted one line per sampled frame as "[idx] t=X.Xs: text"]

Respond with **ONLY** a JSON array (no prose, no code fences). Each element is an object {"text": <the complete, merged caption text>, "start_s": <seconds this message **FIRST** appears, across all its merged occurrences>, "end_s": <seconds it **LAST** appears, across all its merged occurrences>}. Use the frame timing provided to fill start_s/end_s. If the array is empty, respond with [].