Project Video Studio

Comprehensive Technical Overview

Updated 21 October 2026 • Zahidul Islam

01 | Latest project snapshot

Project Video Studio — Comprehensive Technical Overview
Author: Zahidul Islam | Updated: 21 October 2026 | Release scope: local video production

Source repository: https://github.com/Zahidulislam2222/project-video-studio — published under Zahidul Islam. GitHub CI passed all four jobs for source commit 76cfa73: 95 tests were discovered, 94 executed and one optional model test skipped; lint, v2 types, compilation and the three security scanners passed.

The current system turns an approved script, recorded explanation or brief into an editable video project. It prepares narration or transcription, composes reference/style checkpoints, renders layered shots and produces verified viewing masters plus portable source packages. The intended operator creates regular project walkthroughs, technical explanations and social clips, with AI assistance handling routine preparation and the operator retaining editorial approval.

The latest addition is automatic human video matting. A separately configured local model creates synchronized foreground and mask clips; the compositor can place text behind a presenter and another panel in front. Original footage remains part of the editing package. Human cutouts are explicitly treated as derived illustrations rather than unaltered evidence.

Current local evidence: 95 automated tests passed on the configured workstation, including six focused matting checks. Twenty-six v2 modules passed type checks. Whole-source lint, medium/high Bandit, the configured secret-scanning hook and 200 applicable Semgrep rules on 28 v2/test targets passed. The exact 39-package runtime had no known advisories at its recorded audit; that result is time-bound.

Actual portrait and landscape validation films measure 70.733 seconds and 307.133 seconds. Full decoding, frame counts, source hashes, caption timing, package links and measured audio were checked. Later source-note and matting changes have focused real-render checks; earlier full-length artifacts remain intact rather than being relabeled as new renders.

This is local software, with no hosted frontend/backend service or measured public-service availability. Full reference-style reconstruction, seven complete illustration engines, autonomous shot/reviewer orchestration and native layered editor round trips remain unfinished. Natural talking-head edge quality and human listening require separate review. The following pages explain the implemented workflow, its evolution, evidence and remaining roadmap.

Publication verification on 21 October 2026: all 95 installed tests passed in 27.603 seconds. All 36 Python modules passed type checks; the complete staged source passed the configured secret/security hook, Gitleaks, Bandit and Semgrep. Legacy download/process helpers now enforce whole-operation deadlines, bounded raw response bytes and child-process cleanup. These checks do not close the creative capabilities listed as unfinished.

02 | Before → After: dedicated comparison

The comparison preserves the earlier workflow as a historical baseline. “Current” describes implemented local capabilities; it does not turn open creative goals into completed features.

Area	Earlier workflow (v1)	Current local engine (v2)
Production input	Narration-first film files and one main media source per scene.	Script, brief or recorded speech; closed project/shot/layer contracts.
Composition	Primarily flattened scene media, title treatment and captions.	Up to four visual layers, masks, opacity, explicit rectangles and depth-ordered callouts.
Short / long layouts	Portrait and landscape format presets.	Portrait upper/lower, landscape side-by-side, picture-in-picture and custom placement.
Timing	Planning estimates and synthesized-scene timing.	Preparation resolves actual narration before footage offsets; recorded cues preserve continuous timing.
Reference workflow	External references supported manual direction.	Local cut/shot-length/palette analysis and contact sheets; full style understanding remains open.
Style checkpoints	Template-driven motion and editorial boards.	Three actual composed style frames, animation sample, palette profiles and art-direction notes.
Human cutouts	Externally prepared media/masks were necessary.	Local recurrent matting creates synchronized foreground/mask and retains the original source.
3D production	Externally rendered clips could be imported.	Original Blender scene adapter, editable source, opaque MP4 and alpha ProRes4444 output.
Revisions / recovery	Unique exports and last-success pointers.	Hash-bound shot reuse, focused/timestamped revisions, portable restoration and cancellation.
Verification	67 pre-rebuild tests at the recorded baseline.	95 installed tests, real short/long evidence and focused cutout/depth/recovery checks.
Editing interchange	Flattened XML picture/audio interchange.	Portable original inputs plus editable project JSON and Blender sources; picture XML remains baked.
Release boundary	Private workstation setup.	Source-only open-source publication; private media, models and machine configuration excluded.

03 | The production workflow

Start with a script, recorded explanation or brief. A script is spoken as supplied; the engine does not invent claims to fill time. A brief may use the optional local director to propose a structured storyboard. A recording can supply corrected SRT cues or use the configured offline recognition model to draft them. Technical names and factual statements still need review.

Approve source rights and inspect recordings for private content before import. Optional footage is placed in order; a companion view supplies the second side of a comparison or a presenter. Landscape defaults to side-by-side and portrait to upper/lower. A presenter cutout can be requested when a suitable video companion or video recording exists. Missing coverage stops preparation instead of silently looping clips.

Preparation resolves generated narration before assigning source offsets. Recorded speech preserves its shared timeline across shots. A companion video advances continuously, avoiding a restart at every shot boundary. The resulting project records stable shot identifiers, durations, layouts, source notes and audio cues.

Review the readable storyboard, then inspect three composed style frames and an animation sample. These previews use the production compositor, including real footage, masks, callouts and transitions. New visual directions can begin with a consistent reference-image palette profile and explicit typography/texture/motion notes. A palette alone does not establish a complete art style.

Build after those checkpoints. The renderer creates separate shot pictures/audio, assembles them, normalizes sound, generates captions where cues are available and checks the exported masters. A failure keeps its diagnostics and partial files; only a verified export advances the last-success pointer. The operator reviews the full picture, voice, claims and source privacy before sharing.

For regular use, the workstation launcher opens the local menu; option 10 prepares a video from inputs. Command-line users can run prepare, plan, style-test, animation-test and build. A completed package can be revised by stable shot ID or by a timestamp resolved against its prepared timeline. Earlier drafts remain available.

04 | Architecture and data ownership

The application is a local modular Python system. FFmpeg owns deterministic media conversion/composition/decoding; Pillow draws original procedural graphics; optional local engines provide speech, recognition, planning and human matting. Blender is an optional native 3D renderer. The main production path remains usable without a hosted account or cloud inference service.

Project contracts define version, ratio, style, voice, captions and shots. Each shot owns its narration or recorded audio, timing, labels, beats, layers, callouts, transition and source note. Layers define source identity, normalized bounds, fit, trim, opacity and optional mask. A matting origin additionally records the original asset, processing interval and model identity.

The local store uses SQLite transactions for project revisions, imported asset metadata and durable job states. Owner-scoped lookups prevent references to another local account's assets. Revision conflicts fail explicitly. Queued/running/complete/failed/cancelled/interrupted states describe work; a file's existence alone never establishes a completed film.

Configuration has distinct owners: maintained JSON/templates contain production defaults, themes, prompts and content; ignored machine configuration or supported environment overrides contain workstation paths and optional local tool/model selections. Credentials and private recovery notes are excluded from source control. Model, font, tool and source hashes contribute to reuse decisions.

The render recipe binds the source bytes, shot/project inputs, engine code, relevant packages, configuration, tools, fonts and model assets. Compatible completed shots can be reused in a new export; a changed dependency invalidates that reuse. This favors correctness over an unsafe fast path that could attach current metadata to an old picture.

Trust boundaries remain explicit: source media and AI output are untrusted inputs; the operator owns rights and factual approval; optional AI planning has no publishing or shell-action tools; local media tools receive argument vectors rather than untrusted shell expressions. A future hosted architecture would require additional network identity, tenant and operational boundaries.

05 | Visual composition and animation

The v2 compositor supports multiple image/video layers, contain/cover fitting, normalized rectangles, opacity, matching grayscale masks, depth-ordered callouts and authored or narration-bound visual beats. A mask and foreground must remain synchronized; automatic camera motion on masked sources is rejected rather than shifting the subject relative to its alpha.

Portrait and landscape are separately composed. Upper/lower, side-by-side and picture-in-picture layouts use explicit safe placements instead of a universal center crop. Custom rectangles support specialized project walkthroughs. Readability depends on the actual source resolution and UI density, so important controls/code must be reviewed in the delivered frame.

Titles, diagrams, accent reveals and callouts are original procedural graphics. Stills can use bounded push/pull camera movement. Incoming cuts, fades and slides are supported; semantic holds need not follow a fixed rapid-edit cadence. Text that cannot fit at the supported minimum size fails with an actionable error rather than becoming unreadably small.

The seven current presets—technical, editorial, pastel, minimal, pixel, paper and cinematic—provide configurable starting points. They are not seven independent illustration engines. Complex custom illustrations, exact typography/texture recreation and documentary-style camera grammar remain creative-engine work. A named creator reference communicates direction, not a promise to reproduce an entire channel.

Foreground source notes use a bounded contrast-backed region above footage. A focused picture-in-picture check and actual-frame pixel regression cover a prior occlusion problem. Captions use an explicit pixel canvas and ratio-specific safe areas; the portrait correction prevents default subtitle scaling from covering the lower view.

Text can appear behind a supplied or automatically extracted person through lower callout depth, and above the person through foreground depth. The actual cutout validation shows both relationships. This proves compositing behavior; it does not independently establish subject-mask accuracy for every person, background or camera movement.

06 | Narration, captions and audio evidence

Generated narration uses approved male stock voices. The configured choices include American and British English options, with delivery presets that adjust speed and pauses. This is not voice cloning or an emotion-prompt acting interface. Technical pronunciation and performance must be auditioned; measured loudness cannot establish expressive quality.

Generated speech records real phrase timing. Recorded narration keeps its original speech and can use corrected SRT or an offline recognition draft. These paths have different accuracy limits: model recognition can mishear names, and authored cues can be wrong. The engine validates ordered positive cue intervals and bounds them against the prepared shot.

Captions are exported as SRT and pixel-calibrated ASS for burn-in. The portrait validation contains24 cues with at most two lines; its widest measured line fits the configured safe width. The landscape validation contains108 cues. These are checks on the recorded validation artifacts, not a universal claim that any future script will fit.

No music is generated or added. Optional non-musical clicks/whooshes have their own effects stem. Voice, effects/mix and normalized master remain separately available for finer editing. Two-pass normalization is followed by measured integrated loudness and true-peak checks; a silent technical test is identified as silent rather than receiving a fabricated loudness value.

The portrait master measures70.733 seconds, 2,122 frames and1080×1920 output. Its recorded final audio is−16.03 LUFS integrated with−1.73 dBTP true peak. The landscape master measures307.133 seconds, 9,214 frames and1920×1080 output, with−16.03 LUFS and−3.44 dBTP. Both clean/captioned masters passed full decoding and their recorded checks.

These full-length films preceded the last source-credit foreground adjustment and later matting addition. Those later changes have their own focused renders/regressions, while earlier artifacts remain unchanged. Human listening to the entire final film, intelligibility on real devices and editorial approval remain separate responsibilities before publication.

07 | References, 3D and editor boundaries

Local reference analysis measures video duration/dimensions, detects scene-change candidates, calculates shot lengths, estimates a sampled palette and generates a contact sheet. Cut detection can miss dissolves or confuse camera movement with an edit. The measurements are useful evidence for pacing decisions, not a complete reconstruction of transitions, camera language, typography or narrative intent.

Reference-image profiles retain source hashes, extracted colors, readable text contrast and operator art-direction notes. Consistent references help establish a direction; three composed frames and an animation test check that direction against the real script/media. Full style understanding and general online-link acquisition remain unfinished. The local analyzer requires an approved local video.

Public creator references were researched through selected browser page reads and sampled frames. They were not all downloaded, listened to or measured shot by shot. Reference imagery is not included in the source release. A public URL does not grant permission to redistribute another creator’s material, use their identity or imply endorsement.

The Blender adapter has rendered an original editable component scene, native frames, an opaque MP4 and a transparent ProRes4444 MOV. The native .blend source remains editable. These checks establish the implemented adapter and original scene path, not an unlimited library of cinematic documentary environments or verified GPU throughput.

Final Cut Pro7 XML contains picture clips and normalized master audio. The picture graphics are baked into those clips. Source project JSON, masks, original media and Blender scenes remain editable through their own workflows, and sources can be restored from a portable export. XML validity and existing media links do not prove a native editor import/re-export.

Installed Resolve external scripting did not connect in the checked workstation state. OBS remote control was disabled. CapCut can receive ordinary MP4/audio/SRT handoff files, but native project automation was not executed. After Effects/Premiere layered round trips remain unverified. StykAnimate exports may be imported as approved media; its application code is not bundled.

08 | Automatic human matting

The focused cutout upgrade uses a separately configured, hash-pinned RVM ONNX model through local CPU inference. No automatic download, paid endpoint or remote inference fallback is used. The original publisher release was retrieved over HTTPS and its observed bytes pinned locally; the upstream release did not supply an independent signed digest.

Frames are processed sequentially. Four recurrent outputs become the next frame’s inputs and reset for each independent shot. The worker produces a foreground and alpha estimate; those become synchronized lossless Foreground.mkv and grayscale Mask.mkv. Both outputs must match the expected frame count/dimensions and fully decode before a verification receipt is written.

The default processing envelope is30 seconds per shot,1,280 pixels on the long edge, two CPU threads and a30 minute deadline, with bounded source pixels, disk estimates and free-space reserve. Split long presenter segments into shorter shots. Generated clips have an8 GiB per-file bound distinct from ordinary512 MiB inputs, while imported account assets remain subject to the10 GiB transactional quota.

The actual installed check used a2-second,60-frame,512×512 animated crop of a NASA public-domain portrait. CPU inference measured18.148 seconds on that workload. Sample foreground coverage was approximately49.25%, 48.33% and50.64%. These are measured alpha statistics, not confidence scores or a semantic quality rating.

A separate1920×1080 composition visibly places rear text behind the extracted person and a front panel above the torso. Its60-frame master fully decodes; the original animated source, foreground and mask travel in the portable package, and restoration remaps them correctly. Derived human footage cannot claim to be unaltered proof.

Actual cancellation stopped the inference child, retained incomplete work without a success receipt and released the workspace lock. A deadline also propagated safely. Successful scratch frames are reclaimed; failed intermediates remain for diagnosis. The source was an animated photograph, not natural talking-head footage: moving hair, hands, blur, occlusion and difficult backgrounds still need representative visual review.

09 | Reliability, recovery and verification

Every export has a unique destination. Previous successful films, earlier revisions and failed runs remain intact. The last-success pointer advances only after verification; a partial narration file, picture clip or zero-exit subprocess is insufficient. Controlled job states and one-renderer-per-workspace locking bound conflicting local work.

Cancellation and timeouts propagate to active child processes. Durable jobs can be recovered as interrupted after restart. Resume creates a new export and reuses only compatible completed shots whose recorded recipe still matches. This avoids silently mixing an old visual with changed source, layout, labels, fonts, renderer configuration or model assets.

Portable restoration validates the package root, source containment and hashes before importing assets into the current workspace. A regression rejects an Inputs directory redirected outside the selected package. Original sources for derived matting are included in render recipes and copied alongside foreground/mask; restoration preserves that lineage.

The recorded installed suite contains67 legacy checks,5 release-boundary checks,17 core checks and6 matting checks. Coverage includes contracts, finite values, ownership/revision conflicts, path boundaries, real media composition, caption safe areas, source-note visibility, recovery, model integrity, quota separation and restoration. Configured model-dependent paths ran locally; a clean CI environment reports skips when optional assets are absent.

Security/type/lint tools complement these behavior checks. The configured hook runs secret scanning and Python analysis; Semgrep found no issue in its recorded scope. The public workflow adds standard-runner checks with read-only source permissions and no paid inference. A scanner pass is not proof of blanket security or a substitute for real media/user-flow tests.

Historical preservation was independently checked across1,641 files totaling approximately3.79 GiB before publication work. Source and model runtime parity were recorded separately. The publication changes update guides, notices, automation and the legacy process/download boundaries; historical films remain excluded from the repository and unchanged. New changes should repeat the affected checks and receive a fresh review before release.

10 | Security, privacy, rights and AI boundaries

This release is local workstation software, not a public multi-tenant application. Its source and configuration schemas do not establish hosted authentication, isolation or compliance. The local store protects owner-scoped references, but a public service would need authenticated per-request authorization, upload validation/abuse controls, tenant-aware storage, operational monitoring and deployment hardening.

Inspect screen recordings for credentials, cookies, customer information and private browser state before import. Do not film login secrets or copy another project's recovery data into this workspace. Local SQLite/media are not automatically encrypted: use suitable operating-system permissions, disk encryption and backups. Define retention/deletion for original footage, derived masks, transcripts, captions and exports.

The optional director is identified as AI assistance. It accepts a bounded brief, validates structured output, uses only an explicitly configured loopback model and cannot invent source assets or verified proof. It has no publishing, external messaging, purchasing or general shell-action tools. Its output still needs factual review; no unsupported answer or completed action should be invented.

Original project code is published under GPL-3.0. Dependencies, optional models, fonts and footage retain their own terms. Kokoro model materials identify Apache 2.0 and its inference package uses MIT upstream terms; the optional RVM repository identifies GPL-3.0 for code, while particular model redistribution terms require separate checking. No model weights, workstation fonts or proprietary editor binaries are bundled.

Before processing real customer data or hosting a service, identify controller/processor roles, user/data locations, vendors, processing regions and applicable law. Assess EU GDPR where its scope applies and relevant US federal/state privacy, copyright and consumer requirements. Children's or sensitive/regulated material needs additional analysis. This is a review plan, not legal advice or a compliance certification.

Maintain privacy notices, lawful processing grounds, access/erasure procedures, retention rules, vendor agreements and breach response where applicable. Reference OWASP guidance for software/upload/AI threats and regulator guidance for jurisdictional duties. Do not assume a scanner, license file, model disclaimer or signed agreement alone establishes compliance.

11 | Future scale, availability and cost roadmap

The current capacity unit is a local production workspace with one active renderer. No public registered-user, daily-active-user, request-per-second or concurrent-render workload has been measured. The two-hour project envelope is a validation ceiling, not evidence that arbitrary two-hour productions are performant. The2-second matting measurement must not be extrapolated into a million-user claim.

A future hosted design should distinguish1 million registered users,1 million daily viewers and1 million simultaneous active sessions. If1 million simultaneous sessions made one metadata request every30 seconds, the modeled demand would be about33,333 requests/second before bursts. That is an explicit assumption for sizing an API tier; it says nothing about render-worker capacity.

For render demand, define submission rate, average duration, resolution, effects, CPU/GPU profile, output size, retention and peak burst duration. Worker concurrency follows measured service time, not user count alone. Separate lightweight browsing/metadata traffic from expensive jobs; use admission limits, per-project quotas, bounded queues, idempotency, backpressure, cancellation and failure recovery.

A measured expansion could begin with stateless API instances, durable metadata storage, project-scoped object storage and independent rendering workers. Apply tenant authorization before every read/write. Cache only suitable public content; private media must not leak through caches or logs. Add replicas, regional redundancy or orchestration only when reliability/workload evidence justifies their cost and complexity.

A future 99% monthly availability objective would permit approximately 7 hours 12 minutes of unavailable time in a 30-day month. It is a proposed SLO, not achieved uptime or an SLA. Define successful-request and completed-job indicators separately, plus latency/error targets, acceptable data loss, recovery time, alert ownership and error-budget response. Test failure, restore and rollback before making service commitments.

Budget all inference, compute, storage, egress, database, logs, backups, queues and security tooling. Free allowances and temporary credits do not prove sustainable zero cost; alerts are not spending caps. No cloud provisioning, paid inference, billing activation or automatic paid scaling is part of this release. Reprice the selected workload and obtain explicit informed approval before any chargeable step.

12 | History, priorities and maintenance

The earlier stages remain part of the project's history. October 12 established a reusable local narration/caption/export workflow. October 15 added flexible short/long formats, configurable motion, non-musical sound and licensed stock-media handling. October 17 expanded original editorial motion boards, with additional source/label/cache/watchdog checks. Those earlier workflows remain retained rather than being represented as features invented by the latest rebuild.

October 21 introduced the v2 layer/shot contracts, recorded-speech preparation, optional local director, composed style/animation checkpoints, local reference measurements, Blender adapter, portable source restoration and bounded recovery. Subsequent frame review corrected portrait caption sizing and source-credit occlusion. The focused human-matting upgrade then added local recurrent extraction, source lineage and verified text depth.

The next creative priorities are stronger reference grammar analysis, more distinct illustration/animation authoring, autonomous shot production with an independent reviewer role and verified native editor round trips. Do not describe the current seven presets or developer fresh review as a completed runtime multi-agent production system. Natural presenter footage needs its own quality review.

Frontend/backend hosting is a separate future release. It requires identity/tenant/upload controls, privacy/vendor review, capacity measurements, owned monitoring and recovery evidence. Source publication and the technical document do not establish a deployed service. Preserve the local workflow while adding these capabilities in focused, reviewable increments.

For maintenance, keep dependency/tool/model provenance current, re-run affected behavior/security gates, inspect actual final frames and full audio, and record escaped defects with their missing gate. Keep a last-known-good editing package and test restoration. The operator remains responsible for claims, rights, private content and final editorial judgment.

Primary references: FFmpeg filters and legal guidance; Blender command-line rendering and license; Kokoro model/voice inventory and kokoro-onnx; RobustVideoMatting inference documentation; OBS websocket documentation; OWASP Cheat Sheet Series; EDPB small-business privacy guidance. These sources inform implementation or future assessment; they do not certify this application. Repository checks and per-export evidence describe the actually executed scope.

Official source links
FFmpeg filters: https://ffmpeg.org/ffmpeg-filters.html
FFmpeg legal: https://ffmpeg.org/legal.html
Blender rendering: https://docs.blender.org/manual/en/latest/advanced/command_line/render.html
Blender license: https://www.blender.org/about/license/
Kokoro model: https://huggingface.co/hexgrad/Kokoro-82M
kokoro-onnx: https://github.com/thewh1teagle/kokoro-onnx
RVM inference:
https://github.com/PeterL1n/RobustVideoMatting/blob/master/documentation/inference.md
OBS control: https://github.com/obsproject/obs-websocket
OWASP: https://cheatsheetseries.owasp.org/
EDPB: https://www.edpb.europa.eu/sme_en