

TL;DR “UMO stills” refers to a production pattern for multi-identity consistency: curate identity-clean stills with a strong image base (e.g., an OmniGen2-class generator), index them in an identity bank, then use retrieval-guided conditioning (plus optional LoRA adapters) to keep subjects consistent across long or multi-scene videos.

What problem does it solve?

Long or multi-scene videos with several recurring subjects (actors, hosts, avatars) often drift in face details, outfits, or accessories. Typical video models can maintain a single identity in a short clip, but multi-identity and long-form projects need stronger anchors and repeatable retrieval.

“UMO stills” addresses this by:

- Creating identity-clean, high-SNR still frames for each subject
- Indexing those stills with robust embeddings for retrieval
- Feeding anchor imagery and retrieval hints back into the video pipeline

This elevates identity stability while keeping creative control (pose, camera motion, scene swaps) intact.

Core components

1. OmniGen2-class image base (stills)

- Use a modern text-image generator (OmniGen2-class) to produce or refine identity-clean stills (front/³/₄ profile, neutral expression, key outfits).
- Enforce quality gates: resolution ≥ 1024 px, sharpness, exposure, and minimal motion blur; crop to consistent head/torso framing.

2. Identity bank (embeddings + metadata)

- Compute embeddings with a face/ID model and a general visual encoder (CLIP-family). Store per-identity vectors plus rich tags (hair, outfit, glasses, accessories).
- Deduplicate with cosine thresholding; maintain a curated “gold” set.

3. Retrieval-guided conditioning (video)

- At generation time, query the bank by prompt + rough frame description to fetch the nearest anchor still(s).
- Condition the video model with anchor crops (concat channels, reference frames, or adapter inputs) and prompt constraints.
- Optionally blend LoRA adapters per identity/outfit for stronger lock-in.

4. Consistency checks and feedback

- During generation, run face/ID similarity on sampled frames. If drift $> \tau$, nudge guidance (increase identity weight, swap anchor, or re-seed).
 - For long-takes, insert “refresh” keyframes (UMO stills) at scene boundaries.
-

Suggested pipeline (high level)

1. Curate stills

- Generate/refine 6–12 stills per identity with the image base (clean backgrounds, neutral to expressive variants).
- Normalize framing and lighting; run ESR/face-restore only if truly needed.

2. Index

- Embed: face ID model + CLIP-family encoder
- Store: vectors + tags + source prompt + crop boxes
- Prune: cosine dedup; flag near-duplicates and low-SNR samples

3. Generate video

- Per shot: retrieve top-k anchors per identity matching the shot prompt
- Condition the video model with anchors (reference frames/controls)
- If available, attach per-identity LoRA (light-weight, composable)

4. Monitor & correct

- Sample frames every N steps; compute similarity vs. bank
 - If drift: raise identity guidance, swap anchor, or short backtrack
-

Practical tips

- Stills coverage: capture hair up/down, accessory on/off, and key outfit variants; avoid training the bank on artifacts you don't want reproduced.
 - Multi-identity scenes: stagger identity anchors (A-first frames, B-later), or use per-identity regions for stronger separation.
 - Retrieval hygiene: keep bank balanced (no identity overwhelms with hundreds of near-duplicates). Curate, don't just crawl.
 - LoRA adapters: lightweight per-identity LoRA can stabilize tough cases; freeze if the base model already handles the look.
 - Long-form projects: put anchor refresh points at scene cuts or timeboxes.
-

Where this pattern fits

- Talk shows, vlogs with recurring hosts/guests
 - Multi-character shorts with outfit continuity
 - Branded spokespersons used across episodes
 - Avatar pipelines combining pose drivers with identity control
-

Notes and sourcing

This post describes an industry pattern often nicknamed “UMO stills”. Specific implementations vary across teams and products. The “OmniGen2-class” mention is generic: use a strong contemporary image generator for identity-clean stills, then integrate anchors with the video model of choice. Always validate results on your prompts, assets, and hardware.

References

- [UMO \(GitHub\)](#).
- [UMO \(arXiv\)](#).
- [AnimateDiff \(GitHub\)](#).
- [AnimateDiff \(arXiv\)](#).