THE GENERATOR IS THE TRACKER: MULTI-OBJECT TRACKING BY PAINTING PERSISTENT IDENTITY COLOURS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-object tracking (MOT) is conventionally decomposed into detection followed by association, with object identity maintained as external state: track buffers, motion models, appearance embeddings. We ask whether a video *generator* can maintain that state *in pixels*. We fine-tune a 22B text-to-video diffusion model (LTX-2.3) with a lightweight in-context LoRA to translate an RGB clip into an *ID-map* clip, a video in which every person is painted a flat, distinct colour that persists over time (same colour, same identity). Long videos are generated as chained windows, each conditioned on the cleaned tail of the previous one; a brief *continuation* fine-tune teaches the model to extend a given colouring, after which identity flows through the chain with no separately trained detector, no motion model, and no association module. On the DanceTrack test server the system reaches **40.3 HOTA**, to our knowledge the first entry whose tracks are read directly off video produced by a pretrained video generator. This is well below today's specialists (66–76 HOTA), but with an error profile no baseline shares: its association score (**AssA 44.1**) exceeds every tracker of the 2020–22 benchmark suite while detection is the deficit. Controlled comparisons on the validation split show where the identity comes from: windows generated without generation-time state and linked by post-hoc association score 18.2 HOTA against 34.8 for the chain, while re-linking the chain's own windows post hoc changes nothing (34.7), so the colours already carry what overlap association can recover, and motion- and appearance-based associators run on the system's own boxes reach only 22–24 AssA, an ordering that holds again on SportsMOT validation. On 383 mined occlusion events the chained generator re-acquires 42% of the identities it held on both sides of the gap, including gaps longer than its 49-frame context, by keeping the colour alive in pixels through the gap. We release code, checkpoints, and the full pinned experiment log.

1 INTRODUCTION

The dominant paradigm in multi-object tracking is *tracking-by-detection*: a detector proposes boxes per frame, and an association stage stitches them into tracks using motion models and appearance embeddings (Bewley et al., 2016; Zhang et al., 2022). Even end-to-end transformer trackers (Zeng et al., 2022) retain the same ontology: identity is *external state*, carried in query slots or track buffers alongside the video.

Video generative models suggest a different ontology. To synthesize a coherent video, a generator must implicitly solve the binding problem: it must know which pixels belong to the same object across frames, through occlusion, deformation, and camera motion. Such correspondence is known to be present in diffusion features (Tang et al., 2023; Nam et al., 2025). We ask whether it can be read out *as generation*: we ask a video diffusion model to emit, for an input RGB clip, a video of *ID-maps*, every tracked person painted a flat, distinct colour on black, with the constraint that *the same person keeps the same colour over time* (Figure 1). Identity is not a slot, an embedding, or a learned buffer; it is a property of the generated pixels, and decoding tracks requires only nearest-palette colour quantisation. Our contributions are the mechanism and its controlled evaluation:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

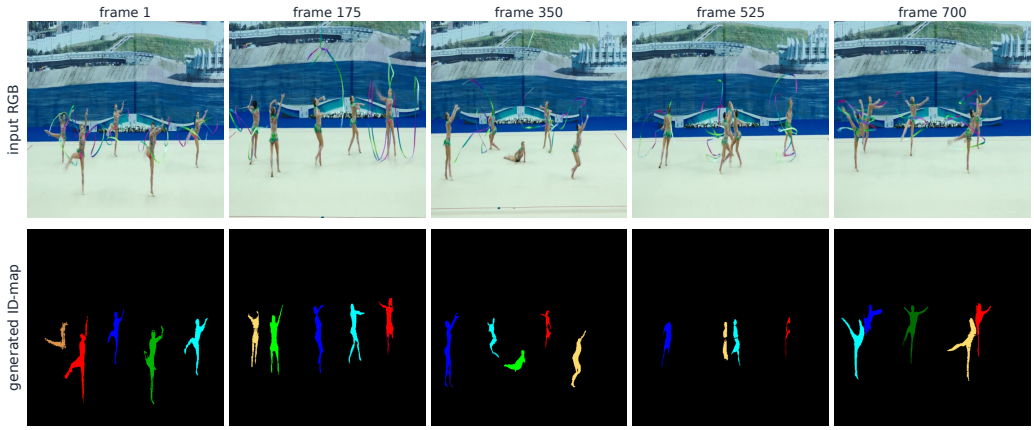


Figure 1: **Tracking by painting.** Input RGB (top) and our generated ID-map video (bottom) for a full 702-frame DanceTrack validation sequence (dancetrack0065, our development video). Identity is carried entirely by the generated colours across 18 chained generation windows. The five principal colours persist over the sequence (one dancer is missed at frame 525); over the full 702 frames the chain scores IDF1 0.47 with 30 predicted track ids for 5 people, i.e. persistence with short-lived fragments, not perfect identity.

- **Identity-as-colour video translation.** We formalise tracking as RGB→ID-map video translation over a fixed $K{=}48$ CIELAB-separated palette with a lossless codec (round-trip mask IoU $\geq$ 0.99), and fine-tune LTX-2.3 with a rank-32 in-context LoRA (positionally aligned reference conditioning) on 40 DanceTrack training videos with silhouette supervision from SAM 3.

- **Continuation training for unbounded video length.** Windows of 49 frames are chained by conditioning each window's first 9 frames on the *cleaned* (decoded and re-encoded) tail of the previous window. Overlap-based chaining suffices for per-frame targets such as depth (Hu et al., 2025); for identity it fails at inference time, because the model fights an imposed colouring it was never trained to continue. A brief teacher-forced *prefix* fine-tune teaches "continue this colouring" and raises full-video IDF1 on the development video from 0.17 to 0.47.

- **A generative tracker on the official evaluation server.** A single frozen submission achieves **HOTA 40.3 / AssA 44.1 / IDF1 45.2**. We do not claim state of the art: 2025–26 specialists reach 66–76 HOTA (Table 1). The result is the error profile: association above every entry of the 2020–22 benchmark suite (maximum 40.2, MOTR) while detection (DetA 37.6) is the deficit, on the benchmark built to break association; Section 4.1 states what this comparison does and does not show.

- **Where the identity comes from.** On the 25 full-length validation videos with identical window geometry, windows generated *without* generation-time state and linked by post-hoc association score 18.2 HOTA against 34.8 for the chain, whereas re-linking the chain's own windows with the same associator changes nothing (34.7), and classical associators run on the chain's own boxes reach AssA 21.6 (motion) and 24.3 (appearance and motion) against 38.3 for the colours. Generation-time state, not post-hoc association, carries identity.

- **Re-acquisition through and beyond the context window.** On 383 mined occlusion and exit events, the chained generator re-binds the correct identity for 42% of the 205 events it tracked on both sides of the gap (22% of all 383); a Kalman/IoU associator and an appearance-bank re-identifier on the same boxes and events reach 23% and 24%. Gaps longer than the 49-frame context still re-acquire at 27% (7 of 26), and controls show how: in every re-acquired event the colour was painted somewhere during the gap, so the generator keeps identity alive in pixels rather than re-deriving it.

We report a failure axis as a finding: in dense scenes (>14 people) the model saturates at 10–12 painted instances per frame at 512^2, a latent-token-capacity limit that resolution-matched training would have to lift. Throughput is not a claim; it is reported in Section 5.

2 Related Work

Multi-object tracking. Tracking-by-detection pairs a detector with association by motion (Bewley et al., 2016), appearance (Wojke et al., 2017), or both at scale (Zhang et al., 2022; 2021; Zhou et al., 2020). Transformer trackers make association end-to-end but keep identity in query state (Zeng et al., 2022; Sun et al., 2020; Gao et al., 2025; Segu et al., 2025; Shao et al., 2026). Diffusion has entered this stack as a box or motion denoiser inside tracking-by-detection (Luo et al., 2024; Lv et al., 2024); the generated object is a box, not the video. TrackDiffusion (Li et al., 2025) conditions video generation on tracklets; we invert that mapping. DanceTrack (Sun et al., 2022) was designed to defeat appearance cues (uniform costumes, chaotic motion) and exposes association as the weak point of this stack: strong detectors reach DetA 70–80 while AssA collapses to 20–40, which makes it the natural testbed for a method whose only identity mechanism is the generator itself.

Segmentation and foundation-model tracking. SAM 2 (Ravi et al., 2025) and SAM 3 (Carion et al., 2026) track prompted objects with a per-object memory; SAM2MOT (Jiang et al., 2026) and SAMURAI++ (Yang et al., 2026) turn this into multi-object trackers by seeding SAM 2 from a detector and propagating one memory per object, reaching 72–76 HOTA on DanceTrack without MOT training. These are the closest neighbours to our system: identity also lives inside a foundation model rather than in a tracker. They still need a box detector to instantiate objects and propagate objects one at a time; we instantiate and bind all people jointly in one generated clip.

Perception as generation. Painter (Wang et al., 2023a) and SegGPT (Wang et al., 2023b) cast perception as in-context image painting, SegGPT including video object segmentation with masks painted as colours; Pix2Seq-D (Chen et al., 2023) generates panoptic masks with a diffusion model and, in video, conditions on past predictions so that instance ids persist frame to frame, the ancestor of our decode-and-recondition chain. Tracking as colour propagation was observed in self-supervised colourisation (Vondrick et al., 2018). Diffusion image generators have been repurposed for dense prediction by light fine-tuning (Ke et al., 2024), and video diffusion models for video depth with segment-wise inference and stitching (Hu et al., 2025; Chen et al., 2025). Most directly, Gabeur et al. (2026) argue that image-generation pretraining plays the role LLM pretraining plays for language, matching or surpassing specialists such as SAM 3 on several segmentation benchmarks by instruction tuning an image generator, and name video input as future work. Relative to this line we contribute (i) a pretrained 22B video generator rather than a task model, (ii) whole-window generation with three-dimensional attention so occlusions are bridged inside the context, (iii) explicit continuation training for unbounded chaining, and (iv) evaluation as a tracker on an MOT test server.

Correspondence in generative features. DIFT (Tang et al., 2023) showed that correspondence emerges in image diffusion features, and DiffTrack (Nam et al., 2025) that video diffusion transformers encode temporal correspondence usable for zero-shot point tracking. Our results extend this evidence to instance identity through full occlusion, read out as generated pixels and measured on a tracking benchmark.

Backbone. We build on latent video diffusion (Rombach et al., 2022; Ho et al., 2022) and specifically LTX-2 (HaCohen et al., 2026; 2024), whose in-context LoRA (Huang et al., 2024) provides positionally aligned video-to-video conditioning and whose timestep-based latent conditioning supports our prefix training. Flow matching (Lipman et al., 2023) and LoRA (Hu et al., 2022) keep the fine-tune below 1% of parameters.

3 Method

3.1 Identity as colour: the ID-map codec

An *ID-map* for frame t paints instance i's silhouette with palette colour $c_{a(i)}$ on black, where a is a per-video assignment into a fixed palette of $K{=}48$ colours chosen to maximise pairwise CIELAB separation (minimum $\Delta E = 26.4$). The assignment is *sequence-stable*: one colour per identity for

the whole video. In training data it maps a sequence's track ids to palette indices in sorted order; at inference the generator chooses colours freely. Decoding inverts this by nearest-palette quantisation in CIELAB followed by per-colour connected components with a minimum-area filter (Appendix A); boxes are component bounding boxes and the palette index is the track id. The codec is frozen infrastructure: encode–decode round-trips preserve masks at IoU ≥ 0.99 with 100% identity agreement, and ID-maps are stored only in lossless formats (PNG, FFV1), since lossy compression smears flat colours and silently corrupts decoding.

3.2 Video-to-video fine-tuning with in-context LoRA

We fine-tune LTX-2.3 (HaCohen et al., 2026), a 22B-parameter joint audio-video latent diffusion transformer (48 dual-stream blocks; causal video VAE with $32\times$ spatial and $8\times$ temporal compression), using its native *in-context LoRA* mechanism: the VAE-encoded reference video (RGB) is concatenated to the target token sequence with *aligned* positional encodings, kept clean (timestep 0), and excluded from the loss, so bidirectional self-attention learns a positionally locked translation from reference to target. We adapt only the video-stream projections (attention QKV and output, and the feed-forward layers; rank 32; 164M trainable parameters, 0.74% of the model) with a flow-matching objective. Text conditioning is a fixed prompt whose embedding is precomputed, which drops the 12B Gemma-3 text encoder from training memory: the entire fine-tune fits on a single 80GB A100.

Training data are 777 clips of 49 frames at 512^2 from the 40 DanceTrack training videos, with silhouette supervision produced once by box-prompted SAM 3 (Carion et al., 2026) on ground-truth boxes (97% of boxes matched back by IoU, filled-box fallback otherwise), and occlusion-spanning clips oversampled. The validation split (25 videos) is never used for training; it is used for checkpoint selection and method development (Section 4).

3.3 Continuation training: teaching the model to extend a colouring

A 49-frame window is far shorter than a video, and generating longer clips directly degrades identity (Section 4.5); full videos therefore require *chained* windows with 9-frame overlaps (two latent frames). The chain conditions each window on the previous one through the model's native prefix conditioning, which fixes the first latent frames as clean tokens. The prefix is the previous window's cleaned tail: its last 9 frames are decoded to palette colours, speckle-filtered, and re-encoded, so the condition is always a valid ID-map rather than raw generation. We call this the decode, correct, re-encode loop.

Applied at inference time only, this fails in an instructive way. The anchor binds the overlap frames, but beyond them the model reverts to its own appearance-derived colours: it was trained to *invent* colourings, never to *continue* one, and the errors compound down the chain. We therefore resume training with an additional teacher-forced prefix condition: with probability 0.5, the target's own first two latent frames (9 pixel frames) are given clean and excluded from the loss, teaching the skill "propagate the colouring you were handed." On our development video (dancetrack0065, a 5-person, 702-frame validation sequence used for all single-video diagnostics) this single change lifts full-video IDF1 from 0.172 to 0.466, while the $p{=}0.5$ mixture preserves single-window mask quality for the chain's first window (foreground IoU 0.589 vs. 0.605) at a cost in unconditioned identity quality (Section 4.5).

3.4 Inference: the full-video chain

Figure 2 summarises the pipeline. Given a test video, we (i) split it into 49-frame windows at stride 40; (ii) generate window k conditioned on [reference RGB_k; prefix = cleaned tail of window $k{-}1$] with 20 denoising steps, classifier-free guidance 4.0 and spatio-temporal guidance 1.0, using the fixed prompt "segmentation ID map: paint each person a distinct flat colour on a black background"; (iii) decode, correct, re-encode; (iv) emit boxes from the frames each window owns (overlaps split at their midpoint), with same-colour components in a frame merged into one box. The colour index *is* the global track id: no separately trained detector and no association module exist in the system; identity state lives in the generated pixels and is re-injected as the prefix. New entrants receive whatever colour the model paints; nothing prevents the reuse of a colour that belonged to someone

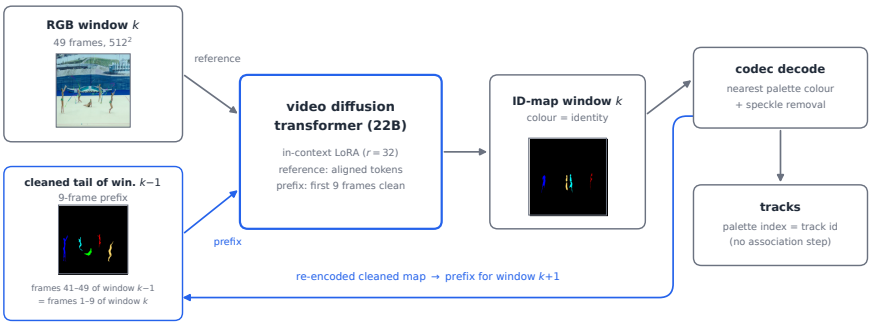


Figure 2: **System overview.** Each 49-frame window is generated by a video diffusion transformer conditioned on (i) the positionally aligned RGB reference (in-context LoRA) and (ii) a 9-frame *prefix*: the decoded, corrected (nearest palette colour, speckle removal), and re-encoded tail of the previous window. The generated ID-map is decoded by nearest-palette quantisation; the palette index *is* the track id, and the cleaned map seeds the next window's prefix.

who left, and such reuse surfaces as a merged box. The palette bounds cumulative identities at $K-1 = 47$ per video (Appendix B).

4 Experiments

Setup. Our primary benchmark is DanceTrack (Sun et al., 2022): 40 train / 25 val / 35 test videos (700–2,400 frames each) of group dance with near-identical costumes and chaotic motion; Section 4.4 repeats the recipe on SportsMOT (Cui et al., 2023). Metrics are HOTA/DetA/AssA and their components (Luiten et al., 2021), IDF1, and MOTA, computed with TrackEval (Appendix A). Test numbers come from the official evaluation server (single submission, method frozen on validation, no test-side iteration). The validation split is our model-selection and development set: checkpoints were selected by validation foreground IoU, and the associator, decode thresholds, and the continuation-training decision were developed on dancetrack0065, one of the 25 videos in Tables 2–3; the test row is the only selection-free number. "Pinned" means fixed in a versioned, released log before the corresponding run; any later change started a new labelled run. All reported numbers use sampling seed 0; a three-seed check on five videos is in Section 5. Total compute is about 154 A100-hours (Appendix G).

4.1 DanceTrack test-server results

Table 1 and Figure 3 show the result. This is a paradigm result, not a state-of-the-art result: 2025–26 specialists reach 66–76 HOTA, and our 40.3 sits with the 2020–21 detector-based systems. What no other entry offers is the profile. Ours is, to our knowledge, the only submission that runs neither a box detector nor an association or propagation module, and the only one whose association exceeds its detection. AssA 44.1 is above every tracker of the 2020–22 suite, including the end-to-end transformer MOTR (40.2) and the motion specialist OC-SORT (38.3 in the benchmark table; 40.2 in its own paper), while DetA (37.6 against 80 for specialists) is the entire deficit. Modern systems reach higher association (52.9–72.2) through stronger association machinery: adaptive re-identification (Maggiolino et al., 2023), non-linear motion models (Lv et al., 2024), long-term query memory (Gao & Wang, 2023), detector proposals (Zhang et al., 2023), or a per-object foundation-model memory (Jiang et al., 2026); their AssA/DetA ratio stays below one (SAM2MOT is closest at 0.91), whereas ours is 1.17.

What the profile does and does not show. AssA is computed only over matched detections, so a tracker that detects the easy people and keeps their tracks whole scores high AssA at low DetA. Our recall is low (validation DetRe 41.4 / DetPr 51.3; LocA 77.1, so localisation of what is found is not the problem) and lowest in crowds, where association also falls (Figure 4, Appendix B). AssA above the 2020–22 suite is therefore a statement about association on the people the system finds, which over-represents sparse scenes; the like-for-like test is association on *identical* detections, where

Table 1: DanceTrack **test** results (official server). Baseline values are from the DanceTrack benchmark table as updated by the benchmark authors (arXiv v3 of Sun et al., 2022, Table 3, which re-ran QDTrack with YOLOX and added OC-SORT after publication; reproduced in Zhang et al., 2023) and from each cited method paper. Det.: sep. = separately trained box detector; e2e = detection queries trained jointly; SAM 2 = detector-seeded SAM 2 memory. TrackTrack does not report DetA.

	Method	Year	Det.	HOTA↑	DetA↑	AssA↑	IDF1↑	MOTA↑
2020–22 suite	CenterTrack (Zhou et al., 2020)	2020	sep.	41.8	78.1	22.6	35.7	86.8
	TransTrack (Sun et al., 2020)	2020	e2e	45.5	75.9	27.5	45.2	88.4
	FairMOT (Zhang et al., 2021)	2021	sep.	39.7	66.7	23.8	40.8	82.2
	TraDeS (Wu et al., 2021)	2021	sep.	43.3	74.5	25.4	41.2	86.2
	QDTrack (YOLOX re-run)	2021	sep.	54.2	80.1	36.8	50.4	87.7
	ByteTrack (Zhang et al., 2022)	2022	sep.	47.7	71.0	32.1	53.9	89.6
	GTR (Zhou et al., 2022)	2022	sep.	48.0	72.5	31.9	50.3	84.7
	MOTR (Zeng et al., 2022)	2022	e2e	54.2	73.5	40.2	51.5	79.7
	OC-SORT (Cao et al., 2023)	2023	sep.	55.1	80.3	38.3	54.6	92.0
2023–24	Deep OC-SORT (Maggiolino et al., 2023)	2023	sep.	61.3	82.2	45.8	61.5	92.3
	MeMOTR (Gao & Wang, 2023)	2023	e2e	68.5	80.5	58.4	71.2	89.9
	MOTRv2 (Zhang et al., 2023)	2023	sep.+e2e	69.9	83.0	59.0	71.7	91.9
	DiffMOT (Lv et al., 2024)	2024	sep.	62.3	82.5	47.2	63.0	92.8
2025–26	TrackTrack (Shim et al., 2025)	2025	sep.	66.5	–	52.9	67.8	93.6
	MOTIP (Gao et al., 2025)	2025	e2e	69.6	80.4	60.4	74.7	90.6
	FDTA (Shao et al., 2026)	2026	e2e	71.7	81.0	63.5	77.2	91.3
	SAMURAI++ (Yang et al., 2026)	2026	sep.+SAM 2	72.8	82.2	64.7	77.7	92.1
	SAM2MOT (Jiang et al., 2026)	2026	sep.+SAM 2	75.8	79.7	72.2	83.9	88.5
	Ours (generation only)	2026	none	40.3	37.6	**44.1**	45.2	16.2

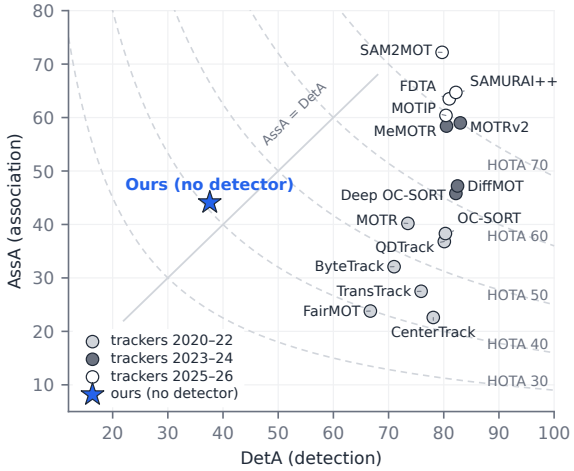


Figure 3: **The generative tracker inverts the classical error profile.** DanceTrack test decomposition: every detector-based tracker sits below the AssA = DetA diagonal (detection ahead of association), while ours (star) is the only method above it and the only one without a detector. Modern specialists (upper right) exceed our absolute scores on both axes. Dashed curves are iso-HOTA. Values as in Table 1.

classical associators on our own boxes fall well short of the colours (Section 4.2). Test scores exceed validation scores (34.8 HOTA / 38.3 IDF1) with a single frozen submission; the 35 test videos are on average sparser than the 25 validation videos, as the density dependence predicts.

4.2 Where does the identity come from? Controlled comparisons

Identity could, in principle, be recovered from generated windows by classical means. We test this on the 25 full-length validation videos with the *same window geometry* (640 windows per arm; Table 2). (**A**) The chain of Section 3.4, colour as identity. (**B**) Windows generated *independently* (no

Table 2: Identity mechanisms on identical window geometry: 25 full-length validation videos, TrackEval. Post-hoc association of windows generated without generation-time state (B) halves HOTA relative to the chain (A); the same associator applied to the chain's own windows (A+assoc.) changes nothing.

Identity mechanism	HOTA↑	DetA↑	AssA↑	DetRe	DetPr	AssRe	AssPr	IDF1↑
A: chain, colour = identity (ours)	**34.8**	32.3	**38.3**	41.4	51.3	46.5	55.8	**38.3**
A+assoc.: chain windows re-linked	34.7	32.5	38.0	41.2	52.1	45.1	59.9	37.7
A boxes, Kalman/IoU association	25.9	32.7	21.6	41.8	51.7	23.7	67.2	25.1
A boxes, appearance+motion bank	27.5	32.5	24.3	41.5	51.5	27.9	53.6	27.6
B: independent windows + assoc.	18.2	19.7	17.7	25.0	40.5	20.9	51.5	17.3

prefix) by the translation-only checkpoint of Section 3.2 and linked post hoc by a pinned overlap-IoU associator (Appendix D); arm B uses that checkpoint because the prefix mixture lowers unconditioned identity quality (Section 4.5), so the arms differ in checkpoint and generation mode as well as in the identity mechanism. (**A+assoc.**) The chain's own windows re-linked by the same associator, the genuinely same-windows control. Finally, we drop the ids from arm A's per-frame boxes and re-associate them with a SORT-style Kalman/IoU tracker (Bewley et al., 2016) and with an appearance-and-motion identity bank over DINOv3 embeddings (Appendix A), which compares association on identical detections.

Arm A leads arm B by about $2\times$ on every identity metric. The post-hoc associator, which tied arm A on the 5-person development video, collapses on dense scenes, where many-to-one overlap linking over-merges crowded, identically dressed people into union boxes; that collapse also costs detection (DetA 19.7), so the B row measures the whole consequence of generating without state, not association alone. Applied to the chain's own windows the same associator is neutral (-0.1 HOTA, -0.4 AssA), so the colours already carry everything overlap linking can recover, and generation-time state (the prefix) is what separates A from B. On identical detections (DetA 32.3–32.7), the Kalman/IoU associator reaches AssA 21.6 and the appearance-and-motion bank 24.3 against 38.3 for the colours, with 3,079 and 2,408 identity switches against 1,950: fast dance motion breaks frame-to-frame geometry and near-identical costumes defeat appearance, exactly where the generated colour stays stable. A stronger post-hoc method might narrow the A–B gap; it cannot change the finding that the pixels alone already exceed classical association on the same boxes.

4.3 Long-gap re-acquisition

The hardest identity problem is re-acquisition: a person occluded or out of frame for dozens of frames must return with the same identity. We mined all ground-truth gap events ($\geq$9 absent frames, then reappearance) across the 25 validation videos: 383 events, 360 occlusions and 23 exits (protocol in Appendix C). For each event we ask whether the person was tracked before the gap and re-detected after it and, if so, whether the majority identity matched. Table 3 reports the conditional rate k/n over such events with its Wilson 95% interval and the unconditional rate over all 383, because the conditional denominator differs between systems (205 events for the chain, 154 for the independent-window arm), together with the two classical associators run on the chain's own boxes over the same events.

Four observations follow (a qualitative example is in Figure 5, Appendix C). First, for gaps inside the temporal context (9–48 frames), where attention can see through the occlusion, the chained model re-binds 47% and 38% of the events it holds, the capability that motivated using a video model. Second, on identical boxes the colours beat both classical associators: 86 against 47 (Kalman/IoU) and 50 (appearance bank) of the same 205 events; on the paired events the colours win 41 to 5 and 40 to 6 (exact McNemar $p < 0.001$). Third, the chain leads the independent-window arm (22% against 14% of all 383 events) but the paired difference is not significant: on the 115 events both hold, the chain wins 20 and the independent arm 10 ($p = 0.10$). Fourth, gaps longer than the context (49–150 frames) still re-acquire at 27% (7 of 26), and two controls say why. Across *independently* generated windows the raw palette colour agrees before and after the gap in 12% of events (31 of 255) against a chance level of 5% (one over the number of colours in use), a weak colour preference rather than re-identification; in the chain the pre-gap colour was painted in at least one gap frame for all 86 re-acquired events, including all 7 beyond the context. Re-acquisition is therefore carry-over:

Table 3: Re-acquisition after ground-truth gaps (383 events, validation set). Conditional k/n counts events tracked before and re-detected after the gap on the same boxes that TrackEval scores; the $>$150 bin is included in "all". The 49-frame window is the model's temporal context. The two associators run on arm A's boxes; "raw colour" reads arm B's independently generated windows without any linking.

Gap length (frames)	9–24	25–48	49–150	>150	all
GT gap events	234	88	53	8	383
A: chain (ours), k/n	60/127	19/50	7/26	0/2	86/205
rate (95% CI)	0.47 (0.39–0.56)	0.38 (0.26–0.52)	0.27 (0.14–0.46)	0/2	0.42 (0.35–0.49)
unconditional	0.26	0.22	0.13	0.00	0.22
A boxes, Kalman/IoU, k/n	33/127	12/50	2/26	0/2	47/205
A boxes, appearance+motion bank, k/n	39/127	10/50	1/26	0/2	50/205
B: independent + assoc., k/n	40/96	10/38	3/19	0/1	53/154
rate (95% CI)	0.42 (0.32–0.52)	0.26 (0.15–0.42)	0.16 (0.06–0.38)	0/1	0.34 (0.27–0.42)
unconditional	0.17	0.11	0.06	0.00	0.14
B raw colour (no linking), k/n	21/161	5/64	5/29	0/1	31/255

Table 4: SportsMOT validation with the DanceTrack recipe re-run from the stock backbone, TrackEval. Left: all 45 validation videos (combined). Right: unweighted mean over the 37 videos that took no part in checkpoint selection (a per-video mean, not a combined score). Re-acq.: conditional re-acquisitions k/n over all 45 videos under the protocol of Section 4.3. The zero-shot rows apply the DanceTrack checkpoint with no SportsMOT training to the 8 selection videos only.

	all val videos (TrackEval combined)				held-out mean				
association	HOTA	DetA	AssA	IDF1	HOTA	DetA	AssA	IDF1	re-acq
Zero-shot (DanceTrack checkpoint) (8 selection videos)									
generated colour	17.0	14.0	20.8	17.8	–	–	–	–	4/22
Kalman/IoU on the same boxes	13.3	14.1	13.0	12.7	–	–	–	–	0/22
DINOv3 bank on the same boxes	14.7	14.1	15.8	14.6	–	–	–	–	5/22
Fine-tuned on SportsMOT train (45 videos; 37 held out)									
generated colour	26.9	26.0	28.2	31.9	26.9	26.3	28.4	32.2	18/159
Kalman/IoU on the same boxes	20.9	26.2	17.1	22.3	21.1	26.5	17.7	22.9	0/159
DINOv3 bank on the same boxes	23.0	26.1	20.7	25.9	23.1	26.4	21.3	26.6	16/159

the generator keeps the identity alive in pixels through the occlusion and the prefix carries it across windows, which is the state-in-pixels mechanism this paper set out to test, and also its limit, since an identity the model stops painting is lost.

4.4 Generalisation to SportsMOT

To test whether the recipe transfers, and not only the DanceTrack model, we repeat it end to end on SportsMOT (Cui et al., 2023) (45 train / 45 validation videos of basketball, football and volleyball; validation only, no test submission), re-tuning nothing: the same SAM 3 labelling of the train split (97.9% of boxes matched), a fresh LoRA on the stock backbone with the DanceTrack rank, learning rate and step budget, checkpoints selected by the same foreground-IoU rule on a fixed 8-video subset, and the chain, decoder and evaluation unchanged. Table 4 reports all 45 validation videos, the 37 outside the selection subset, and the DanceTrack checkpoint applied zero-shot. Training on the domain matters: on the eight selection videos mean per-video HOTA rises from 17.3 zero-shot to 28.0, ahead on every video, and the held-out mean (26.9) shows no selection inflation. The ordering on identical boxes replicates: AssA 28.2 for the colours against 20.7 for the appearance bank and 17.1 for Kalman/IoU (2,510 identity switches against 3,334 and 5,684). Long-gap re-acquisition does not carry over, and the reason is instructive: 369 of the 374 SportsMOT gap events are players leaving the frame, not occlusions. Inside the context the colours still re-bind 15 of the 53 events they hold (chance 0.10; bank 11, Kalman/IoU 0); beyond it they re-acquire 3 of 106, below chance (0.03 against 0.09), and across all re-acquired events the pre-gap colour was painted during the gap only half the time. A player who leaves the frame cannot be kept alive in pixels: the limit Section 4.3 names, measured where it dominates.

4.5 Ablations and analysis

Continuation training is necessary. Single-video ablations use the development video (n=1; checkpoints in Appendix A). With the same chain at inference but no prefix training, its full-video IDF1 is 0.172, *below* independent windows, because the model overrides the imposed colouring mid-window and errors compound. Prefix training lifts it to 0.466. The cost is unconditioned quality: single-window foreground IoU drops from 0.605 to 0.589, and independent windows from the prefix-trained checkpoint stitch to IDF1 0.423 against 0.466 from the translation-only checkpoint, which is why arm B uses the latter.

Window length and chain length. Generating 121-frame clips from the 49-frame translation-only checkpoint instead of chaining keeps mask quality but degrades identity sharply, so we chain at the trained length. Along the chain the loss is mild: full-video IDF1 on validation retains 85% of the 49-frame single-window value and the number of distinct colours per window does not grow (Appendix E).

Dense scenes are token-capacity-bound. Failure concentrates in crowded videos: the model saturates at 10–12 painted instances per frame at 512^2, where a frame is 16×16 latent tokens, regardless of ground-truth count, while recall of large people in sparse scenes is 0.86–1.0, so people in crowds are unrepresentable in that token budget rather than too small. Regenerating a dense video at 768^2 separates crowds but lowers box-level recall (0.48 to 0.32), because the 512-trained reference alignment warps at the unseen resolution, so resolution-matched training is the fix (Appendix F).

5 Limitations

Detection in dense scenes (DetA 37.6) and throughput (about 60 s per 49-frame window on one A100, 0.7 new frames per second at stride 40 and about 0.5 fps end to end on the test run, against 20–30 fps for tracking-by-detection) are the two deficits. Every reported number comes from a single sampling seed, and the seed matters: two further seeds on five validation videos give mean per-video HOTA of 37.4, 42.0 and 37.3 across the three, and conditional re-acquisition of 0.48, 0.32 and 0.44 on the same events, a spread of several points that is widest on the small re-acquisition counts. The mechanism comparisons are paired within a seed on identical boxes, and under every seed re-acquisition stays above four times chance with the pre-gap colour painted during the gap in *all* re-acquired events. The validation split served as the model-selection and development set, so its numbers carry selection bias that the test row does not. The palette bounds cumulative identities at 47 per video and colour reuse is unguarded. Two domains are covered, choreographed dance and team sports, the second on validation only; transfer to appearance-diverse pedestrian benchmarks such as MOT17 remains untested, and on SportsMOT the translation fine-tune peaks at half the DanceTrack step count (750 against 1,500), so the smaller clip set overfits sooner. Continuation training teacher-forces ground-truth prefixes, leaving a train/test gap to self-generated prefixes, and quality erodes along long chains.

6 Conclusion

Identity can live in pixels. A video generator, lightly fine-tuned to paint people with persistent colours and to *continue* a colouring it is handed, tracks full-length videos on the official DanceTrack server with no separately trained detector, no motion model, and no association step, with association quality above the 2020–22 benchmark suite on the benchmark built to break association; modern specialists score higher, and none share this error profile. Controlled comparisons locate the identity in generation-time state: windows generated without it and linked classically are twice as bad, post-hoc association adds nothing to the chain's own windows, classical associators on the same boxes fall well short of the colours on two datasets, and the generator keeps occluded identities alive in pixels inside and beyond its context window. We offer this as task-level evidence for the position of Gabeur et al. (2026) that strong generative models are generalist vision learners, and that "generation as the interface" is a viable, measurable path for video perception. Its two deficits, crowd detection bound by latent token capacity and throughput, are properties of the 2026-era backbone rather than of the formulation; whether they move with the next generation of video models is untested.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, to propose and refine hypotheses, to clean and reformat datasets, to support data analysis, and to interpret results; the method itself is the authors' own. Concretely, an LLM coding assistant wrote, modified, and executed the experiment, training, and evaluation scripts under author direction; proposed candidate diagnoses and control experiments that the authors accepted or modified; and produced intermediate analyses of logged outputs. We have not used generative AI tools to design the method, to generate synthetic datasets (silhouette training labels were produced by a non-generative promptable segmentation model on ground-truth boxes, Section 3.2, rather than annotated by hand), to develop theoretical models, to formulate mathematical claims, or to translate text; assisting with proofs is not applicable to this work. Additionally, we used AI tools to create figures, create and edit code, draft and edit parts of this manuscript, search for information, format references, and edit for readability.

We have reviewed all AI-assisted work. Every quantitative result reported here originates from a logged experimental run rather than from model-generated text: the authors fixed all experimental parameters in a pinned log before each run, and each number in the paper was traced back to its run record. LLM-written code is covered by an invariant test suite (including the codec round-trip test of Section 3.1) run before every experiment. All baseline values were verified by the authors against the cited primary sources rather than reproduced from model memory, AI-drafted text was checked by the authors against the experiment log and revised, and each qualitative example was verified against the underlying data before inclusion. Research direction, hypotheses accepted for testing, decisions at experimental gates, compute expenditure, and the final claims were determined by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses the public DanceTrack and SportsMOT benchmarks under their non-commercial research licences; its videos are Internet-sourced footage of identifiable people collected by the dataset authors, and we add no new personal data and release no RGB frames. Released ID-map outputs and checkpoints are for non-commercial research only. No human-subjects study was conducted.

REPRODUCIBILITY STATEMENT

All experiments are logged in a pinned experiment record (config hashes, seeds, costs, and gate verdicts per run) released with the code, ID-map codec, training and inference configurations, and evaluation scripts in the anonymised supplementary material submitted with this paper. Hyperparameters, prompts, decode rules and the TrackEval invocation are in Appendix A; per-video results, the re-acquisition protocol, the post-hoc associator, the dense-scene probe and the compute ledger are in Appendices B–G. The fine-tune uses a single 80GB GPU. The DanceTrack test result is a single frozen submission to the official server, with the submission archive and per-sequence outputs included in the release.

REFERENCES

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE International Conference on Image Processing (ICIP)*, 2016.

Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric SORT: Rethinking SORT for robust multi-object tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with

concepts. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 138846–138923, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/e0982cbc81401df3430ee1ff780dc7a2-Paper-Conference.pdf. arXiv:2511.16719.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J. Fleet. A generalist framework for panoptic segmentation of images and videos. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 909–919, 2023.

Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. SportsMOT: A large multi-object tracking dataset in multiple sports scenes. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, Yanan Bao, Karen Truong, Zhicheng Wang, Wenlei Zhou, Jonathan T. Barron, Kyle Genova, Nithish Kannen, Sherry Ben, Yandong Li, Mandy Guo, Suhas Yogin, Yiming Gu, Huizhong Chen, Oliver Wang, Saining Xie, Howard Zhou, Kaiming He, Thomas Funkhouser, Jean-Baptiste Alayrac, and Radu Soricut. Image generators are generalist vision learners. *arXiv preprint arXiv:2604.20329*, 2026.

Ruopeng Gao and Limin Wang. MeMOTR: Long-term memory-augmented transformer for multi-object tracking. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Ruopeng Gao, Ji Qi, and Limin Wang. Multiple object tracking as ID prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27883–27893, 2025. arXiv:2403.16848.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. LTX-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026. LTX-2.3 weights, trainer and IC-LoRA: https://github.com/Lightricks/LTX-2.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. DepthCrafter: Generating consistent long depth sequences for open-world videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2005–2015, 2025.

Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context LoRA for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024.

Junjie Jiang, Zelin Wang, Manqi Zhao, Yin Li, and Dongsheng Jiang. SAM2MOT: A novel paradigm of multi-object tracking by segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5388–5396, 2026. doi: 10.1609/aaai.v40i7.37455. arXiv:2504.04519.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9492–9502, 2024. arXiv:2312.02145.

Pengxiang Li, Kai Chen, Zhili Liu, Ruiyuan Gao, Lanqing Hong, Guo Zhou, Hua Yao, Dit-Yan Yeung, Huchuan Lu, and Xu Jia. TrackDiffusion: Tracklet-conditioned video generation via diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Jonathon Luiten, Aljoša Ošep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. HOTA: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 2021.

Run Luo, Zikai Song, Lintao Ma, Jinlin Wei, Wei Yang, and Min Yang. DiffusionTrack: Diffusion model for multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3991–3999, 2024. doi: 10.1609/aaai.v38i5.28192. arXiv:2308.09905.

Weiyi Lv, Yuhang Huang, Ning Zhang, Ruei-Sung Lin, Mei Han, and Dan Zeng. DiffMOT: A real-time diffusion-based multiple object tracker with non-linear prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani. Deep OC-SORT: Multi-pedestrian tracking by adaptive re-identification. In *IEEE International Conference on Image Processing (ICIP)*, 2023.

Jisu Nam, Soowon Son, Dahyun Chung, Jiyoung Kim, Siyoon Jin, Junhwa Hur, and Seungryong Kim. Emergent temporal correspondences from video diffusion transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, 2025. URL `https://openreview.net/forum?id=Ha6RTeWMd0`. arXiv:2408.00714.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Mattia Segu, Luigi Piccinelli, Siyuan Li, Yung-Hsu Yang, Luc Van Gool, and Bernt Schiele. Samba: Synchronized set-of-sequences modeling for multiple object tracking. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.01806. Author order follows the ICLR listing; the arXiv record lists Schiele before Van Gool.

Yuqing Shao, Yuchen Yang, Rui Yu, Weilong Li, Xu Guo, Huaicheng Yan, Wei Wang, and Xiao Sun. From detection to association: Learning discriminative object embeddings for multi-object tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6878–6888, 2026. arXiv:2512.02392.

Kyujin Shim, Kangwook Ko, Yujin Yang, and Changick Kim. Focusing on tracks for online multi-object tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11687–11696, 2025.

Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. TransTrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020.

Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. DanceTrack: Multi-object tracking in uniform appearance and diverse motion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *European Conference on Computer Vision (ECCV)*, pp. 402–419, 2018.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023a.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. SegGPT: Towards segmenting everything in context. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1130–1140, 2023b.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *IEEE International Conference on Image Processing (ICIP)*, 2017.

Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. Track to detect and segment: An online multi-object tracker. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12352–12361, 2021. arXiv:2103.08808.

Cheng-Yen Yang, Hsiang-Wei Huang, Kuang-Ming Chen, Kunjun Li, and Jenq-Neng Hwang. Extending segment anything model 2 to multi-object tracking by optimizing hierarchical trajectory memory. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 8358–8367, 2026.

Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. MOTR: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision (ECCV)*, 2022.

Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMOT: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 2021.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision (ECCV)*, 2022.

Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. MOTRv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European Conference on Computer Vision (ECCV)*, 2020.

Xingyi Zhou, Tianwei Yin, Vladlen Koltun, and Philipp Krähenbühl. Global tracking transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8761–8770, 2022. arXiv:2203.13250.

A PINNED HYPERPARAMETERS AND ADDITIONAL DETAILS

Model. LTX-2.3 dev checkpoint (22B joint audio-video transformer, 48 dual-stream blocks); Gemma-3 12B text encoder used only to precompute one fixed prompt embedding. **LoRA.** Rank 32, α 32, video-stream attention (QKV, output) and feed-forward projections only; audio stream excluded; 163.6M trainable parameters. **Training.** Phase A (translation): 3,000 steps, lr 2×10^{-4}, bf16, gradient checkpointing, flow matching with shifted-logit-normal timestep sampling, batch 1, 777 preprocessed clips ($512^2 \times 49$), training seed 0; checkpoint at step 1,500 selected by validation foreground IoU over checkpoints every 500 steps (0.53, 0.56, 0.605, 0.49, 0.58, 0.55). Phase B (continuation): resume from step 1,500 to 3,000 with gradient accumulation 4 and the prefix condition (2 latent frames, $p = 0.5$); checkpoint at step 2,500 selected the same way (foreground IoU 0.589). Foreground IoU is the IoU of the non-black region against the union of ground-truth silhouettes. **Silhouette labels.** SAM 3 prompted with each frame's ground-truth boxes as positive box exemplars, with the returned masks matched back to those boxes by IoU (97% of boxes; filled-box fallback for the rest). **Inference.** 49-frame windows, stride 40 (9-frame overlap = prefix length); 20 steps; classifier-free guidance 4.0; spatio-temporal guidance 1.0 on block 29; sampling seed 0; 512^2; prompt "segmentation ID map: paint each person a distinct flat colour on a black background"; negative prompt "worst quality, inconsistent motion, blurry, jittery, distorted". **Decode.** K=48 palette, index 0 = background; nearest palette colour in CIELAB for every pixel, so near-black snaps to background; per-colour connected components; minimum area 150 px; per-frame same-colour union. **Windows to tracks.** Overlap ownership split at midpoint; colour index = track id. **Association baselines on our own boxes.** The Kalman/IoU baseline is a SORT-style tracker (Bewley et al., 2016) run on arm A's per-frame boxes with the colour ids dropped. The appearance-and-motion baseline re-associates the same boxes with a DINOv3 identity bank, momentum-updated while a track is visible and frozen under occlusion, the bank of our image-model pilots run with its pinned parameters. **Evaluation.** TrackEval with HOTA, CLEAR and Identity metrics, DO_PREPROC False, predictions rescaled from 512^2 to native resolution; motmetrics for IDF1 in the single-video diagnostics. **Checkpoint per experiment.** Arm A, the test submission, the continuation ablation and the 768^2 probe use the phase-B step-2,500 checkpoint; arm B, the 121-frame test, the oracle decomposition and the 17-window independent numbers use the phase-A step-1,500 checkpoint. The SportsMOT rows of Table 4 use the SportsMOT phase-B step-1,500 checkpoint, trained from the stock backbone with a fresh LoRA (phase A: best validation foreground IoU 0.555 at step 750 of 3,000, selected by the same rule; phase B: 0.542 at step 1,500); the zero-shot rows use the DanceTrack phase-B step-2,500 checkpoint with no SportsMOT training. One validation pass run in a single process scores those two checkpoints on the 8 selection windows at foreground IoU 0.555 and 0.485 respectively, the SportsMOT checkpoint ahead on seven of the eight.

B PER-VIDEO VALIDATION RESULTS

Table 5 lists per-video HOTA components for arms A and B with the number of ground-truth people and predicted track ids; Figure 4 plots association and detection against the number of people. Arm A produces 906 predicted ids for 273 ground-truth identities (3.3 per identity); arm B produces 5,552 (20 per identity). Colour reuse is visible in the boxes as well: 32% of arm A's validation boxes span two or more same-colour components, and the 35 test videos use 24 to 47 of the 47 available colours. Across the 25 videos, DetA correlates with the number of ground-truth people at Spearman $\rho = -0.88$ and AssA at $\rho = -0.60$: the 15 videos with at most 9 people average AssA 45.8 and DetA 43.2, and the 9 videos with 14 or more average 27.6 and 24.5.

C RE-ACQUISITION PROTOCOL

Gap events are mined from the ground truth of each validation video: for every track and every pair of consecutive annotated frames f_{i-1}, f_i with $f_i - f_{i-1} - 1 \geq 9$, one event with gap $f_i - f_{i-1} - 1$; it is an "exit" if the last pre-gap box touches the frame border (2% margin), else an "occlusion". For a tracker output, the pre-gap identity is the majority predicted id over the nearest 15 visible ground-truth frames before the gap among frames where a predicted box overlaps the ground-truth box at IoU ≥ 0.3; the post-gap identity is defined symmetrically. *Tracked before* and *re-detected* require at least 3 matched frames on the respective side; an event is conditional if both hold, and re-acquired

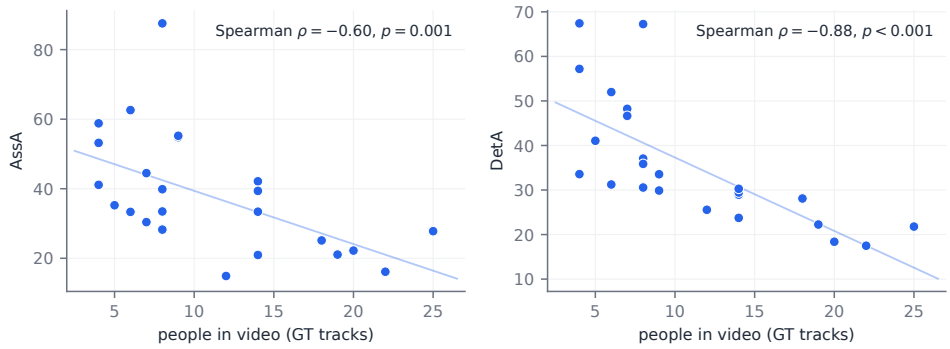


Figure 4: **Association and detection versus scene density** on the 25 validation videos (chained system). Both fall with the number of people; detection falls faster.

Table 5: Per-video validation results. People = ground-truth track count; ids = predicted track ids.

			A: chain					B: independent + assoc.				
video	people	frames	HOTA	DetA	AssA	IDF1	ids	HOTA	DetA	AssA	IDF1	ids
0004	4	1203	37.1	33.6	41.1	38.0	32	20.7	23.3	19.4	26.7	226
0005	4	1203	55.1	57.2	53.2	63.1	32	22.5	26.2	19.6	19.2	191
0007	8	1203	32.3	37.0	28.3	42.9	40	14.4	20.6	10.3	17.1	161
0010	6	1203	57.0	52.0	62.6	68.1	28	43.3	41.2	45.7	49.9	185
0014	12	1203	19.4	25.6	14.9	22.9	43	11.9	19.1	7.6	12.5	276
0018	8	503	76.7	67.3	87.6	86.0	25	51.2	38.5	68.6	60.9	96
0019	7	2402	38.3	48.2	30.4	49.1	36	15.1	24.8	9.2	13.9	510
0025	9	803	39.5	29.9	54.8	40.5	20	28.1	33.3	24.1	31.6	102
0026	19	302	21.0	22.3	21.1	22.0	45	11.3	14.8	9.6	10.9	183
0030	6	1263	32.2	31.2	33.3	35.7	22	18.2	29.0	11.6	17.4	104
0034	14	923	30.9	29.0	33.4	32.5	45	12.1	17.0	8.9	13.0	339
0035	8	703	38.3	37.1	39.9	50.1	34	23.7	28.5	19.8	26.0	219
0041	22	1003	16.7	17.5	16.1	18.1	43	7.4	10.6	5.4	8.1	315
0043	14	183	34.4	29.4	42.1	40.1	37	36.1	28.6	46.7	42.5	76
0047	8	1203	28.9	30.6	28.2	31.4	38	10.0	16.3	6.3	12.5	217
0058	7	1601	45.5	46.7	44.5	51.2	39	16.3	22.3	12.0	15.4	240
0063	8	1000	34.5	35.9	33.5	40.7	45	8.6	10.8	7.4	5.6	458
0065	5	702	38.0	41.1	35.2	46.6	30	21.4	28.5	16.3	22.7	155
0073	14	703	34.1	30.3	39.4	37.1	46	20.2	17.9	23.0	21.9	192
0077	9	1203	42.6	33.5	55.2	47.9	25	14.5	15.0	14.1	15.7	132
0079	18	1202	26.4	28.1	25.1	30.7	39	13.6	17.7	10.7	12.5	181
0081	20	984	19.9	18.4	22.2	22.2	44	8.2	12.7	5.5	11.0	278
0090	14	1004	22.0	23.7	20.9	24.9	46	12.8	13.8	12.2	12.2	343
0094	25	603	24.1	21.8	27.8	23.2	46	8.2	9.4	7.8	7.6	222
0097	4	1203	62.9	67.4	58.8	78.5	26	32.0	41.9	24.6	34.6	151
all	–	–	34.8	32.3	38.3	38.3	906	18.2	19.7	17.7	17.3	5552

if the two majority ids coincide. Bins are by gap length; the 8 events longer than 150 frames (none re-acquired by either arm) appear only in the "all" column of Table 3. All parameters were fixed before the study. Figure 5 shows one event.

D THE POST-HOC ASSOCIATOR (ARM B)

Windows are decoded independently. Between consecutive windows, each colour of window k is linked to the colour of window $k-1$ with the largest summed box IoU over the 9 overlap frames if that sum exceeds 0.5, many-to-one (several current colours may join one previous track, which absorbs within-window colour splits); unlinked colours start new ids, and same-frame boxes of one id are merged by union. The design was selected on the development video by an oracle decomposition: one-to-one Hungarian linking reached IDF1 0.413, many-to-one 0.466, and an oracle stitcher 0.698 (in-window mean 0.661). Global geometric gap healing and colour-affinity healing added nothing. On the development video the associator ties the chain (0.454 against 0.466); at scale it collapses on dense scenes (Table 2).

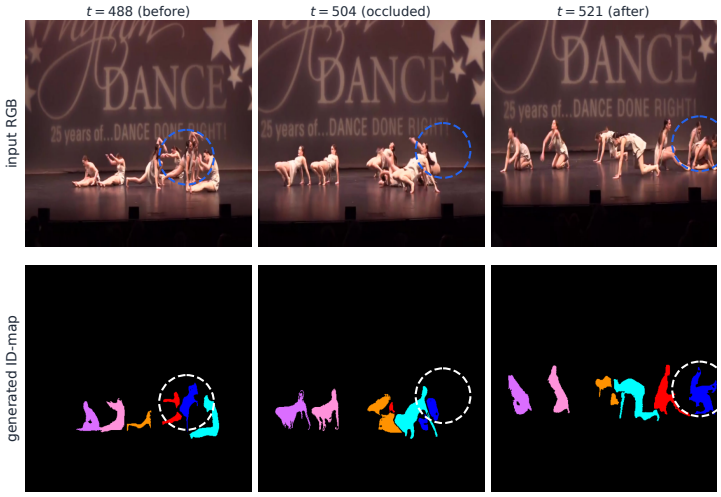


Figure 5: **Re-acquisition through occlusion.** A dancer (dashed circle) is absent from the ground truth for 27 frames and returns painted the *same blue*; the generator, not an association module, re-binds the identity. The small same-colour fragment that persists during the occlusion (middle) is the mechanism: in every re-acquired event the colour was painted somewhere during the gap.

E CHAIN LENGTH

Figure 6 plots, for the chained system, IDF1 over each window's owned frames and the number of distinct colours per window against the window index, averaged over the validation videos that reach that index (25 at window 1, 21 at window 18, 12 at window 30, one at window 60). IDF1 drifts from about 0.53 to about 0.49 over the first 25 windows and the colour count tracks the number of people rather than growing, so fragmentation does not accumulate along the chain; the rise after window 30 reflects the few long videos that reach those indices. Full-video IDF1 averages 0.417 per video, 85% of the 0.488 mean over 49-frame single windows, and the drift above amounts to 0.04 over about 1,000 frames. Generating longer clips instead of chaining is worse: 121-frame clips from the 49-frame translation-only checkpoint keep mask quality but lose identity, scoring IDF1 0.549 over frames 1–49 against 0.312 over frames 50–121, with a colour-switch rate 2.3× higher.

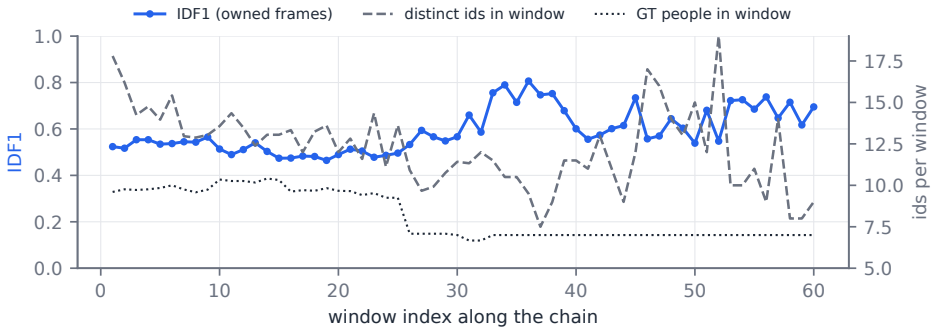


Figure 6: **Identity quality along the chain.** Per-window IDF1 (left axis) and distinct colours per window against ground-truth people (right axis), mean over validation videos reaching each window index.

F DENSE-SCENE PROBE

Per-frame counts and recall of large (≥80 px) people for the videos used in the dense-scene diagnosis at 512^2: dancetrack0041 (16.7 people per frame, 10.1 painted, merge rate 0.19, recall 0.41), 0081 (19.1 / 11.0 / 0.16 / 0.45), 0094 (19.5 / 12.1 / 0.20 / 0.48), 0026 (13.2 / 12.4 / 0.22 / 0.58), 0014 (10.5 / 7.6 / 0.32 / 0.71), 0018 (5.8 / 5.6 / 0.17 / 0.86), 0097 (3.9 / 4.4 / 0.14 / 1.00). Regenerating dancetrack0094 at 768^2 with the same checkpoint separates crowds but loses boxes: 13.1 painted per frame, merge rate 0.20 to 0.11, recall of large people 0.48 to 0.32 (Figure 7). Full-video IDF1 follows the same axis: 0.40–0.86 on the videos with 4–9 people against 0.18–0.33 on those with 14–25, so identity degrades with the number of people the token budget has to represent rather than with their size.

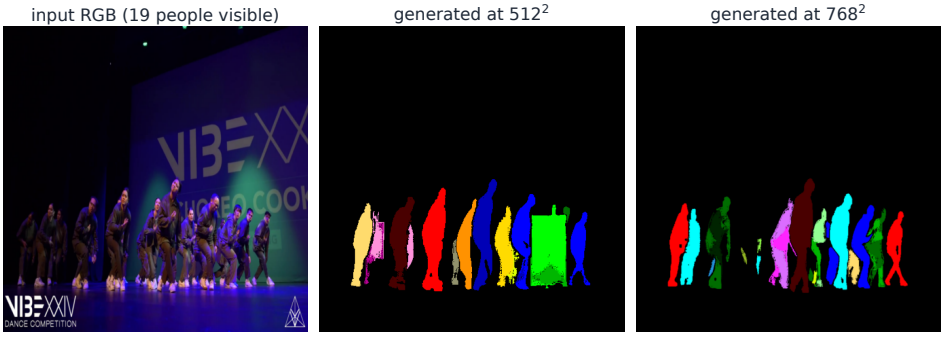


Figure 7: **Dense scenes are token-capacity-bound.** At 512^2 (16×16 latent tokens per frame) the model paints only ∼12 of 19 people, merges neighbours, and hallucinates a non-person region; at 768^2, with the *same* checkpoint, crowd separation improves dramatically, but reference alignment warps at the unseen resolution, so boxes degrade. Resolution-matched training is future work.

G COMPUTE LEDGER

A100-hours per stage, from the run records: SAM 3 silhouette labels 8.5; LTX feasibility and conditioning probes 2.7; Phase A training with six in-loop validation passes over the 25 validation videos ≈26 (estimated from the logged step time and validation duration; the run record holds no wall-clock total); Phase B training with three validation passes ≈17 (same estimate); development-video chains and length tests 0.8; the two full-scale validation arms 22; the 768^2 probe 1.5; test-set generation 26. SportsMOT arm, same recipe from the stock backbone: SAM 3 labels 5.3; zero-shot chain 4.3; Phase A 3.7; Phase B 7.8; the 45-video validation chain 20.2 (two launches, the first reaching 41 videos at its 18-hour limit); the single-process validation pass behind the foreground-IoU comparison 0.5. Three-seed check on five DanceTrack videos 7. Total ≈154 A100-hours, plus about 3.6 T4-hours and CPU time for evaluation. The image-model pilots that preceded the video arm (≈40 A100-hours and about $21 of image-editing API calls) are not included.