TL;DR 3DV-TON replaces per-frame warping with textured, animatable 3D human guidance and a diffusion UNet initialised from Stable Diffusion 1.5 + AnimateDiff. It cuts VFID on ViViD to 10.97 (from 17.29 for CatV2TON), introduces the 130-video HR-VVT benchmark at 720p, and ships code plus weights so teams can stand up consistent try-on pilots.

What is 3DV-TON?

3DV-TON is a video try-on framework from Alibaba DAMO Lab that keeps garment identity and motion consistent across frames. Instead of trusting pure 2D warping, it reconstructs a single textured 3D mesh from a keyframe, animates it with video-driven SMPL sequences, and feeds that textured guidance to a diffusion UNet. The paper debuts at ACM MM 2025 and arrives with a project page, inference code, model weights, and a new evaluation set.

Links:

- Paper: https://arxiv.org/abs/2504.17414
- Code + checkpoints: https://github.com/2y7c3/3DV-TON
- Project page (teasers, comparisons): https://2y7c3.github.io/3DV-TON/
- Launch recap: https://www.linkedin.com/posts/alexandremorgand_tryon-generativeai-imagegeneration-ugcPost-7321511630800920576-D34P

Why it matters for video commerce teams

- Drives conversion by preserving logos, textures, and fabric flow when shoppers see garments on moving bodies.
- Tackles the usual "good stills, jittery footage" failure mode with explicit motion references rather than heavier temporal smoothing.
- Adds HR-VVT (130 videos at 1280×720) so you can evaluate beyond the low-res, single-view ViViD standard.
- Ships open weights and a reproducible preprocessing stack (masking, SMPL fitting, 3D reconstruction), making it viable for in-house experimentation.

Inside the pipeline

1. **Adaptive keyframe selection** chooses the cleanest video frame and runs a 2D image try-on (CatVTON or similar) to produce an initial garment-wearing person.
2. **Animatable textured 3D mesh**: ECON-style reconstruction with SMPL-X refinement (10 iterations) creates a clothed human mesh. The team freezes pose parameters and only optimises shape, translation, and camera scale so reconstruction completes in ~30s.
3. **Video-driven animation**: SMPL sequences from GVHMR/Video-based HPS rig the mesh so textures follow body motion without stretching.
4. **Rectangular masking** expands the edit region, preventing original garment leakage before the diffusion pass.
5. **Diffusion UNet**: Stable Diffusion 1.5 backbone with AnimateDiff temporal blocks. Guidance features from the garment image and textured video are fused through self-attention layers, producing 32-frame clips at 768×576.

Training facts: VITON-HD + DressCode (paired images) and ViViD (paired videos) form the training mix; the model trains for 40k steps on A800 GPUs at lr 1e-5.

Benchmarks and numbers

- **ViViD benchmark** (paired, 768×576): 3DV-TON hits SSIM 0.899 and LPIPS 0.052, while slashing VFID (I3D) to 10.97 versus 17.29 for CatV2TON and 18.20 for ViViD. Even with a larger rectangular mask, it keeps VFID (ResNeXt) at 0.203.
- **HR-VVT (new, 720p)**: Against ViViD and CatV2TON, 3DV-TON records SSIM 0.880, LPIPS 0.086, paired VFID (I3D) 10.77, and paired VFID (ResNeXt) 0.142. In unpaired VFID, the method still leads with 14.55 on I3D.
- **User study (130 clips, 20 raters)**: 3DV-TON scores 69% preference on ViViD and 86% on HR-VVT for overall quality, favoured for both fidelity and motion coherence.
- **Ablations**: Adding SMPL guidance raises SSIM from 0.858 0.880. Injecting textured 3D guidance pushes it to 0.909 and halves VFID (I3D) from 5.24 2.38.

HR-VVT benchmark at a glance

- 130 evaluation videos spanning 50 tops, 40 bottoms, 40 dresses with outdoor/indoor scenarios.
- Distributed via Hugging Face (`2y7c3/HR-VVT`) with garment crops and metadata.
- Challenges legacy baselines that overfit to studio views; expect more occlusions, camera pans, limb crossings, and spectators in frame.

Working with the open-source release

1. Clone the repo and install requirements (`pip install -r requirements.txt`).
2. Pull base assets: Stable Diffusion image variations, AnimateDiff motion module, SD VAE, and 3DV-TON checkpoints into `./ckpts`.
3. Generate agnostic masks with the provided cloth masker (`python preprocess/seg_mask.py --type overall`).
4. Run GVHMR to recover SMPL sequences, then reconstruct textured meshes with the supplied renderer swap.
5. Edit `configs/inference/demo_test.yaml` and run `python infer.py --config ...` to render a 32-frame try-on clip.

Expect full pipeline latency around one minute per 32-frame video (30s for reconstruction + 35s for diffusion without cross-attention).

Implementation notes for production pilots

- The rectangular mask can be tuned; using ViViD-style tight masks gives marginally higher SSIM but loses robustness when garments swing.
- Freezing SMPL pose keeps reconstruction stable when limbs fall out of frame—useful for UGC where camera crops are aggressive.
- HR-VVT pairs well with qualitative review: include motion-specific checks (arm raises, scene transitions) when curating your own QA set.
- Integrate the cloth masker into upload workflows to enforce consistent agnostic conditioning before inference.

Limitations and open questions

- Residual temporal jitter appears when the 3D guidance fails (e.g., occluded arms); diffusion still needs reliable geometry input.
- The rectangular mask increases generated area, which can soften SSIM vs. tighter baselines—worth balancing per product shot.
- Dataset and code currently assume single-person, front-facing videos; multi-person scenes or extreme camera rotations remain open work.
- As the authors note, strong generative power raises misuse risk (e.g., deepfakes); incorporate consent and watermarking policies early.

Next steps for your team

- Stress-test on your catalogue videos: start with 10–20 SKUs mixing textures (denim, prints) to validate the textured guidance advantage.
- Compare against existing try-on stacks using VFID (I3D) and user preference studies; the HR-VVT protocol is a ready-made template.
- If you need faster turnaround, monitor the repo's TODO list (integrated image try-on and 3D guidance release) or swap in lighter SMPL estimators once code lands.
- Feed results back into merchandising flows—successful trials can seed PDP loops, social ads, or live shopping overlays.

References

- 3DV-TON (GitHub)
- 3DV-TON (arXiv)
- Stable Diffusion (GitHub)
- Latent Diffusion Models (arXiv)
- AnimateDiff (GitHub)
- AnimateDiff (arXiv)

CTA: