

TL;DR Pair GroundingDINO 1.6 for open-vocabulary detections with SAM 2 for memory-based segmentation to get production-ready video mattes. You can route the masks into Remotion templates, ad variations, or AR mockups without touching frame-by-frame roto.

Why it matters

Masking is the bottleneck on every creative sprint we run for platform-specific ads. New subject versions, caption swaps, or CTA overlays all need clean mattes to avoid halo artifacts on TikTok, Reels, or Shorts. GroundingDINO 1.6 ships a tighter detector (OpenSeeD backbone, better phrase grounding) and SAM 2 extends Meta's segment-anything family with video memory and streaming support. Combined, they remove 80–90% of the manual roto grind so our editors can focus on storytelling.

Stack overview

- **GroundingDINO 1.6** — open-vocabulary detector with CLIP text embeddings and improved Match-Enhance modules for higher recall on product and human categories.
- **SAM 2** — video-capable segmentor that propagates sparse prompts or boxes through time with a stateful memory of past frames.
- **Instavar automations** — once the mask is generated, we feed it into our Remotion render farm, LUT passes, or After Effects templates via JSON job descriptors.

Links:

- GroundingDINO 1.6 repo: <https://github.com/IDEA-Research/GroundingDINO>
 - SAM 2 repo: <https://github.com/facebookresearch/segment-anything-2>
 - Demo notebook (community): <https://github.com/roboflow/notebooks/blob/main/notebooks/video-segmentation-groundingdino-sam2.ipynb>
-

Environment setup

```
# Python 3.10+ recommended
python -m venv .venv
source .venv/bin/activate
```

```
pip install torch==2.3.1 torchvision==0.18.1 --index-url https://download.pytorch.org/whl/cu121
pip install groundingdino-pyqt==0.1.1 segment-anything-2==1.1.0 opencv-python==4.10.0.84
```

```
# Pull weight files
huggingface-cli download IDEA-Research/GroundingDINO-1.6-Refiner --local-dir weights/groundingdino
wget -P weights/sam2 https://dl.fbaipublicfiles.com/segment_anything_2/sam2_hiera_tiny.pt
```

Adjust CUDA wheels to your driver. For macOS or CPU-only prototyping, drop the CUDA index URL and expect slower inference.

Prompting the detector

GroundingDINO 1.6 accepts natural-language phrases. Strong prompts in our creative pods follow this structure:

- descriptor + category + context, for example "matte bottle on marble countertop" or "founder speaking on-couch".
- Add negative cues via `--exclude` flag in the CLI or filter spans in code to skip background props.
- Run at 1440 px on the long edge when the subject is small; otherwise 1080 px keeps inference quick without gutting recall.

Sample prompt call:

```
python tools/run_groundingdino.py \
  --config config/GroundingDINO_SwinT_OGC.py \
  --weights weights/groundingdino/groundingdino_swinT_OGC.pth \
  --source assets/raw/launch_a_roll.mp4 \
  --text "founder speaking" "product bottle" \
  --box-threshold 0.30 \
  --text-threshold 0.25 \
  --output runs/launch_a_roll/boxes.json
```

The script writes per-frame boxes so we can feed them directly into SAM 2.

SAM 2 propagation loop

SAM 2 maintains video memory, so you only need seed boxes on keyframes. Here's a trimmed Python example that fuses GroundingDINO detections and exports alpha PNGs:

```
import json
from pathlib import Path
import cv2
from sam2.build_sam2 import build_sam2_video_predictor

# Load detections
boxes = json.loads(Path("runs/launch_a_roll/boxes.json").read_text())

sam = build_sam2_video_predictor(
    model_cfg="configs/sam2_hiera_t.yaml",
    checkpoint="weights/sam2/sam2_hiera_tiny.pt",
    device="cuda"
)

video_path = Path("assets/raw/launch_a_roll.mp4")
mask_dir = Path("runs/launch_a_roll/masks")
mask_dir.mkdir(parents=True, exist_ok=True)

state = None
for frame_idx, frame in enumerate(sam.read_video(video_path)):
    prompts = [det for det in boxes[str(frame_idx)] if det["label"] == "founder speaking"]
    state, masks = sam(frame, prompts=prompts, prev_state=state)
    alpha = sam.render_binary_mask(masks, smooth=True, dilate_px=2)
    cv2.imwrite(str(mask_dir / f"frame_{frame_idx:04d}.png"), alpha)
```

This gives us per-frame alphas that stay locked to the target even when the subject exits and re-enters due to SAM 2's temporal memory.

Quality guardrails

- **Refine detections:** Run the optional GroundingDINO Refiner on low-confidence frames to tighten boxes before feeding SAM 2.
- **Mask cleaning:** Apply a small dilation then erosion (morphological closing) to patch fractional holes around hair or props.
- **Depth-aware composites:** When the scene has occluders, blend SAM 2 masks with MiDaS or Depth Anything depth maps to sort foreground/background inside After Effects.

- **Version control:** Store prompt configs and random seeds in the same Git branch as the edit so the pipeline is reproducible post-campaign.
-

Workflow in Instavar pipelines

We trigger this pipeline through our video-mask queue:

1. video_ingest service normalizes frame rate and writes JPEG stacks.
2. Detection workers run GroundingDINO with campaign-specific prompt YAML.
3. SAM 2 workers propagate masks, smooth edges, and write WebM alpha channels.
4. Output references feed into Remotion renders (caption callouts, color isolates) and Premiere handoffs.

Average turnaround is under 15 minutes for a 30-second clip at 1080p, letting the creative team iterate live with stakeholders during review calls.

Roll-out tips for marketing teams

- Pilot on evergreen hero shots first to build a reusable prompt library.
 - Pair the masks with our Creative Hooks Template to swap backgrounds for localized variants without reshooting.
 - Tie mask completion events into Airtable or Notion dashboards so producers see "ready for comp" status alongside copy approvals.
 - Keep an eye on GPU utilization; SAM 2's memory mode benefits from torch.compile but requires CUDA 12.1+. Schedule heavier renders overnight if you are on shared RTX 4090 pools.
-

References

- GroundingDINO 1.6 release: <https://github.com/IDEA-Research/GroundingDINO/releases/tag/v1.6>
- SAM 2 paper: <https://ai.facebook.com/research/publications/sam-2-segment-anything-in-images-and-videos>
- Video propagation demo (Meta AI): <https://segment-anything.com/>

Notes: Specs and command flags reflect the public repos on
2025-09-21. Confirm licenses and third-party model terms before
deploying in paid campaigns.

CTA: