THE FOUR PHOTOGRAPHIC TOKENS

Measured with Six vision models

SAM — subject masks (S)

DINOv3 — attention saliency (S)

InSPyReNet — salient-object mask (S)

CLIP — composition-prompt similarity (C)

LAION / FGAesQ — aesthetic assessment (A)

Border accounting — masks → captured vs lost mass (I)

Geometry — saliency → thirds · balance · lines (C)

= scored with the four photographic tokens

→ scores every frame, drives the refinement gates, and guides the compositional-aware explanation



Subject
Is there a clear main subject?

Composition
Is the main subject well placed?

Integrity
Is the subject kept whole?

Aesthetic
Does the frame look pleasing overall?

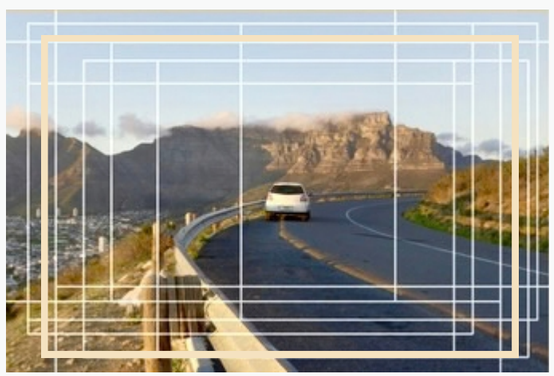
1 Input A 360° scene





Score = sum of the four token summands

2 Generate & shortlist



white: shortlisted boxes / light-yellow: chosen crop

Learned ranker

Each candidate scored by its own photographic tokens and the best crop of the candidates is chosen.

3 Rank & refine



light-yellow: chosen crop / yellow: refined crop

Learned refinement gates

Accept a refinement proposal only if the predicted IoU (against human-preferred crops) gain is positive.

4 Final Crop



camera parameters: **pan +165°, tilt -0°, fov 60°**

LoRA **Explanation**

*"... placing the car **centrally** while allowing for a dynamic **leading line** ..."*

underlined claims verified with the features: subject centre at 52% of width