

From Video to Controllable 3D

Geometry, semantics and verifiable feedback

Bingchen Gong

Postdoctoral researcher · LIX, École Polytechnique

13 October 2026

A convincing image is only the beginning

Can a captured object become an asset
we can **measure, name and change**?

Geometry

Recover a usable surface



Meaning

Locate a named part



Control

Edit and verify the result

My contribution: connect learned visual representations
to explicit geometric and semantic structure.

Long-term goal: controllable 3D and motion from visual observations.

A research trajectory across vision and graphics

2016–2021

HKU · PhD

Yizhou Yu
Visual learning and graphics

2022–2024

CUHK · Postdoc

Qi Dou
Neural reconstruction

2024–present

LIX · Postdoc

Maks Ovsjanikov
Geometry and multimodal learning

Scientific ownership

ZeroKey: conception, implementation and evaluation.
Good Match: scientific lead since April 2026.

Collaborative contribution

MILO / Spokes, PatchAlign3D, 3DHarnessBench.
Six co-mentored researchers and students.

MILo: put the mesh inside training



Extracted mesh and Gaussian rendering from the same reconstruction

A surface that constrains the fit

Sample pivots from the Gaussians.
Interpolate an explicit mesh.

Compare mesh and splat
depth and normals.

Gradients update Gaussian parameters.
Delaunay connectivity is refreshed discretely.

Image fitting gains an explicit surface that can be inspected and measured.

Spokes: orient Gaussians around the surface

An outward orientation helps wrap both sides of thin structures.



Reconstruction examples from the Spokes teaser

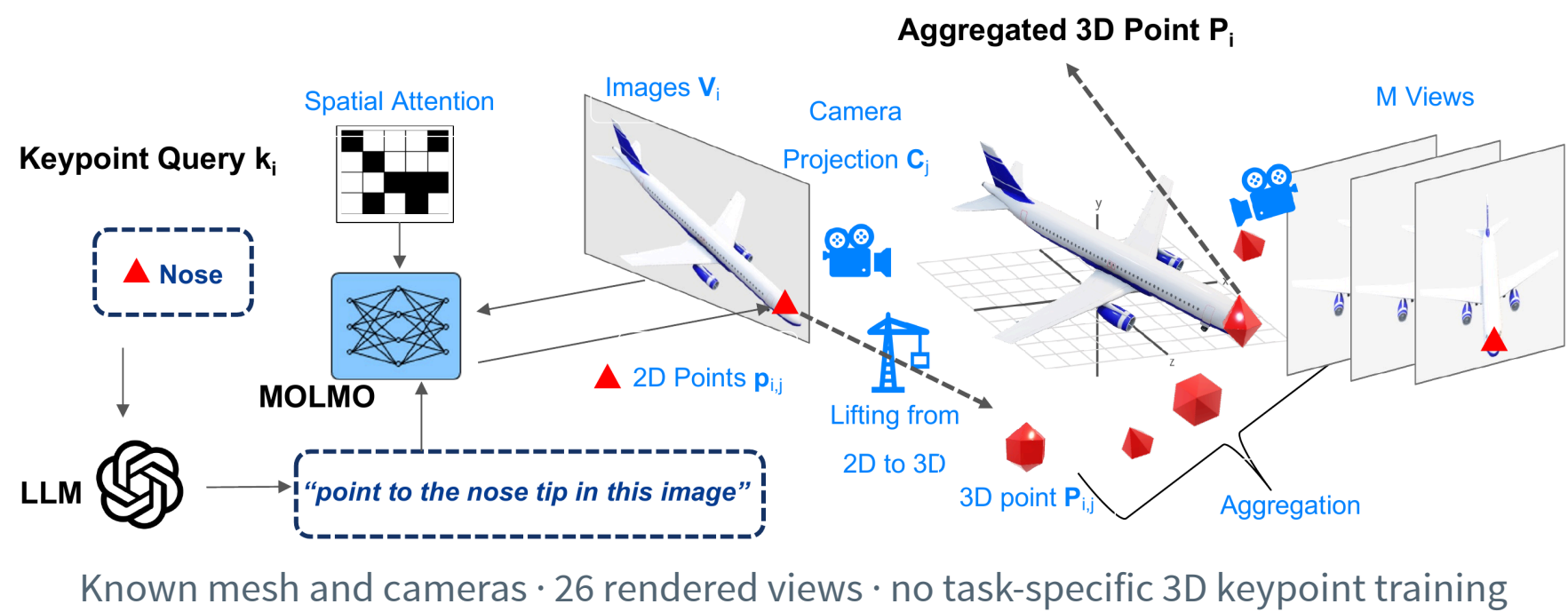
Evaluate the reconstructed surface, not just the rendered image.

Tanks & Temples · six-scene mean F1 ↑

Method	Uniform sampling	Virtual scanning
MILo	0.34	0.42
Spokes · 2 pivots	0.48	0.53

Same-paper surface protocols; no PAM.
Separate from mesh-size comparisons.

ZeroKey: lift language-guided points into 3D



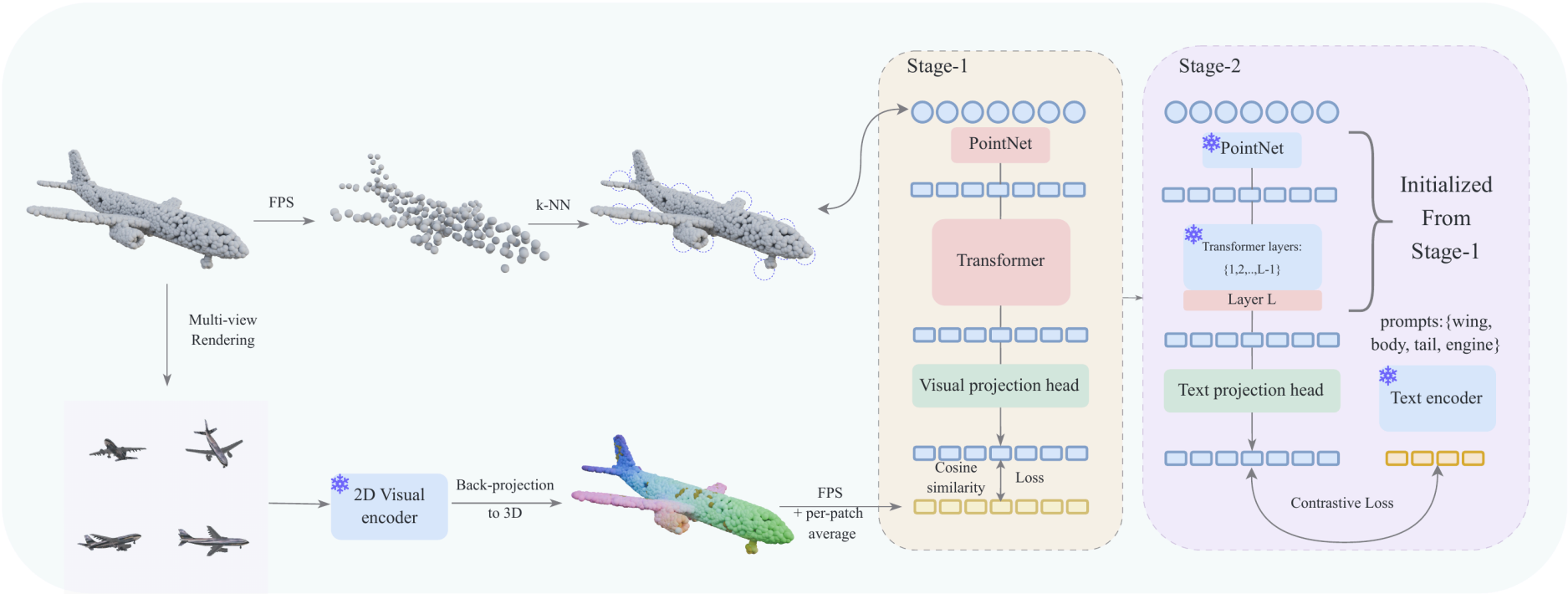
Soft voting absorbs pixel error

Lift a weighted pixel neighborhood,
then cluster hypotheses across views.

Voting weights	IoU@0.10 ↑
Constant · ablation	54.64
Gaussian · ZeroKey	79.43

KeypointNet: airplane, chair, table · IoU (%) · manually assigned benchmark text queries.

PatchAlign3D: visual structure, then language



2,048 XYZ points → 128 local patches → language-aligned features

Stage 1: distill DINOv2 features.

Stage 2: align patches with part names.

At inference: one encoder pass, no rendering.

Training	ShapeNetPart mIoU ↑
Joint	50.2
Two stages	56.9

The tested hybrid exposes a precision tradeoff

Patch–text similarity adjusts ZeroKey’s votes;
local features help cluster its 3D proposals.

KeypointNet comparison · IoU (%) at geodesic thresholds ↑

Method	0.01 · strict	0.05	0.10
ZeroKey	13.16	56.60	79.43
PatchAlign3D + ZeroKey	10.42	58.87	81.27

Better regional semantics can still reduce precise localization.

The trained part encoder complements the existing multiview pipeline.

3DHarnessBench: execute, inspect and revise

Observe → write Blender code
→ execute → inspect → revise.

100 static procedural objects; source programs withheld.

10-object Opus 5 ablation · normalized Chamfer ↓

Supplied evidence	Error
Four images	0.0150
Images + global / part measurements	0.0084

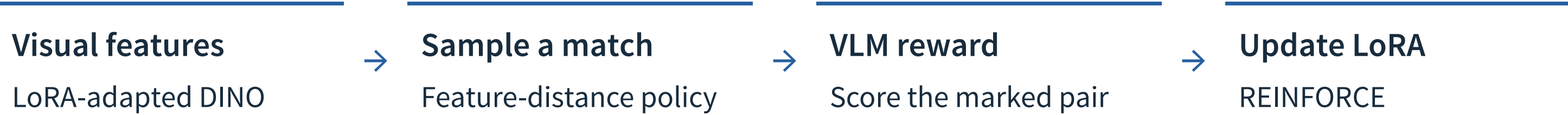
Geometry improves here; topology does not follow the same ranking.

Explicit measurements can support the next revision.



An executed reconstruction,
not the recovered source program

Good Match: learn from a VLM’s feedback



SPair-71k cat test · matched evaluation · PCK@bbox 0.10 (%) ↑

Method	PCK
Frozen SD + DINOv3	74.87
Geometry-aware LoRA · Kimi reward	77.74 ± 0.64

+2.87 points; a semantic reward still needs geometric calibration.

Source keypoints are given. No VLM at inference; precision and 3D transfer remain limitations.

Controllable 3D within MM

Build captured assets whose **geometry and meaning** remain usable as we compress, supervise and edit them.

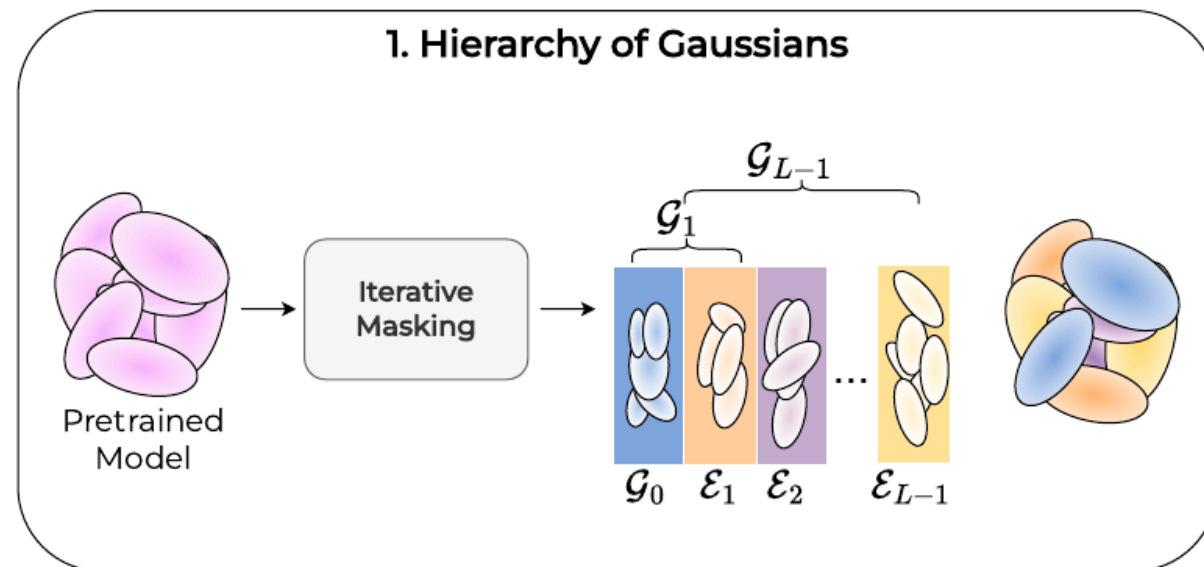
Research question	Starting point at Télécom Paris
What geometry survives a smaller model?	MM compression + our explicit surface models
When is visual feedback precise enough?	MM multimodal evaluation + our correspondence learning
Which controls make an asset usable?	IDS interaction and simulation + our reconstructed geometry

My research identity: representations with measurable 3D structure.

Longer-term outcome: compact, semantically addressable assets that support editing and interaction.

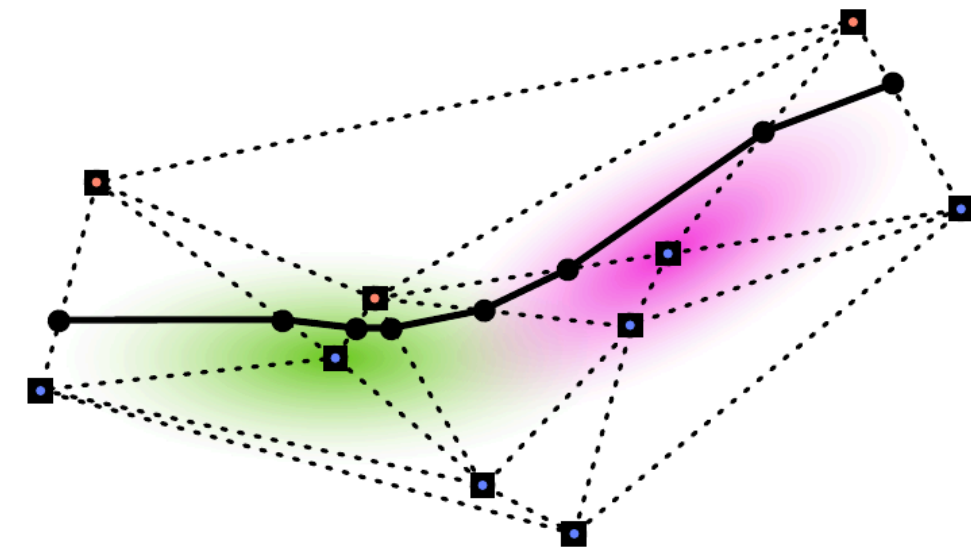
Does pruning preserve the reconstructed surface?

GoDe: nested Gaussian budgets



Rank primitives by rendering-loss gradients.
Jointly refine the nested levels.

MILo: Gaussians also support a mesh



Each retained Gaussian carries nine pivot values.
Changing the subset changes the triangulation.

First target: preserve thin surfaces in a progressively compressed **MILo-base** model.

Proposed extension with Enzo Tartaglione's compression methods; no combined result yet.

Test whether mesh feedback changes what we keep

Proposed ranking:
image-gradient rank + β · mesh-gradient rank

Mesh feedback: MILO’s depth / normal consistency, anti-erosion and interior losses.

Controlled comparison	Hierarchy	Joint refinement
A · image-only baseline	Image gradients	Image loss
B · isolate mesh feedback	Same as A	+ MILO losses
C · test the proposed ranking	Image + mesh gradients	Same as B

At matched bytes: image quality · surface F1 · thin-part recall

Tanks & Temples · fixed codec, precision and training steps.
Count pivot data and dictionaries; measure decoding and mesh-rebuild time.

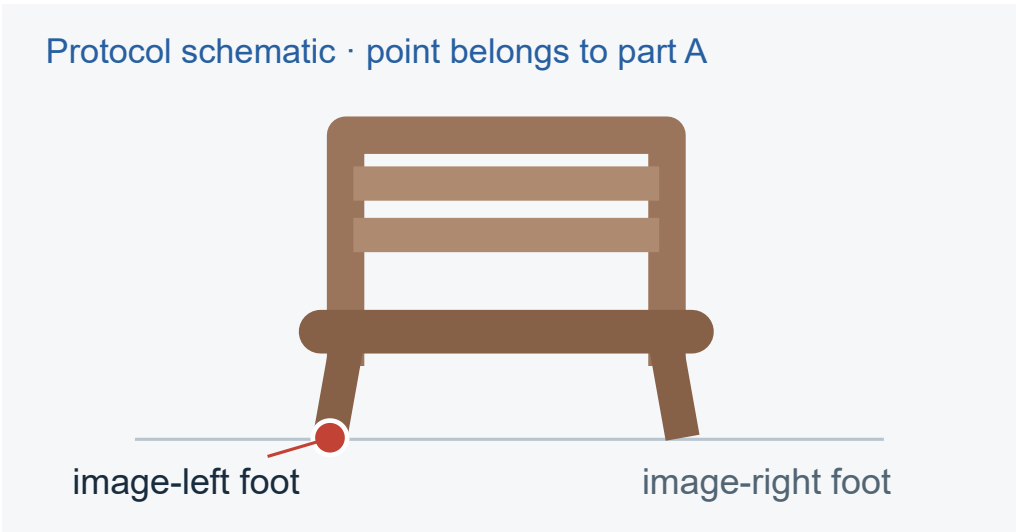
Can the judge reject one wrong detail?

Stephan Alaniz and coauthors · FINER

Pair a correct question with a minimally altered false premise. **Both answers must be correct.**

Published paired accuracy	Multi-rel
InternVL3.5-14B	47.0%
+ FINER-Tuning	71.2%

FINER-CompreCap · Table 1
Five-choice questions; DPO trains answers.



Proposed extension: known part IDs and visibility

- Q+ Is the point on the brown chair’s **image-left** foot? **Yes.**
- Q– Is the point on the brown chair’s **image-right** foot? **No; image-left.**

Illustrative answers; the test uses shuffled MCQs.

My executable-scene tools could give their negative-query research controlled 3D counterexamples.

Test the score before it trains the matcher

Freeze the image pair and source point; perturb only the target.

Target candidates	Required evidence
True point + offsets of 1%, 3%, 10%	Score should rank smaller errors higher
Same part type, wrong limb / repeated part	Reject a semantic look-alike
Background or occluded target	Reject or abstain; report separately

Offsets use the object-box scale. Compare Qwen2.5-VL-7B with its released FINER checkpoint at equal budget.

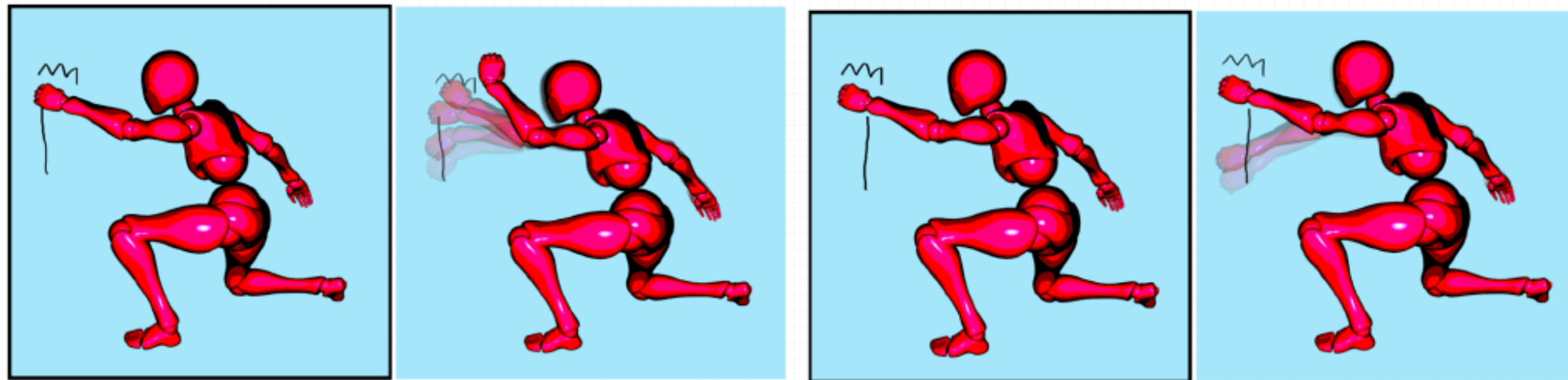
Proposed score → labeled calibration → held-out decision

- P(same precise landmark) from fixed-choice logits.
- Held-out synthetic + real pairs: test score ranking first.
- Then compare matcher PCK after otherwise identical Good Match training.

Use it in Good Match only if coordinate accuracy improves. Otherwise retain FINER as a diagnostic.

Which joint should this stroke control?

With Amal Dev Parakkat: address a documented SMACC failure.



SMACC, Figure 14 · the trajectory and impact can select different fingers.

The attachment fails first

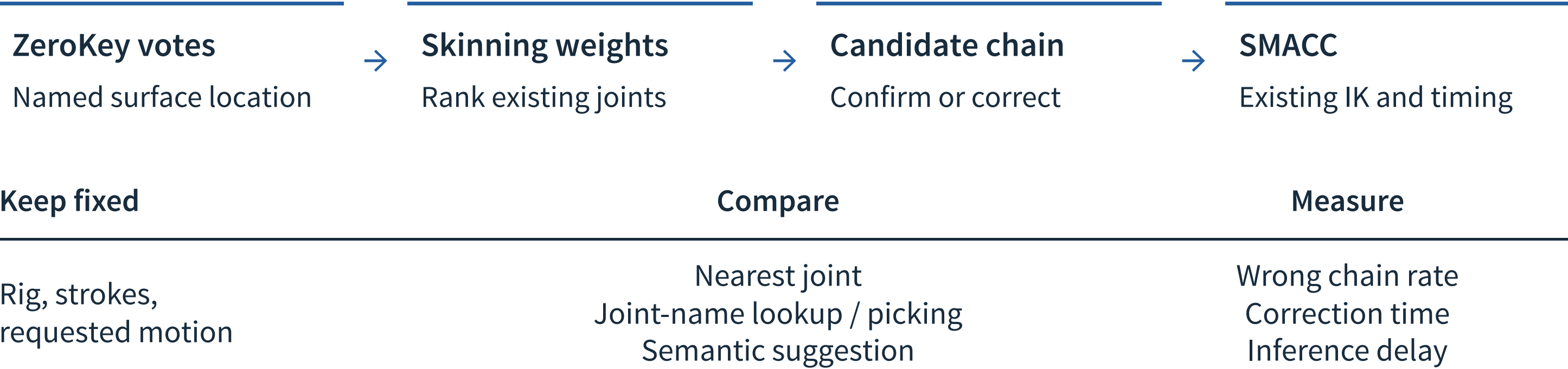
Nearest projected joints
can belong to different chains.

Name the intended control

Use “the character’s left wrist”
to restrict the candidate chain.

Resolve the target before the existing motion solver runs.

Map surface evidence to a candidate chain

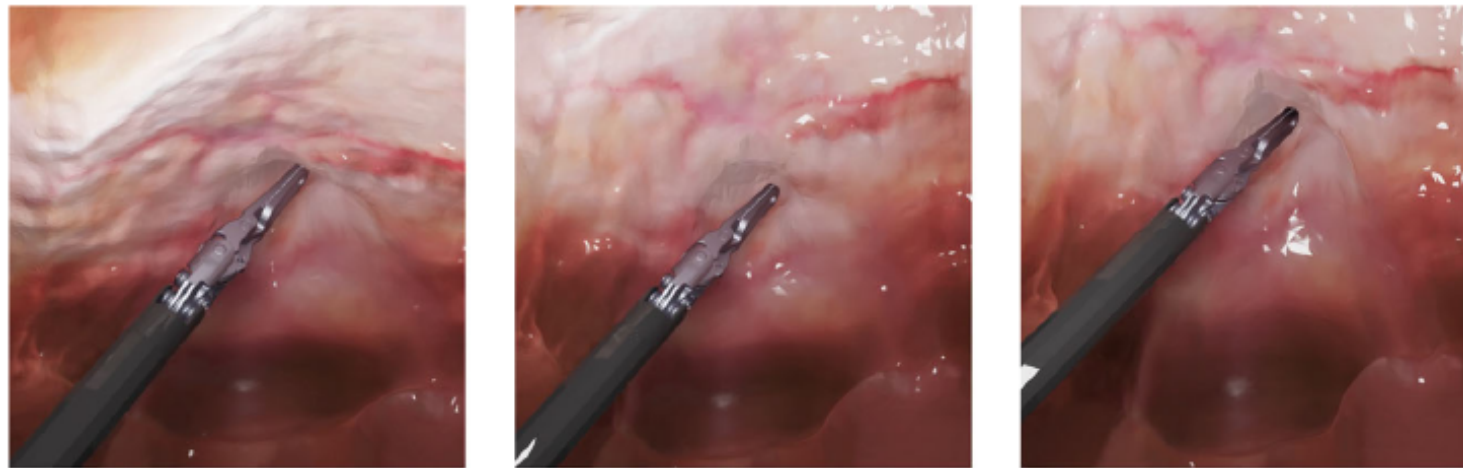


Show uncertainty; fall back to a click when views disagree.

The point identifies a surface region, not an internal joint center. Test whether vision adds value beyond rig names.

A reconstructed surface must survive interaction

Efficient EndoNeRF · coauthored work



Reconstructed surface → closed mesh → FEM

Wang, Gong, Long, Fan & Dou · 2024, Figure 4

The hidden base and material settings are supplied.

A question for Kiwon Um

When does reconstruction error
make relaxation fail?

Dimension Expansion · 2025

3D springs → auxiliary coordinates
→ damp back toward 3D.

More freedom can escape a trapped configuration.

Successful untangling can still distort the intended motion.

Separate recovery from unwanted physical change

One synthetic deforming beam, rendered and reconstructed.

Initial state	3D spring relaxation	Auxiliary-coordinate relaxation
Reference initial state	Reference behavior	Does the motion change?
Reconstructed counterpart	Does the solver fail?	Recovery at what cost?

Hold the mechanics fixed

Rest graph and correspondences supplied.
Clone loads, materials, boundary conditions and time step.

Read out both outcomes

Constraint error and runtime.
Shape / motion error and energy change.

Fewer failures count only alongside the distortion they introduce.

First year: establish two useful research tools

Months 1–3	Months 4–8	Months 9–12
Reproduce GoDe on explicit Gaussians; retain MLo’s surface state.	Separate pruning and quantization; test geometry at matched bytes.	Release a geometric rate–distortion benchmark.
Reproduce FINER’s paired test; construct known point errors.	Measure reward sensitivity on held-out synthetic and real pairs.	Release a calibration protocol; train only with a validated reward.

Then select one IDS extension: rig attachment or deformation robustness.

A student can own one baseline, adapter or evaluation module.
Continue only if the measured benefit justifies computation and human effort.

Teaching experience and preparation

HKU · 2016–2021

Eight courses

Teaching assistant.
Tutorials, programming labs,
assignments and projects.

Completed during my PhD

CTLHE

Certificate in Teaching and
Learning in Higher Education.
Planning, feedback and reflection.

MVA M2 · autumn 2026

TA appointment

Geometry Processing and
Geometric Deep Learning.
Materials, practical work and
assessment.

Teaching: make the reasoning visible

Align the learning outcome, the exercise and what is assessed.

Predict

What changes when the camera moves?



Implement

Render, reconstruct and inspect errors.



Explain

Defend a diagnostic and revise the result.

Give feedback while students can still use it.

Share session plans, assessment criteria and peer observation with the teaching team.

Contribute through existing teaching

Existing course area	Contribution I can offer
Deep Learning for Computer Vision APM_5DA12_TP	Visual features, limited-data learning and rigorous evaluation
Multimodal foundation models / Image Generation IA709 / IA719	Spatial grounding, neural rendering and failure analysis
Vision 3D et Vidéo CSC_4IM08_TP	Camera geometry, motion and reconstruction

Agree a concrete lab or module with the course leaders.

English teaching now. French teaching objective: within two years of arrival.

A lab on reconstruction failure

Explain why a reconstruction fails
and test a correction.

Given

A short scene video
and a starter notebook.



Investigate

Vary the available views.
Predict and explain a failure.



Submit

A reproducible result,
diagnostic and revision.

Review the prototype before the final submission.

A common rubric and an individual explanation; adapt prerequisites for continuing education.

Develop the programme through MM, IDS and teaching

Setting	A contribution colleagues and students can use
MM research	3D assets with measured storage, surface and grounding errors; reproducible compression and feedback experiments.
IDS collaboration	Build on the Parakkat connection through 3DHarnessBench; test specific interfaces for editing and deformation.
Teaching and supervision	Turn a baseline or diagnostic into an M2 project and practical lab; develop a doctoral question after the first evidence.

External partnerships can start with a measured need:
asset size, editing effort or the reliability of a visual check.

A programme that others can build on

Geometry you can measure.

Semantics you can test.

Assets you can control.

MM provides a strong setting to connect these capabilities with efficient, reliable visual learning.

Thank you.