GA3CE: Unconstrained 3D Gaze Estimation with Gaze-Aware 3D Context Encoding

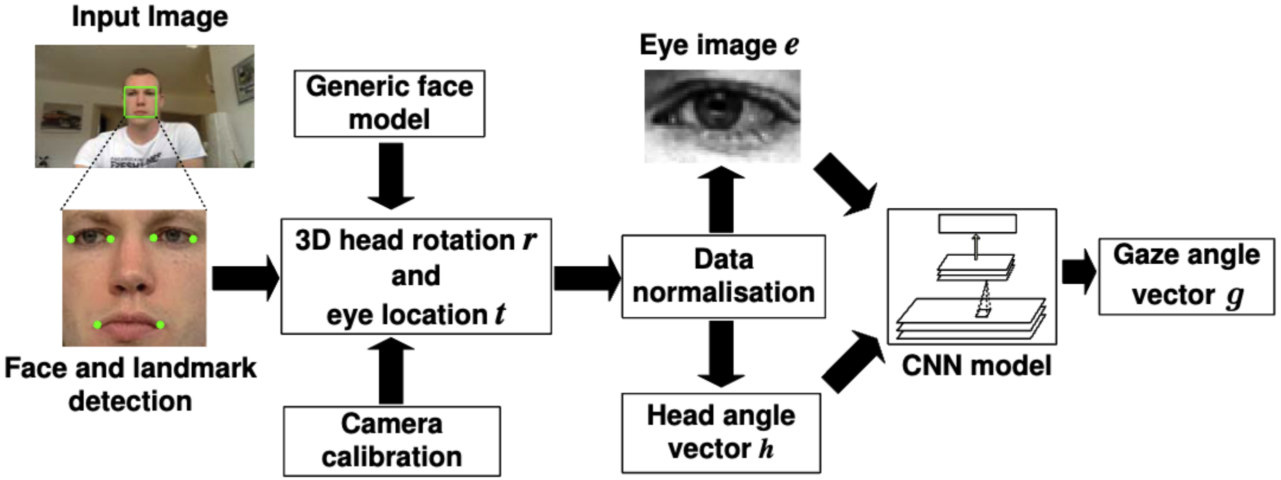
Yuki Kawana   Shintaro Shiba   Quan Kong   Norimasa Kobori

Woven by Toyota

Unconstrained 3D gaze estimation

- Estimating the 3D direction of gaze
- Without assuming access to close-up views of the eyes

Unconstrained setting [1]

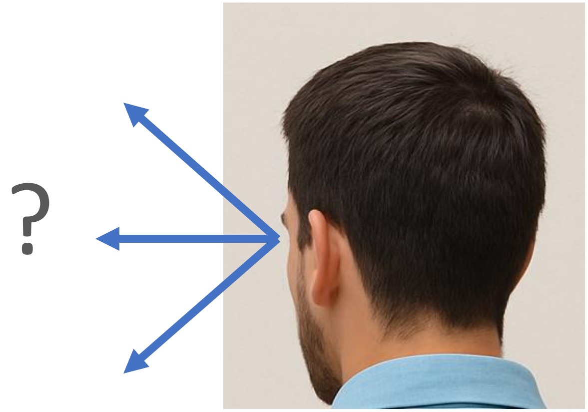


Constrained setting [2]

[1] Kellnhofer et al. Gaze360: Physically Unconstrained Gaze Estimation in the Wild. ICCV 2019.
[2] Zhang et al. Appearance-Based Gaze Estimation in the Wild. CVPR 2015.

Challenge #1: Ambiguity

- Ambiguous 3D gaze under the unconstrained setting

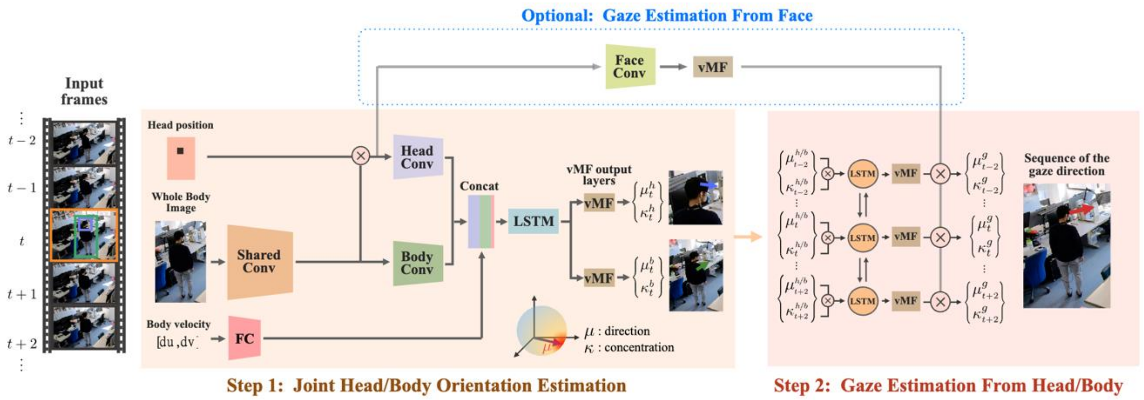


Previous works: 3D gaze from spatial context cues

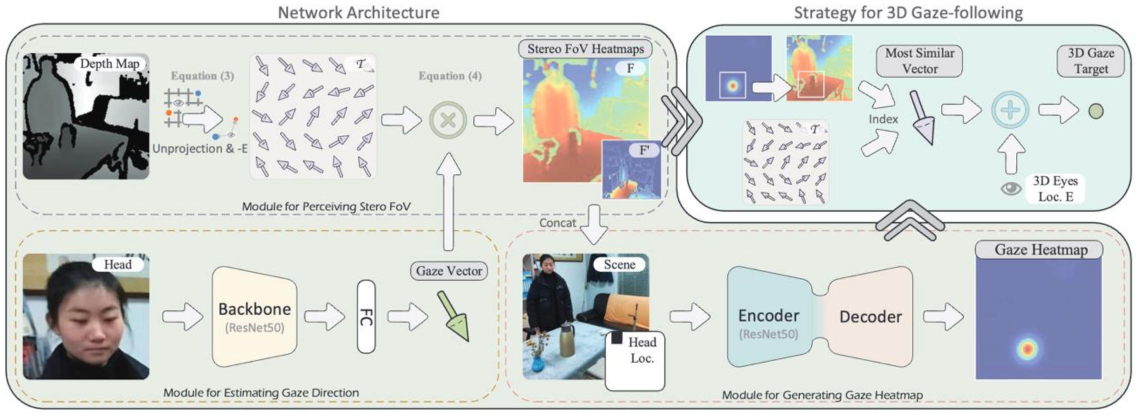
- Leverages spatial context cues for gaze estimation

Previous works: 3D gaze from spatial context cues

Body crop as input [1]



Scene RGBD image as 2D features + geometric post-processing [2]

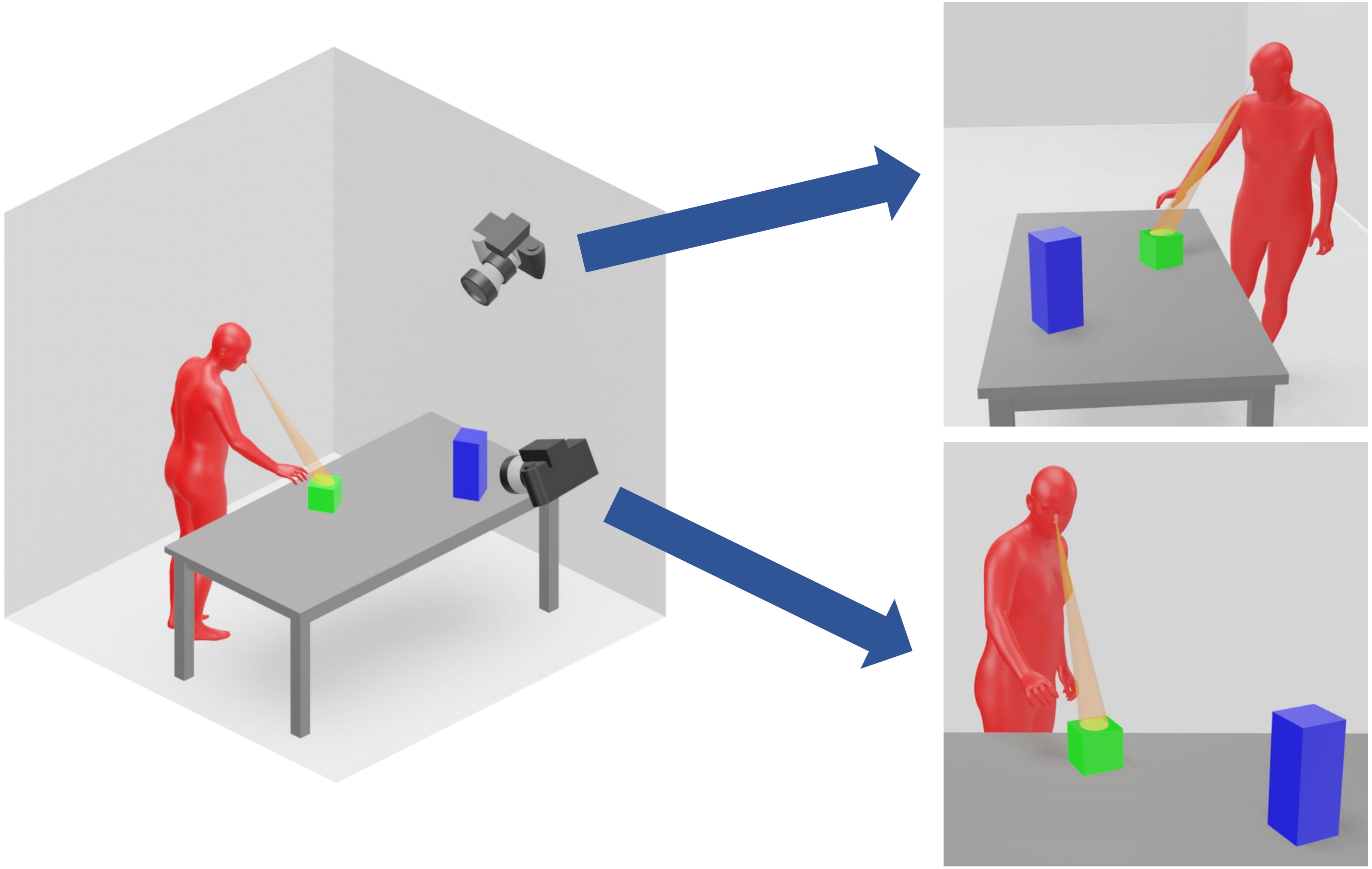


✗ Limited 3D understanding

[1] Nonaka et al. Dynamic 3D Gaze From Afar: Deep Gaze Estimation From Temporal Eye-Head-Body Coordination. CVPR 2022.
[2] Hu et al. GFIE: A Dataset and Baseline for Gaze-Following From 2D to 3D in Indoor Environments. CVPR 2023.

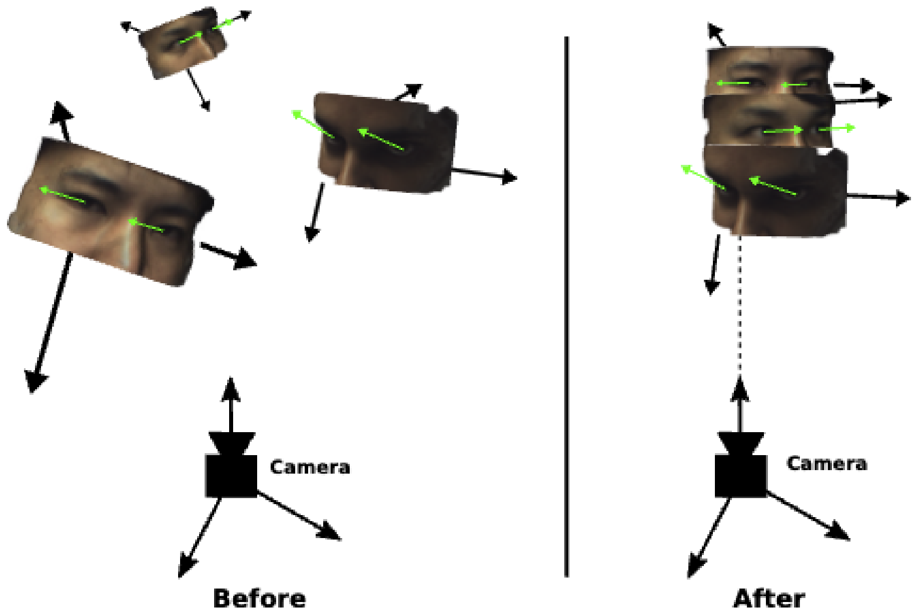
Challenge #2: Variability in viewpoint

- Diverse 2D observations and 3D gaze due to different viewpoints



Previous works: Normalization

- Head appearance normalization in constrained setting [1,2,3]



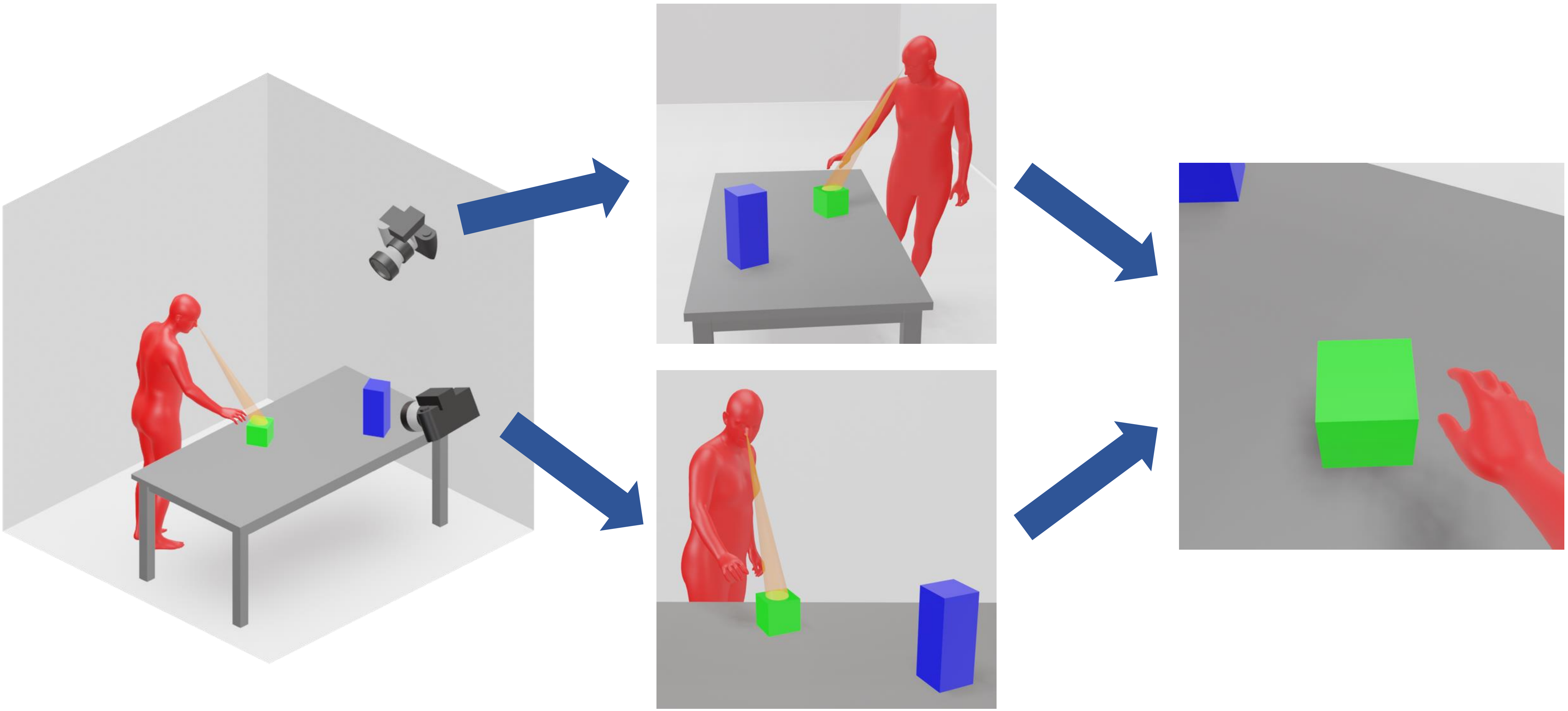
Appearance normalization [3]

✗ Not applicable to spatial context cues

[1] Sugano et al. Learning-by-Synthesis for Appearance-based 3D Gaze Estimation. CVPR 2014.
[2] Zhang et al. Appearance-Based Gaze Estimation in the Wild. CVPR 2015.
[3] Zhang et al. Revisiting Data Normalization for Appearance-Based Gaze Estimation. ETRA 2018.

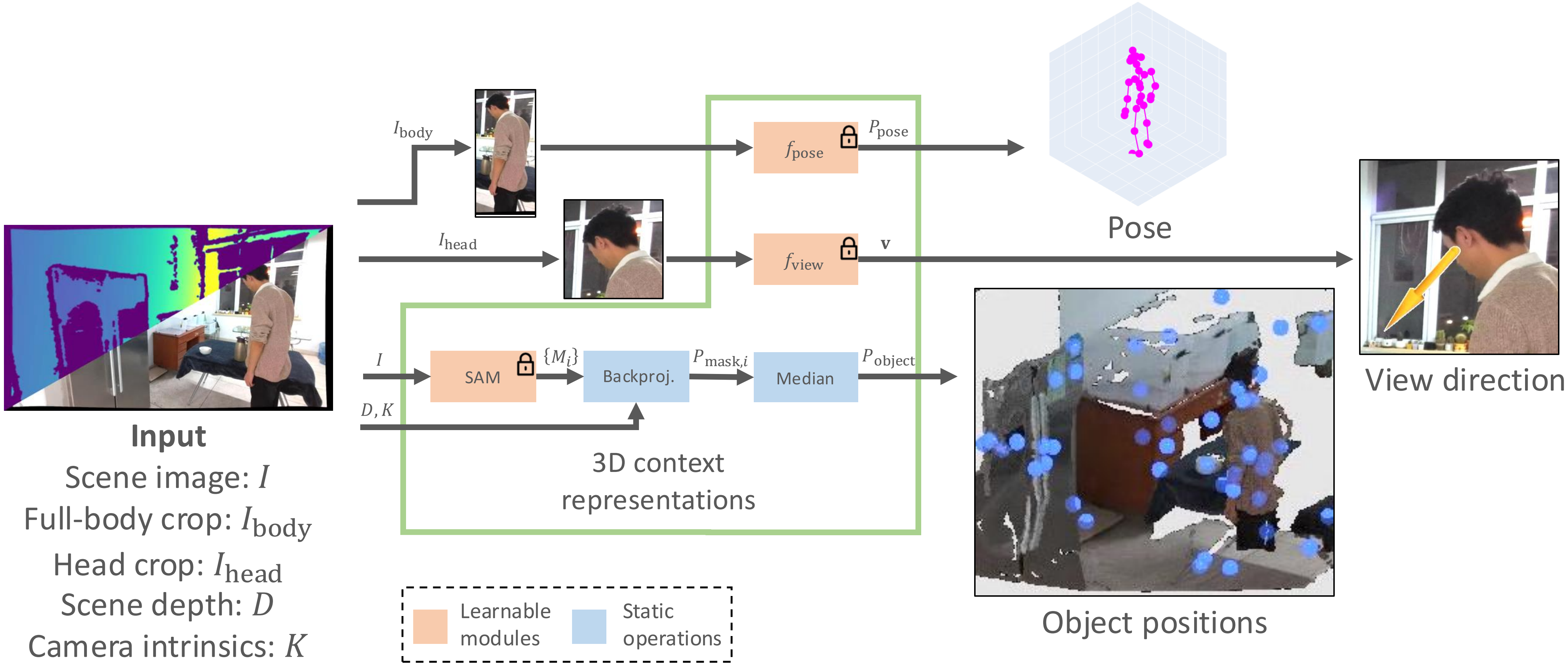
Our approach: Core idea

- Learning spatial reasoning **within egocentric 3D space**



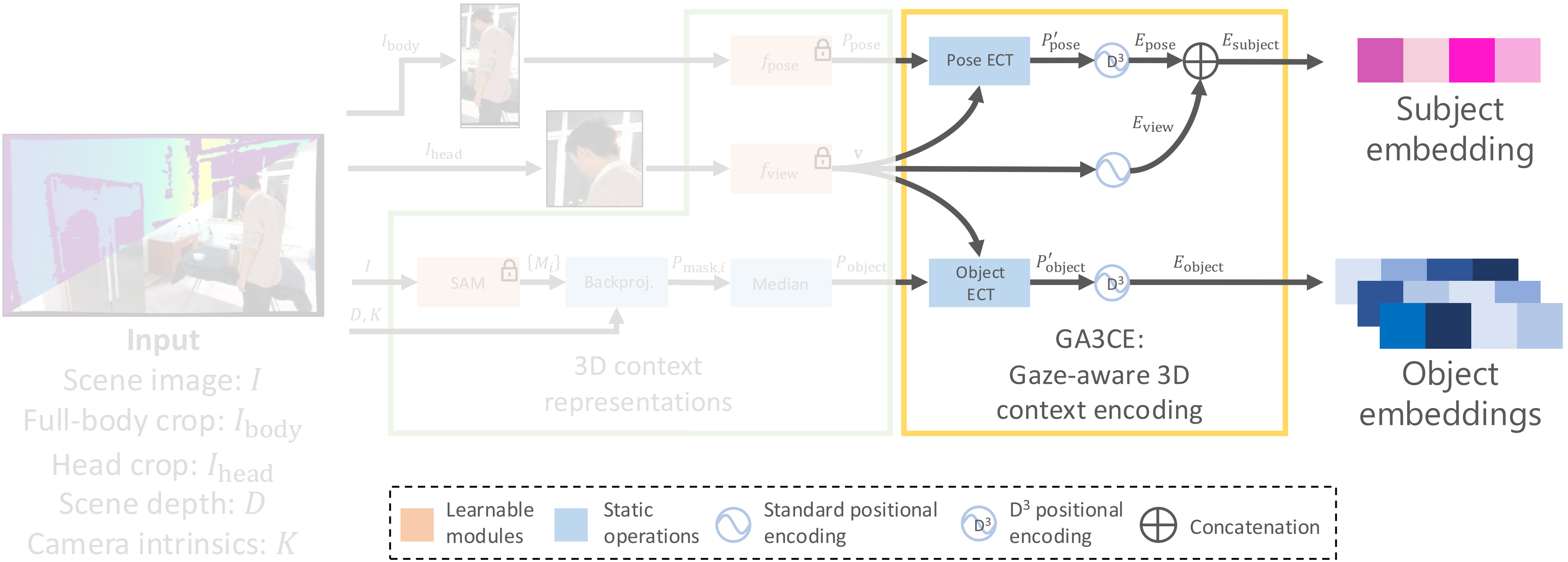
Our approach #1: 3D context representation

- Use 3D representations as intermediate features



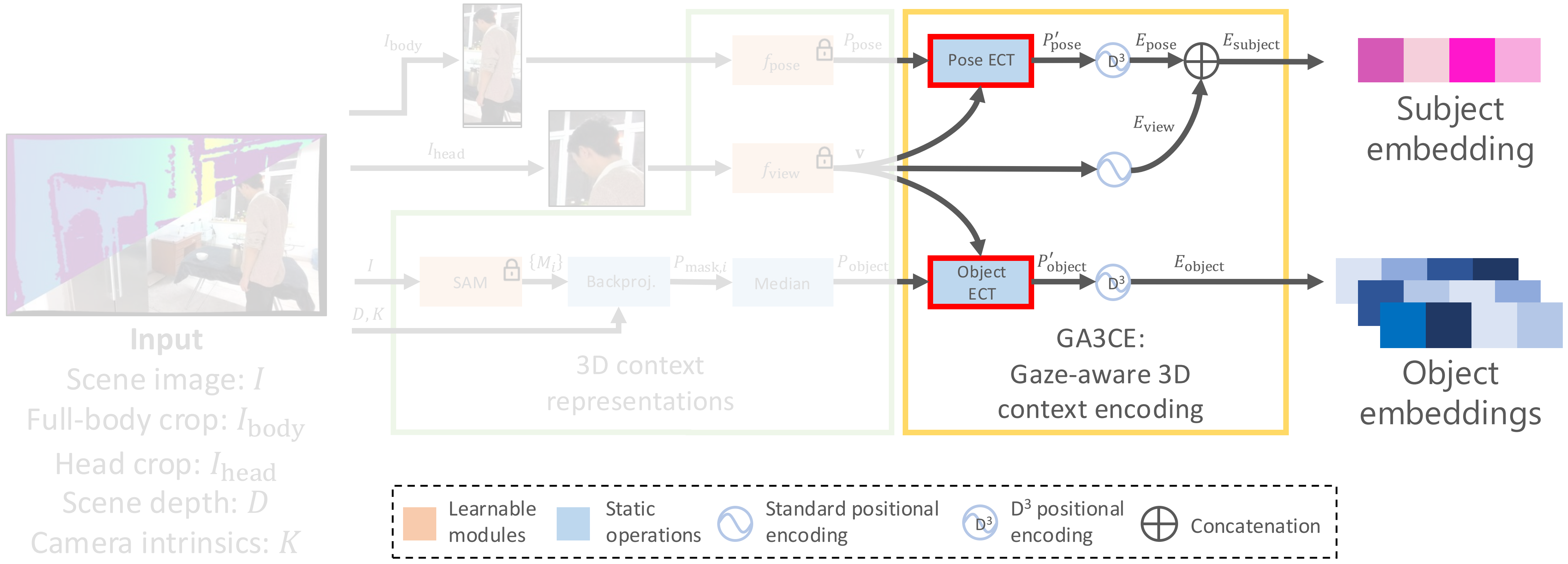
Our approach #2: **G**aze-**A**ware **3**D **C**ontext **E**ncoding

- Encode 3D context into egocentric, gaze-aware embeddings



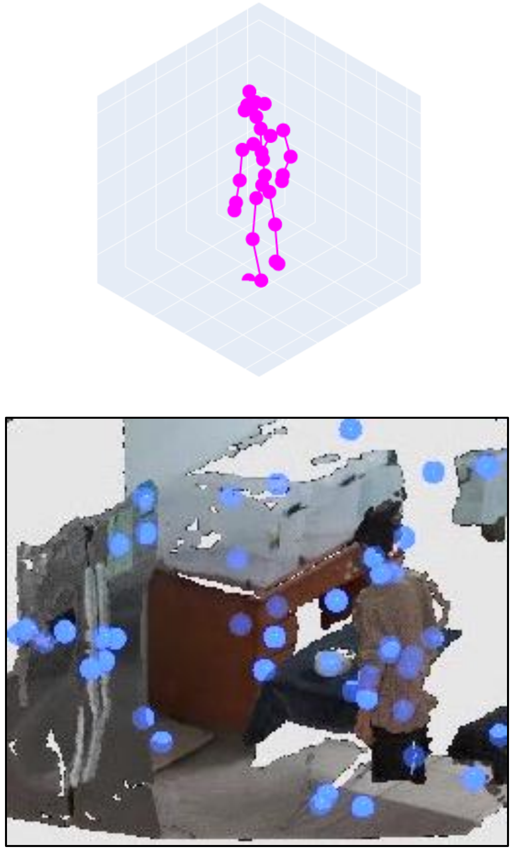
Our approach #2: **G**aze-**A**ware **3**D **C**ontext **E**ncoding

- Encode 3D context into egocentric, gaze-aware embeddings

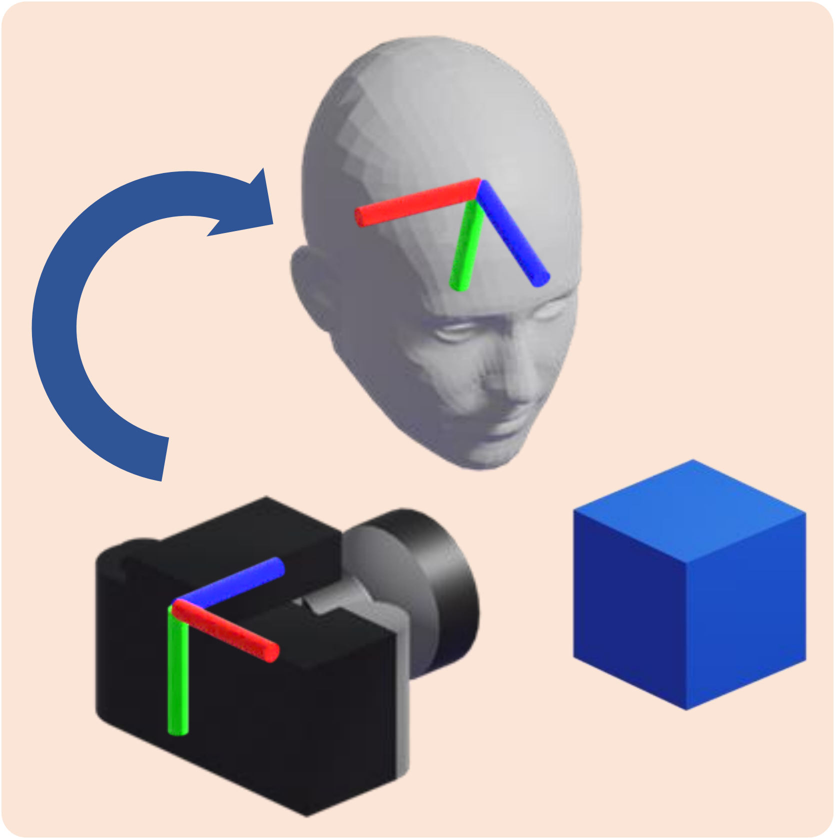


GA3CE #1: Egocentric transformation (ECT)

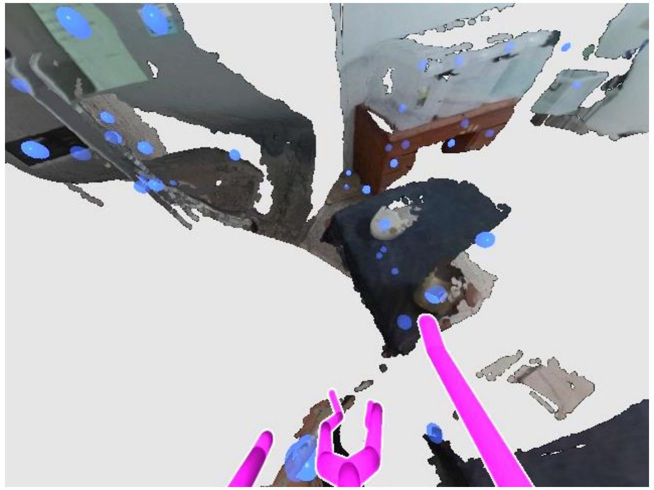
- Normalize pose and object positions into egocentric space



Camera space



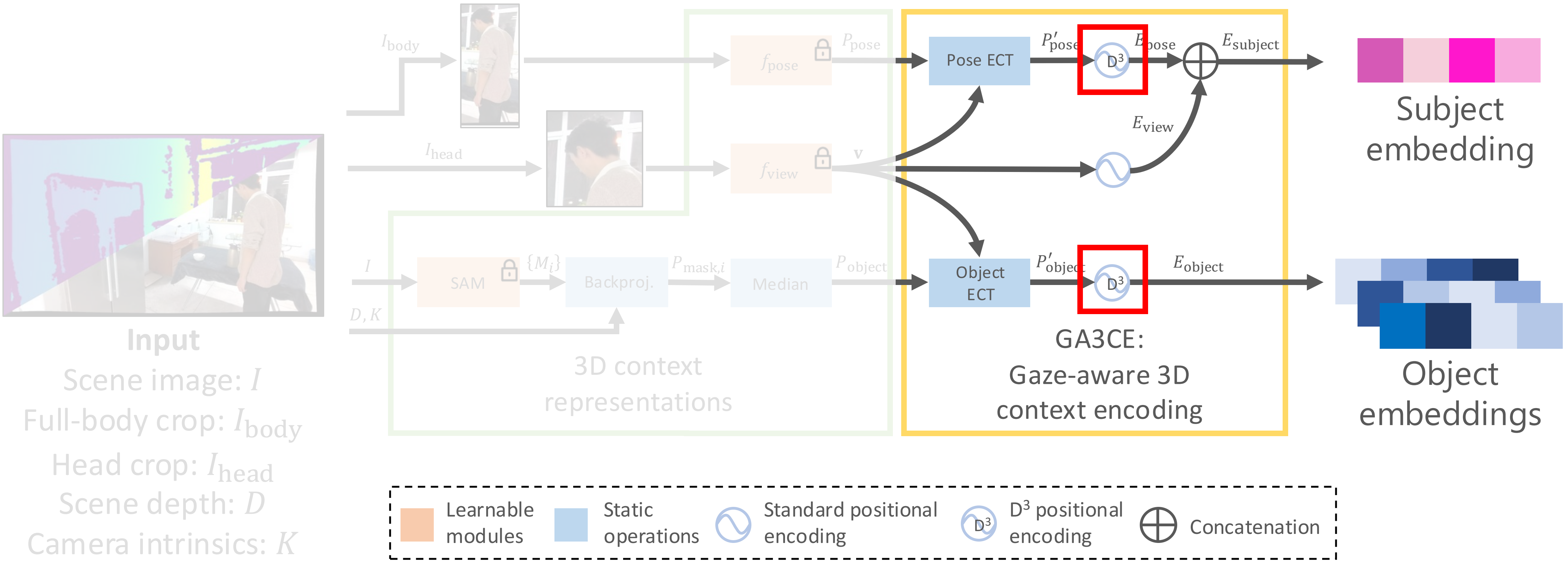
Egocentric transformation



Egocentric space

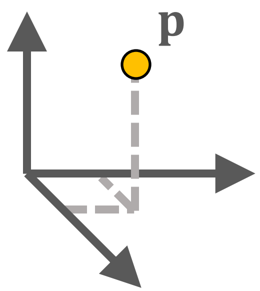
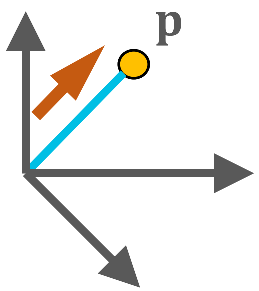
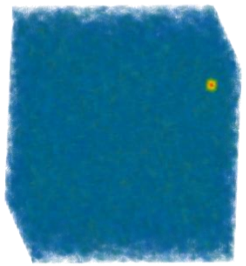
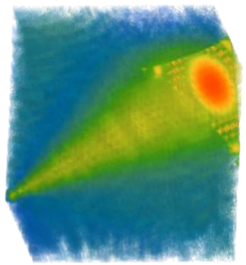
Our approach #2: **G**aze-**A**ware **3**D **C**ontext **E**ncoding

- Encode 3D context into egocentric, gaze-aware embeddings



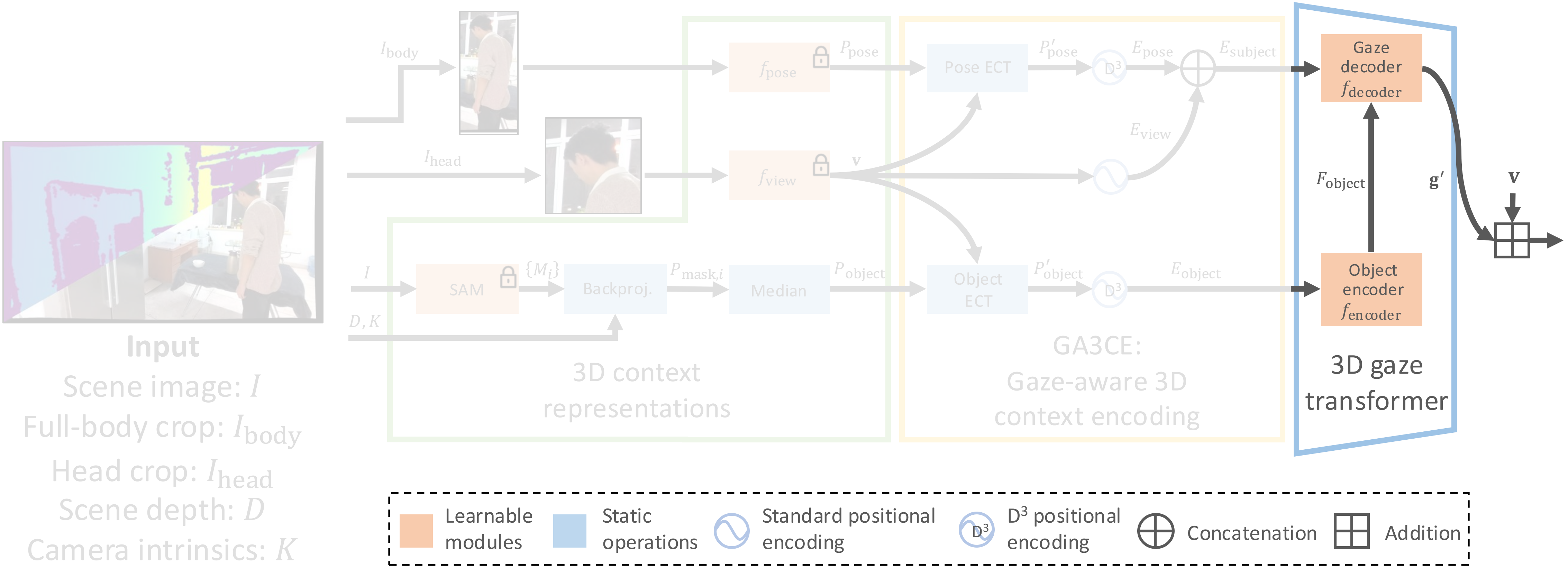
GA3CE #2: D^3 positional encoding

- Positional encoding (PE) for direction-distance-decomposed (D^3) embedding space

	Standard PE	**D^3 PE (Ours)**
Representation		
Formulation	$\gamma(\mathbf{p})$	$\tilde{\gamma}(\mathbf{p}) = \gamma(\frac{\mathbf{p}}{\lVert\mathbf{p}\rVert}) \oplus \gamma(\lVert\mathbf{p}\rVert)$
Correlation		
	Positional	Positional + **directional**

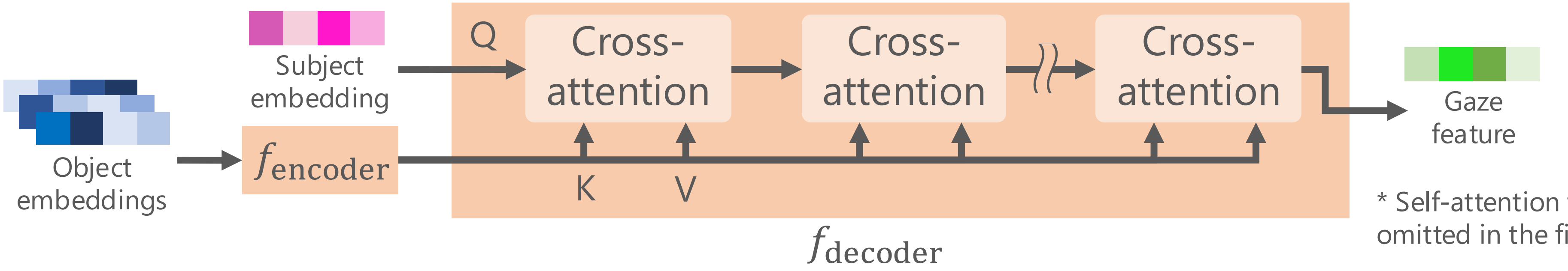
Our approach #3: 3D gaze transformer

- Learn spatial reasoning by cross-attention between subject and object embeddings

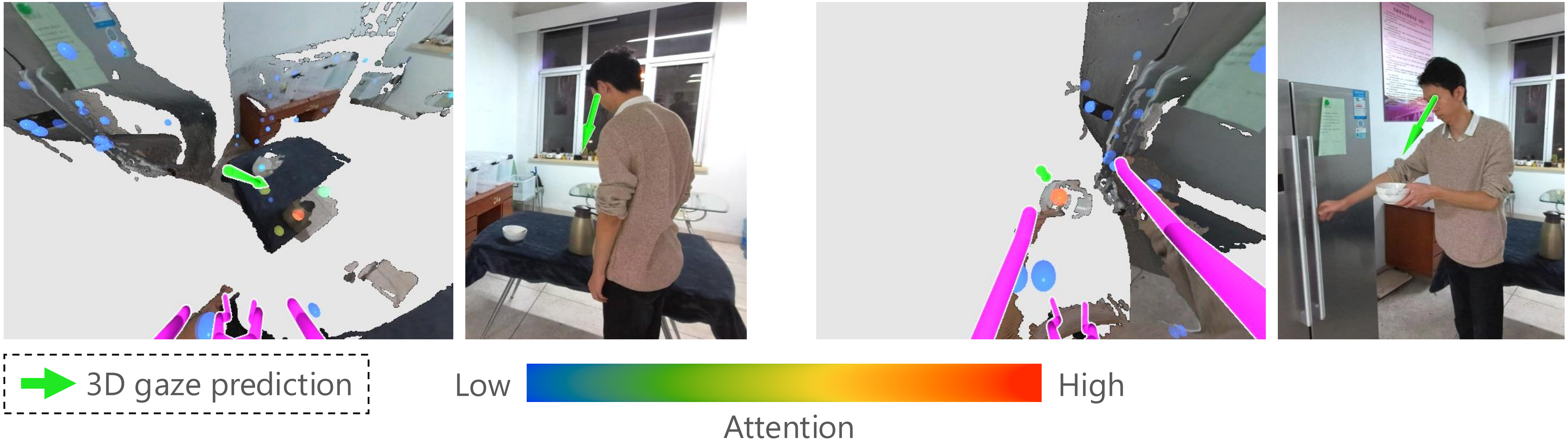


3D gaze transformer: Spatial reasoning

- Learn to leverage object positions relevant to the subject's pose and view direction in egocentric space

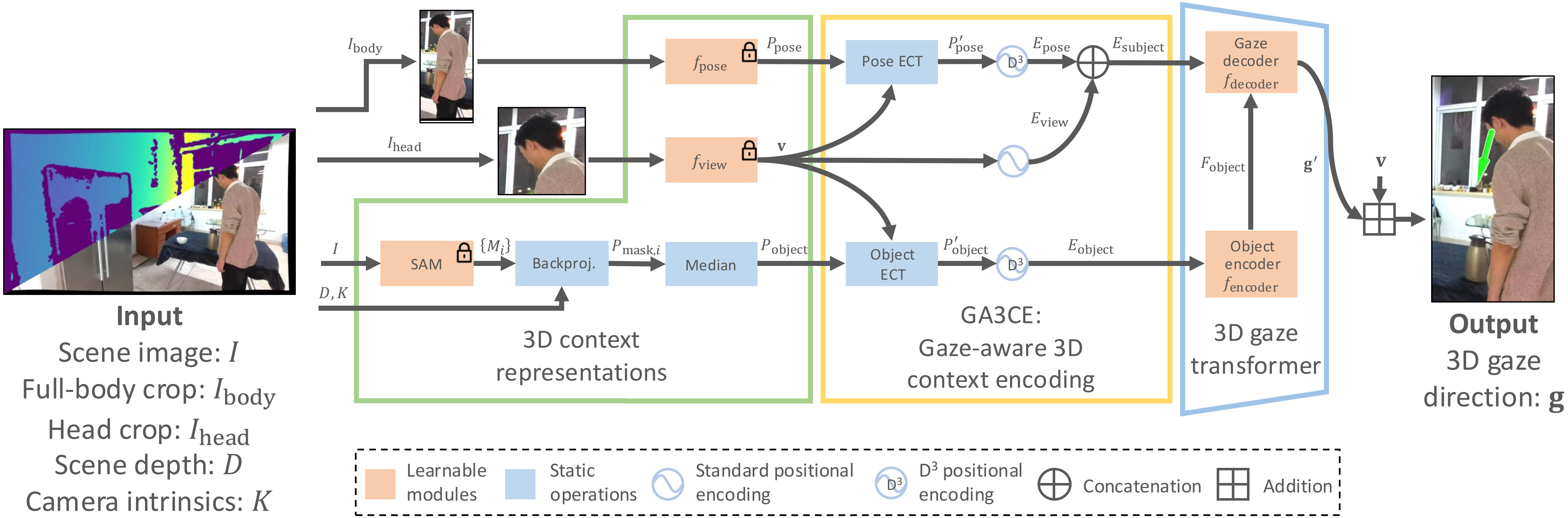


* Self-attention that precedes cross-attention is omitted in the figure for clarity.



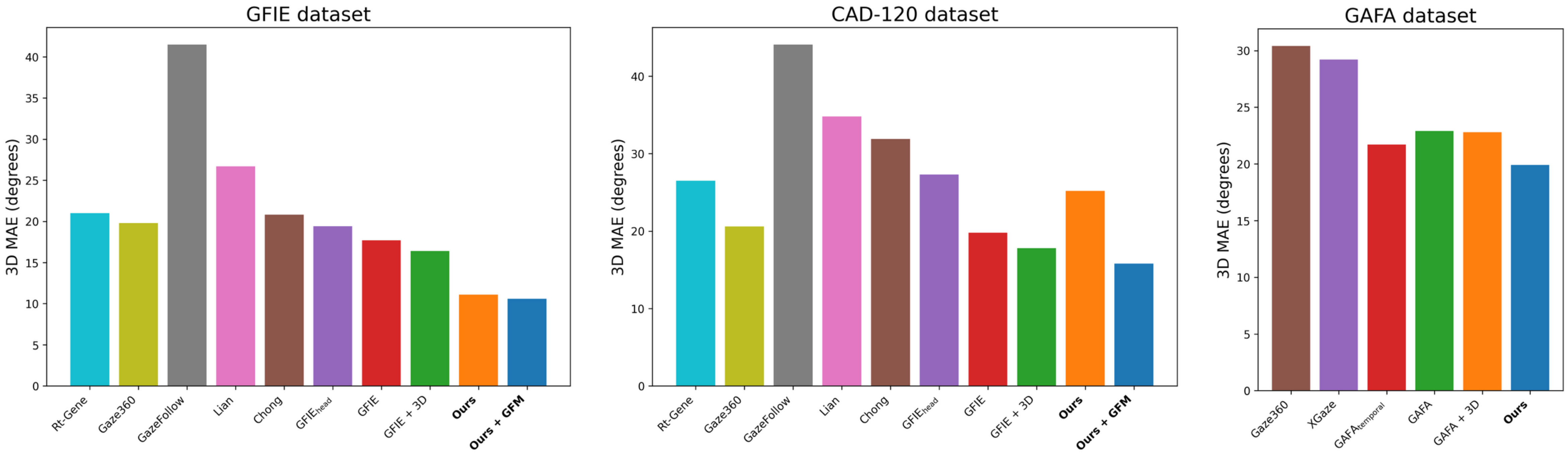
Our approach: Full pipeline

✓Learnable 3D understanding: geometrically grounded estimation
✓Normalization for unconstrained setting: simplified learning



Quantitative results

- 13%–37% improvement over leading baselines



Qualitative results

- Less viewpoint variability by egocentric transformation

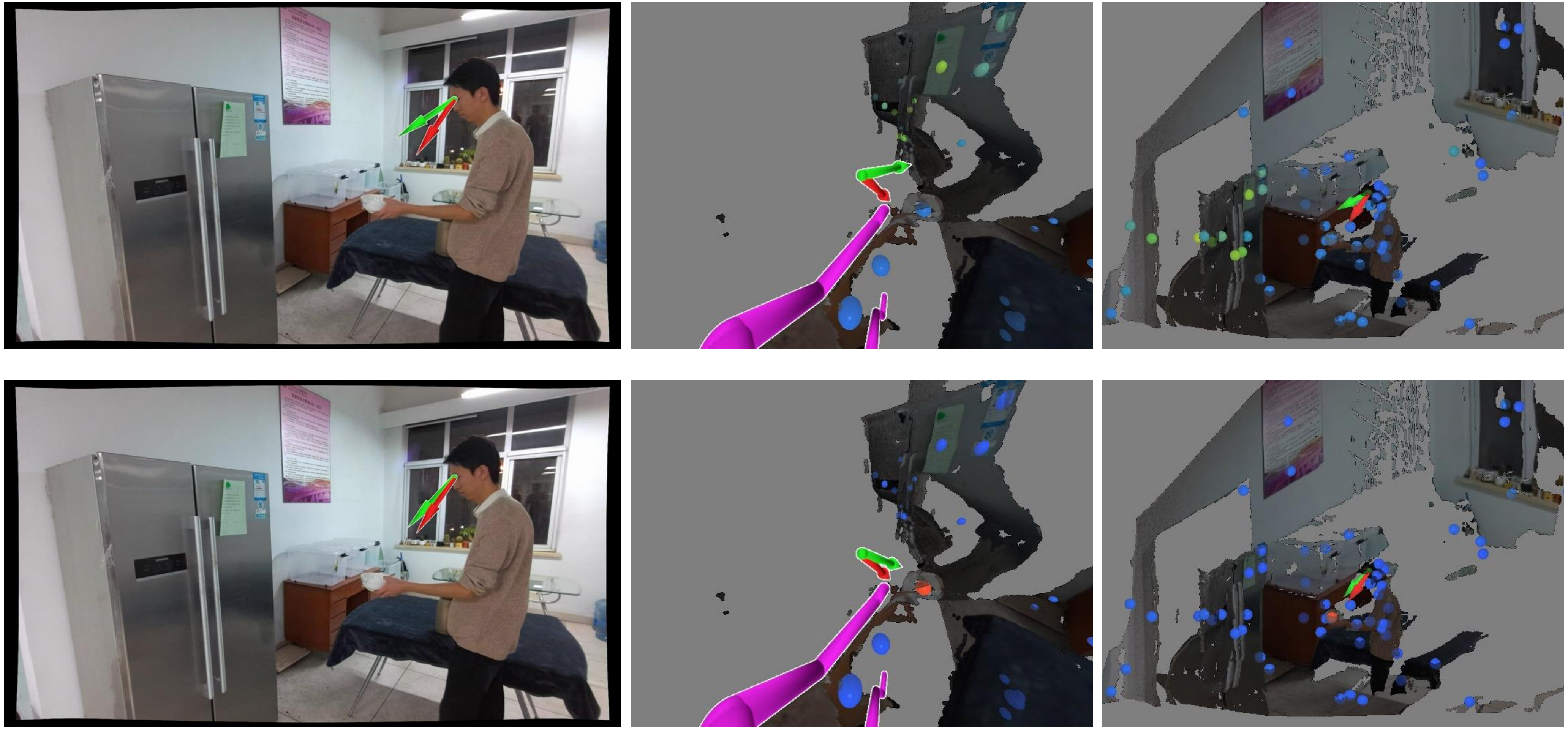
Camera view | Egocentric view | Orthographic view

Ablation results of GA3CE

- Improved spatial reasoning

Without GA3CE
Avg. MAE: 18.9
(on GFIE [1])

With GA3CE
Avg. MAE: **11.1**
(on GFIE [1])



Camera view | Egocentric view | Orthographic view

[1] Hu et al. GFIE: A Dataset and Baseline for Gaze-Following From 2D to 3D in Indoor Environments. CVPR 2023.

Conclusion

- We propose a novel 3D gaze estimation approach for the unconstrained setting
- Gaze-Aware 3D Context Encoding (GA3CE) enhances geometrically grounded estimation
- Our method outperforms leading baselines on three benchmark datasets by **13%–37%**

Project page: woven-visionai.github.io/ga3ce-project