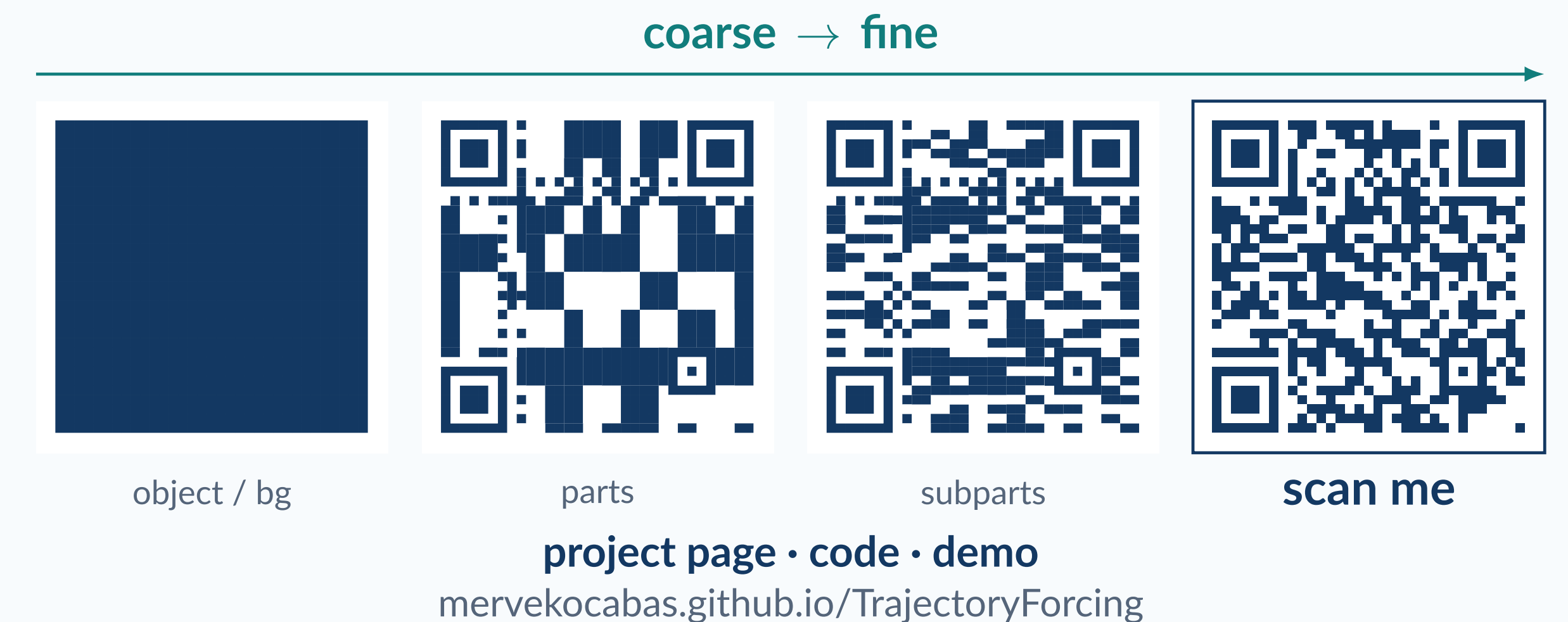


Trajectory Forcing

Structure-First Generation with Controllable Semantic Trajectories

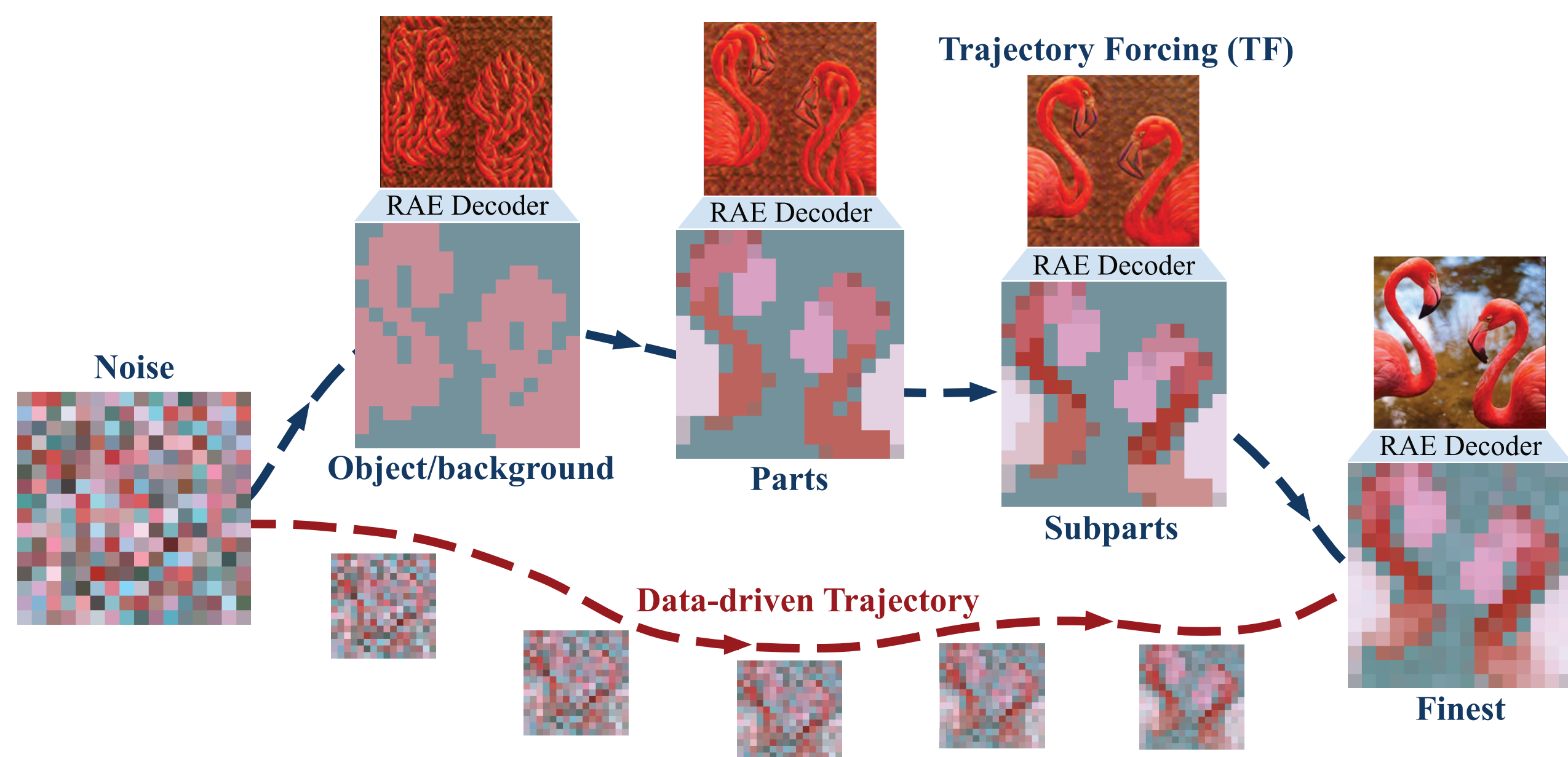
Merve Kocabas^{1,3,5} Gege Gao^{1,2,4,†} Bernhard Schölkopf^{3,4,5,6,7,‡} Andreas Geiger^{1,2,6,7,‡}

¹University of Tübingen ²Tübingen AI Center ³Zuse School ELIZA ⁴ETH Zürich ⁵MPI for Intelligent Systems ⁶ELLIS Institute ⁷KE:SAI [†]Project lead [‡]Shared last authorship



1 The path is hidden

Modern generators map noise to an image in *one opaque sweep*. Intermediates exist only as by-products: unstructured, uninspectable, discarded once the output appears.

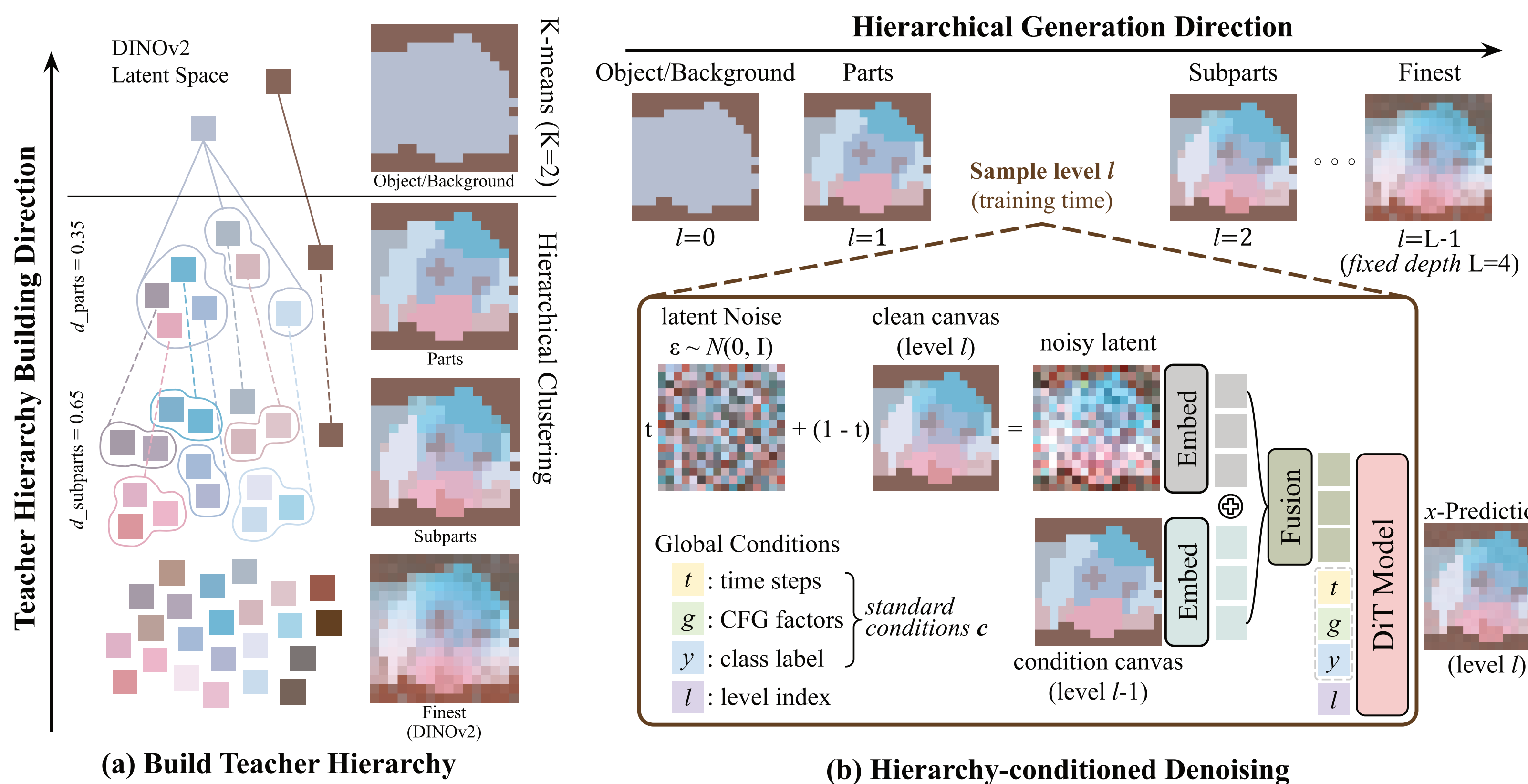


TF (top, blue) replaces that trajectory with a coarse-to-fine progression through **object/background** → **parts** → **subparts** → **finest**. Every intermediate decodes through a *shared* RAE decoder into an image, editable before the next begins.

2 Contributions

- **Semantic Trajectories**: Generation becomes an explicit, decodable coarse-to-fine process rather than an opaque noise-to-image mapping.
- **Efficient Trajectory Learning**: Annotation-free semantic hierarchies are learned with a shared one-step flow model, requiring one evaluation per level.
- **Controllable and Measurable Generation**: Intermediate states enable localized editing, while LIS and PMR quantify edit locality and hierarchical consistency.

3 Method



(a) A teacher hierarchy clusters DINOv2 features into level canvases. (b) One shared network trains over all levels, conditioned on the level below.

(i) Flow loss: $\mathcal{L}_{\text{flow}} = \mathbb{E}_{t,r,l} \left[\left\| \mathbf{V}_\theta - \mathbf{v}_g \right\|^2 \right]$.

Prediction type: $\hat{\mathbf{x}}_g(\mathbf{z}^{(l)}(t), l, \mathbf{z}^{(l-1)}; \mathbf{c})$

Average velocity: $\mathbf{u}_\theta = \frac{\mathbf{z}^{(l)}(t) - \hat{\mathbf{x}}_g}{t}$

Compound prediction: $\mathbf{V}_\theta = \mathbf{u}_\theta + (t-r) \cdot \text{JVP}_{\text{sg}}$

Overall objective: $\mathcal{L} = \mathcal{L}_{\text{flow}} + \lambda \mathcal{L}_{\text{struct}}$.

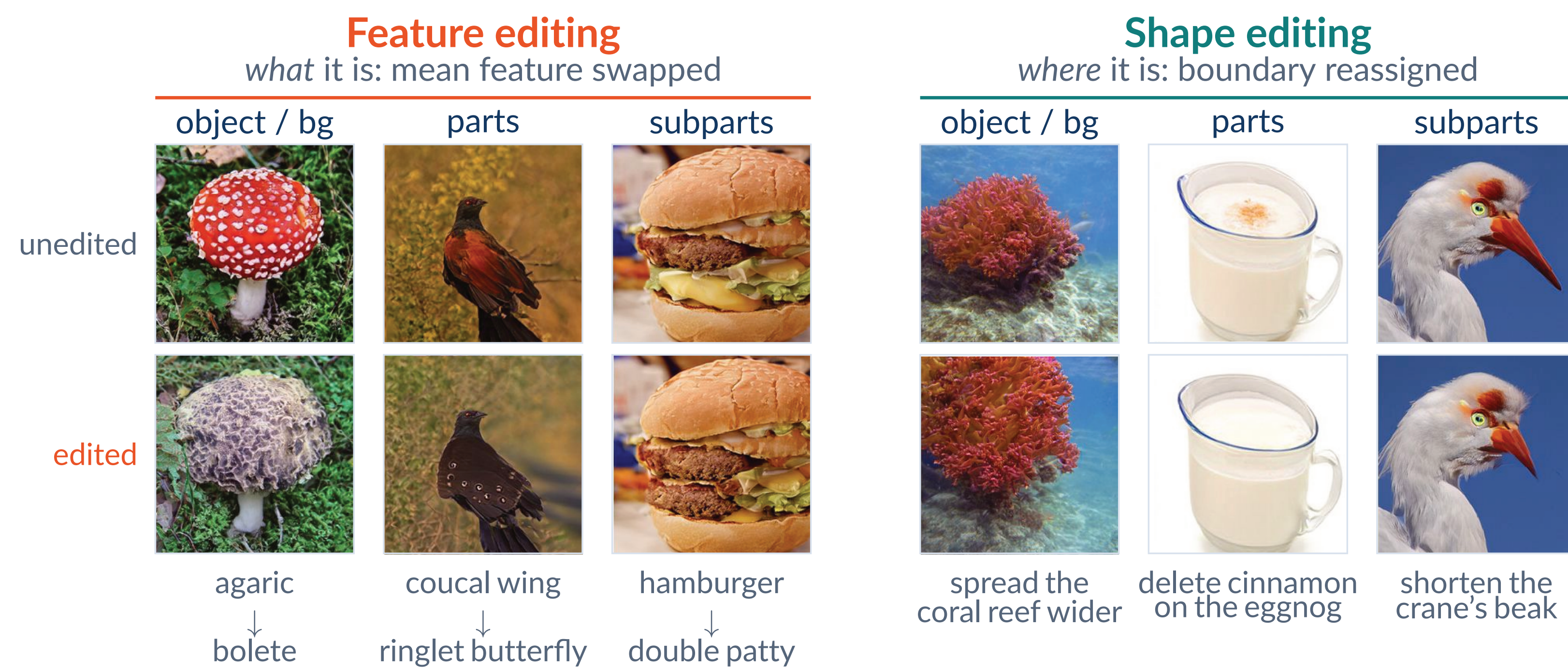
(ii) Structural loss:

Penalizes the deviation of each predicted token from the ground-truth mean $\mu_k^{(l)}$ of its assigned region.

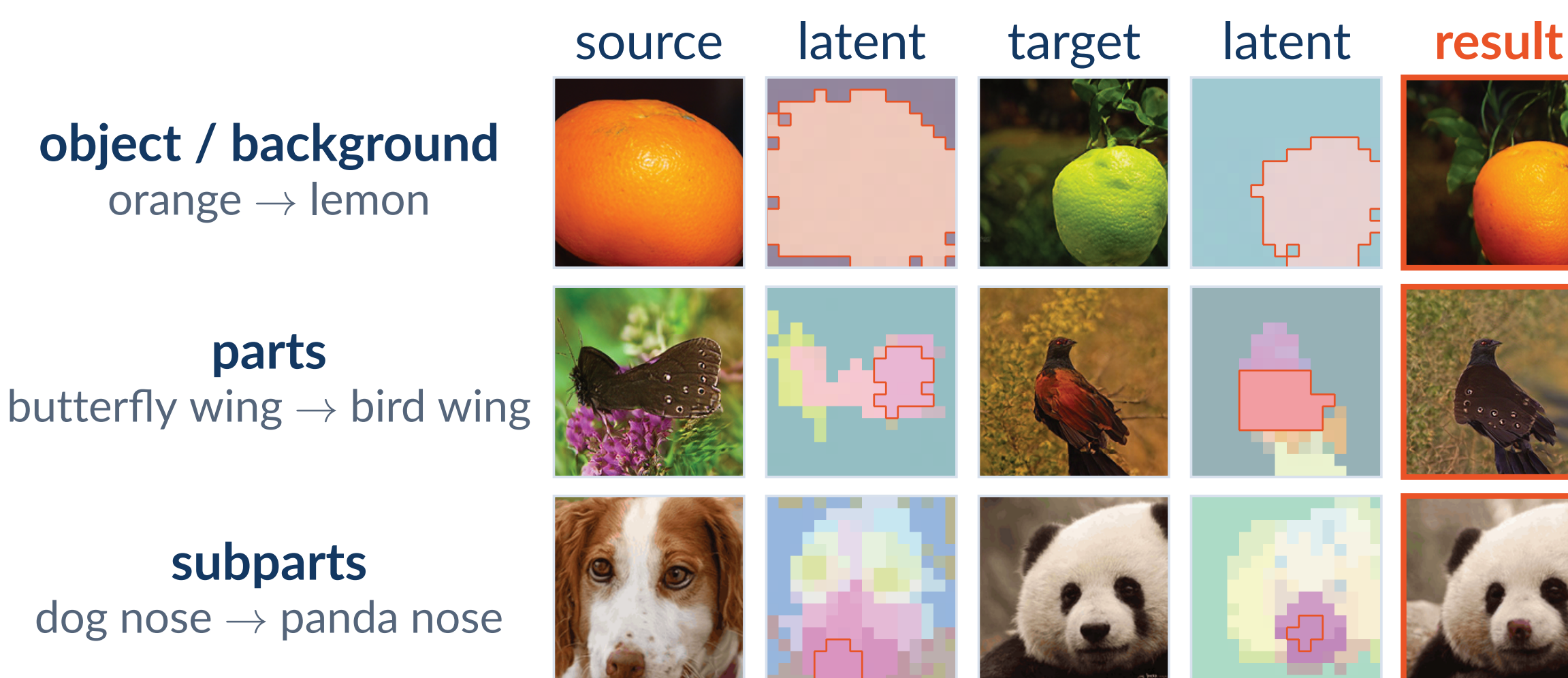
$$\mathcal{L}_{\text{struct}} = \frac{1}{K_l} \sum_{k=1}^{K_l} \frac{1}{|R_k^{(l)}|} \sum_{i \in R_k^{(l)}} \left(1 - \cos \left\langle \mathbf{x}_{\theta,i}, \mu_k^{(l)} \right\rangle \right)^2$$

where K_l is the number of valid regions.

4 Editing



Both operations, at every level. Each pair re-generates from the canvas edited at that level — coarse edits restate the scene, subpart edits stay local, and shape edits move a boundary without changing what it depicts.



The **outlined** tokens are copied out of the source's canvas and into the target's, which is then continued to generate subsequent levels.

5 Results



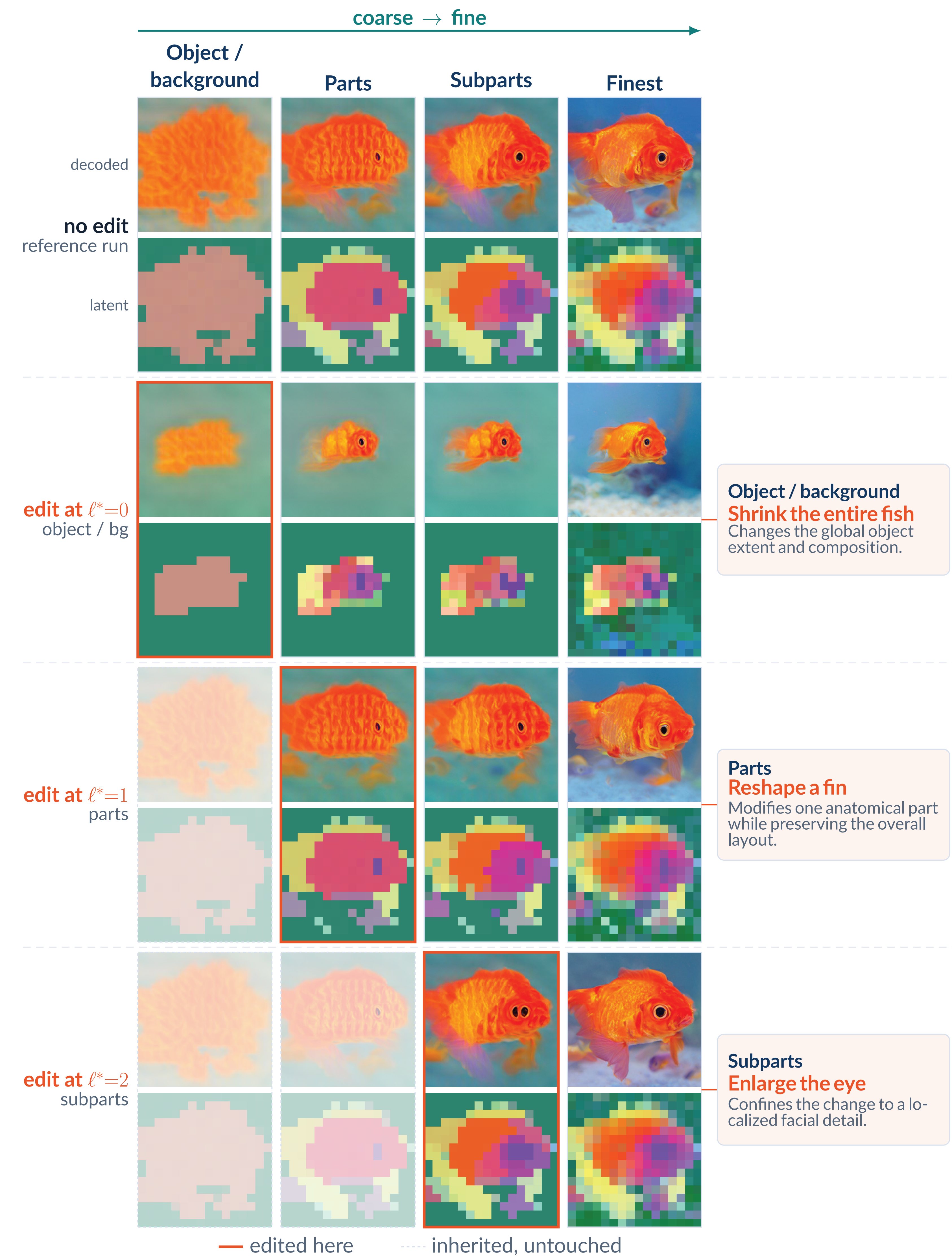
Generated trajectories. Four samples, four levels each, decoded output above its PCA latent. Every stage is already a recognisable image.

Method (ImageNet 256 × 256)	Space	NFE	Params	Ep.	FID ↓
JIT-L/16 (Heun)	pixel	100×2	459M	600	2.36
RAE + DiT-XL/2	DINOv2	50×2	839M	800	1.13
NVG-d20	VQ-VAE	184*	497M	250	2.44
pMF-L/16	pixel	1	411M	320	2.52
IMF-L/2	VAE	1	409M	640	1.86
TF-B/16	DINOv2	4×1	177M	80	7.93
TF-B/16 + FD [‡]	DINOv2	4×1	177M	80	2.52
TF-L/16	DINOv2	4×1	515M	80	6.38
TF-L/16 + FD [‡]	DINOv2	4×1	515M	80	2.02
TF-H/16	DINOv2	4×1	1.1B	80	5.97
TF-H/16 + FD [‡]	DINOv2	4×1	1.1B	80	1.98

*NVG needs 9 + 7×25 = 184 NFE.

[‡]Inception-space FD-loss post-training.

6 Scope control



Generation is Markov, so an edit at ℓ^* leaves every coarser canvas *identical* and re-lays out only what follows, at its own scale — the latent is what makes that checkable by eye. The editing level sets the breadth of the change.

7 Conclusion

Structure need not cost quality. Four semantic levels, one evaluation each, keep TF competitive on ImageNet while every intermediate stays a legible, editable image. Control comes from the process, not a probe on a finished model: an edit inherits its level's *scope*.

Limitations: Fixed level depth and a centre-prior split suit centred single-object scenes best; each level sees only its predecessor.

Next: Learned hierarchies (SAM), a per-image number of levels set by how much detail the image carries, and text conditioning — so an edit's scope can be decided between *prompting* and navigating the latent.