MIRU 2026 · RESEARCH PRESENTATION

Dual-Stream Discrete Geometric Gating Network for Alleviating VAE Coupling in Image Editing

Peilin Xiong · Honghui Yuan · Junwen Chen · Keiji Yanai

The University of Electro-Communications

A coarse mask should **localize**
not shape.

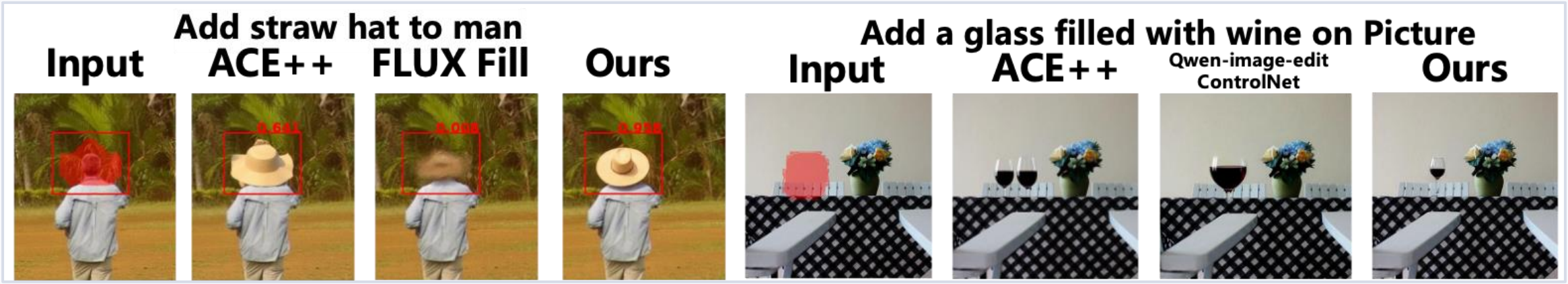
Preserve the background. Free the subject geometry.



Input · coarse support · generated result

Coarse masks localize—but should not shape

A rough spatial hint becomes a geometric prior when its contour enters the generator too literally.



Hint

Users sketch approximate support—not silhouettes.

Bias

Literal contour conditioning creates truncation, boxes, and boundary artifacts.

Goal

Preserve the scene while the subject follows the instruction's natural geometry.

[1] Black Forest Labs, “FLUX.1 Fill [dev],” 2024.
[2] C. Mao et al., “ACE++: Instruction-Based Image Creation and Editing via Context-Aware Content Filling,” ICCV Workshops, 2025.
[3] Z. Duan et al., “Diffusion Templates: A Unified Plugin Framework for Controllable Diffusion,” arXiv:2604.24351, 2026.

The task is a Two-Zone Constraint

Preserve everything outside the edit while allowing natural geometry inside it.

Background zone

$\Omega b = 1 - M$

Preserve appearance, texture, and layout.



Editing zone

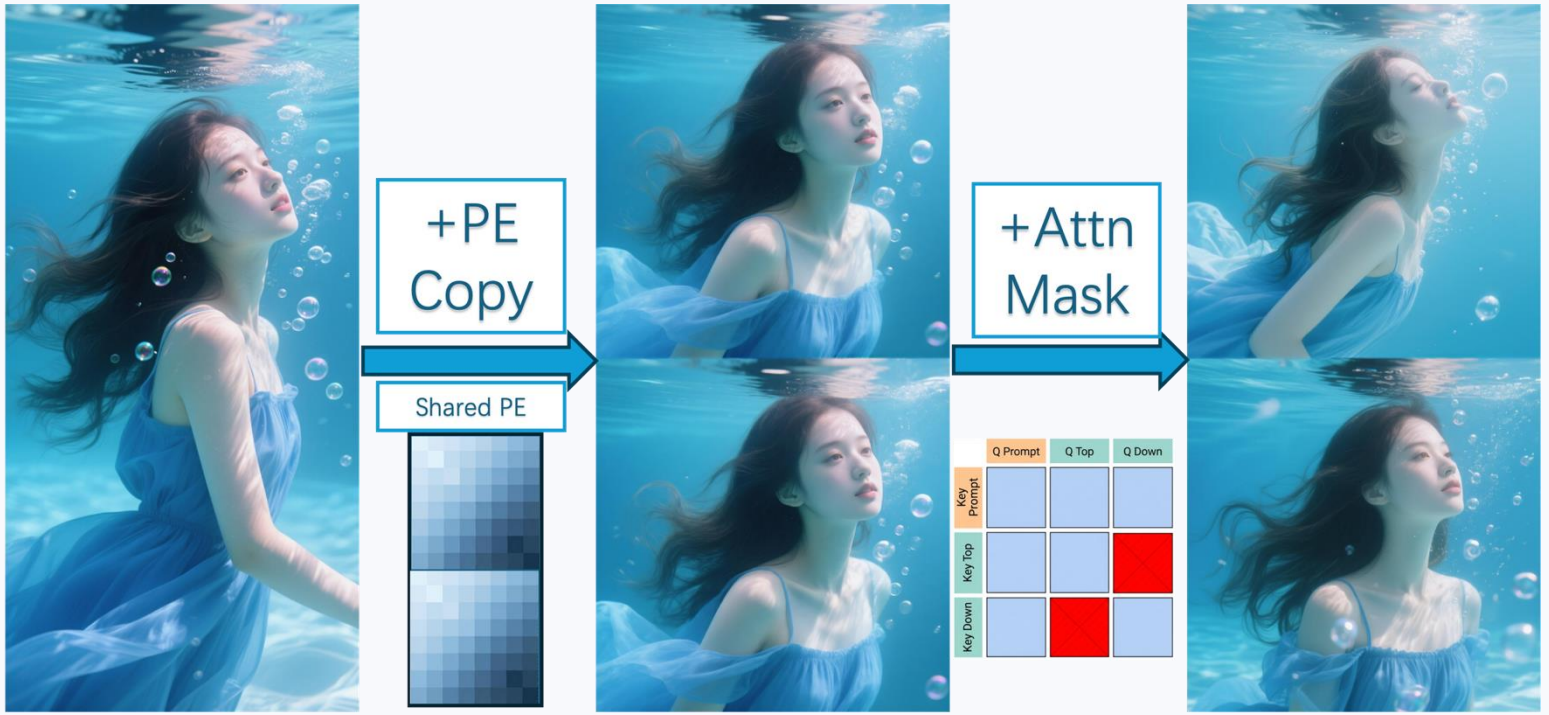
$\Omega e = M$

Follow the instruction; do not copy the contour.

The mask is a localization hint—not a target boundary.

Positional embeddings select reusable context

1. **Same PE**
   same context cue
2. **Aliasing**
   duplicate content
3. **Suppress**
   limit cross-region reuse



Positional embeddings select reusable context

1. **Same PE**
   same context cue

2. **Ignore the Prompt**
   prompt control becomes weak

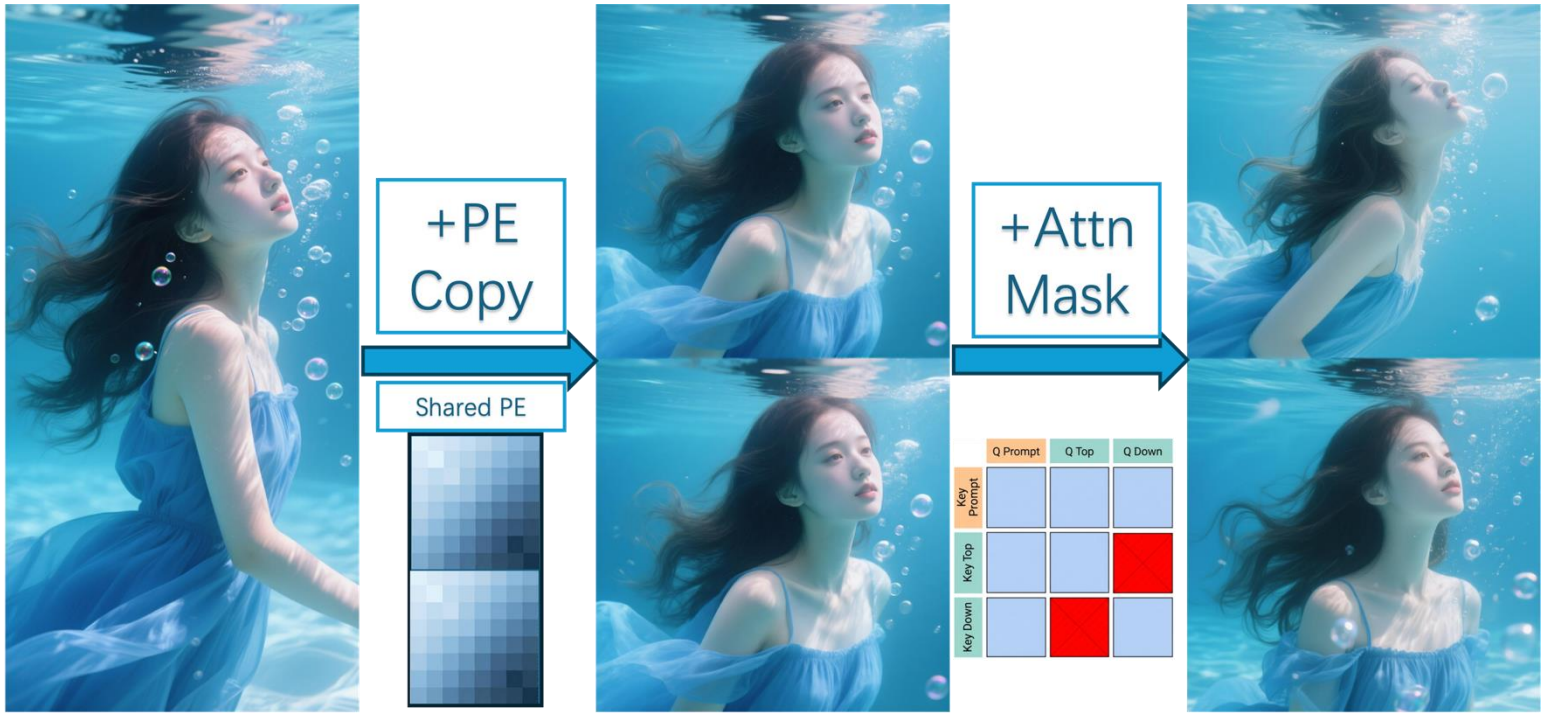
3. **Cannot Fit the Scene**
   lighting and appearance do not match



Xiong et al., “PosBridge: Multi-View Positional Embedding Transplant for Identity-Aware Image Editing,” BMVC 2025.

Positional embeddings select reusable context

1. **Same PE**
   same context cue
2. **Aliasing**
   duplicate content
3. **Suppress**
   limit cross-region reuse



BridgePath separates support from geometry

Different paths, one DiT forward pass.

1

Main Path

source trajectory
preserves context

2

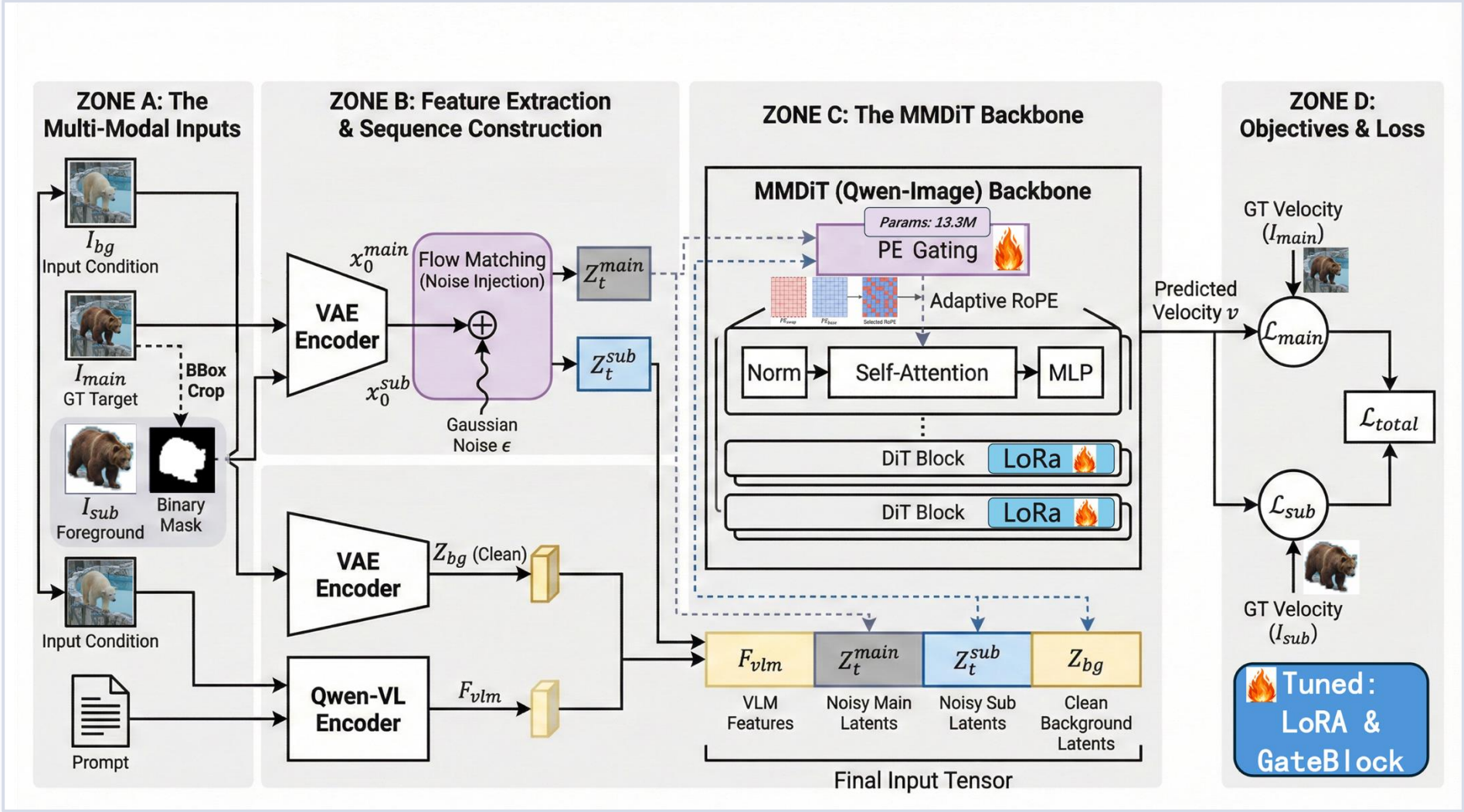
Subject Path

independent noise
inside edit bbox

3

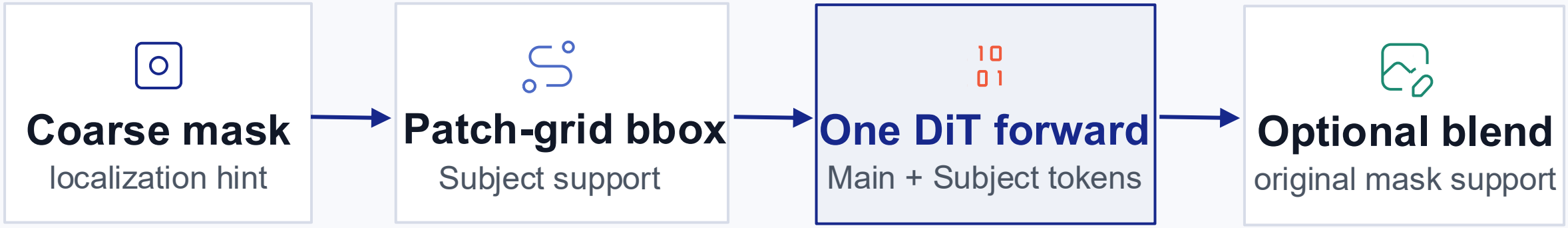
Discrete Gate

fusion ↔ freedom



Masks stay outside the DiT

But they remain in the workflow—and still define support.



FLOW-MATCHING TRAINING

Clean target
concatenate Main + Subject

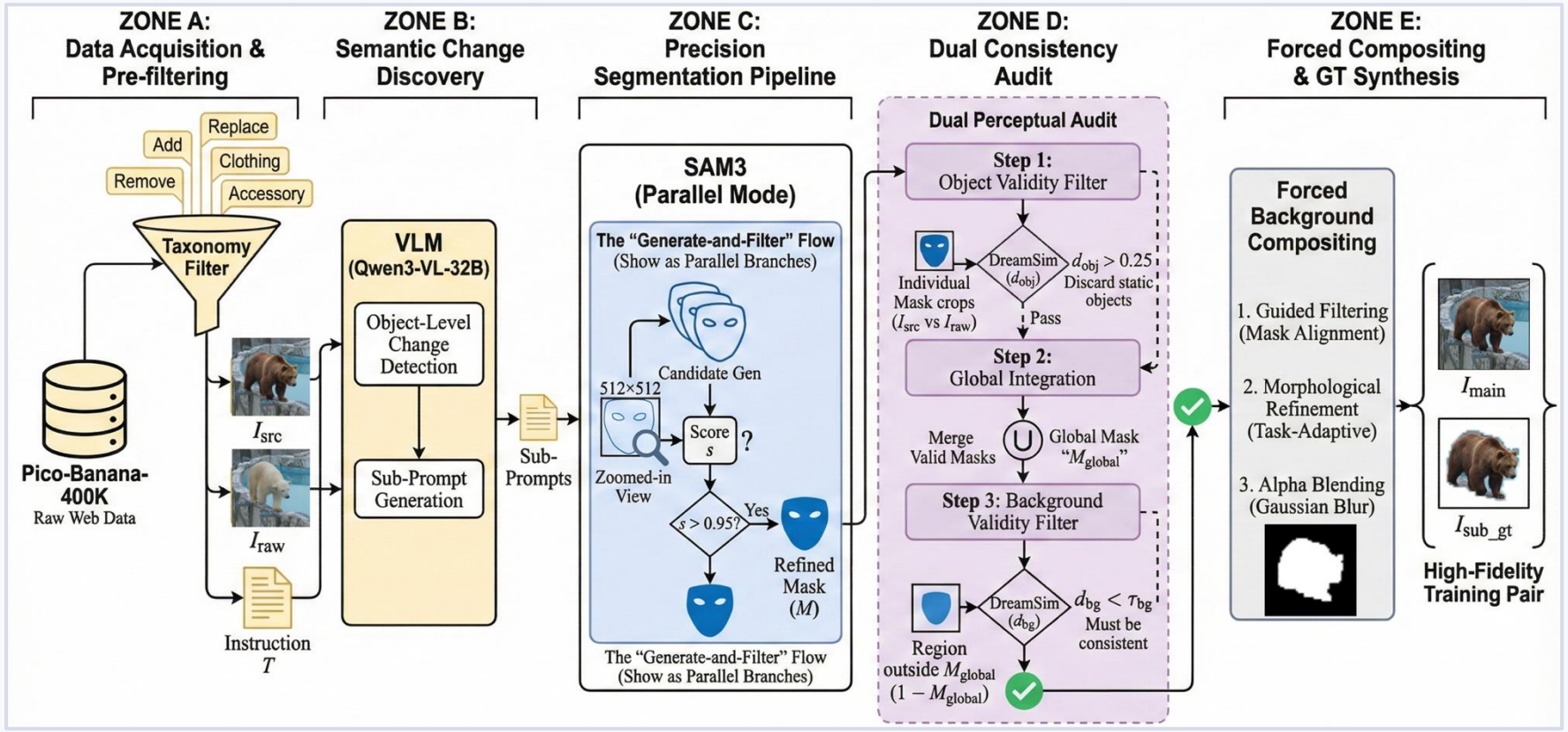
Independent noise
matching sequence dimensions

Optimize
rank-512 LoRA + GateBlocks

⚠ **Guardrail: masks still define support, perturbation, and blending.**

Training pairs must change the subject

Not the scene: foreground change, background stability.



42,425
training pairs

1,444
held-out test pairs

No evaluation pair is used for training.

VLM discovery → SAM masks → DreamSim filtering → forced compositing → mask perturbation

Local precision improves

Without sacrificing global quality.



LOCAL SIGLIP2-T ↑

0.503

text alignment

higher is better

LOCAL DINO ↑

0.902

feature identity

higher is better

LOCAL DREAMSIM ↓

0.175

perceptual error

lower is better

LOCAL SIGLIP2-I ↑

0.907

image similarity

higher is better

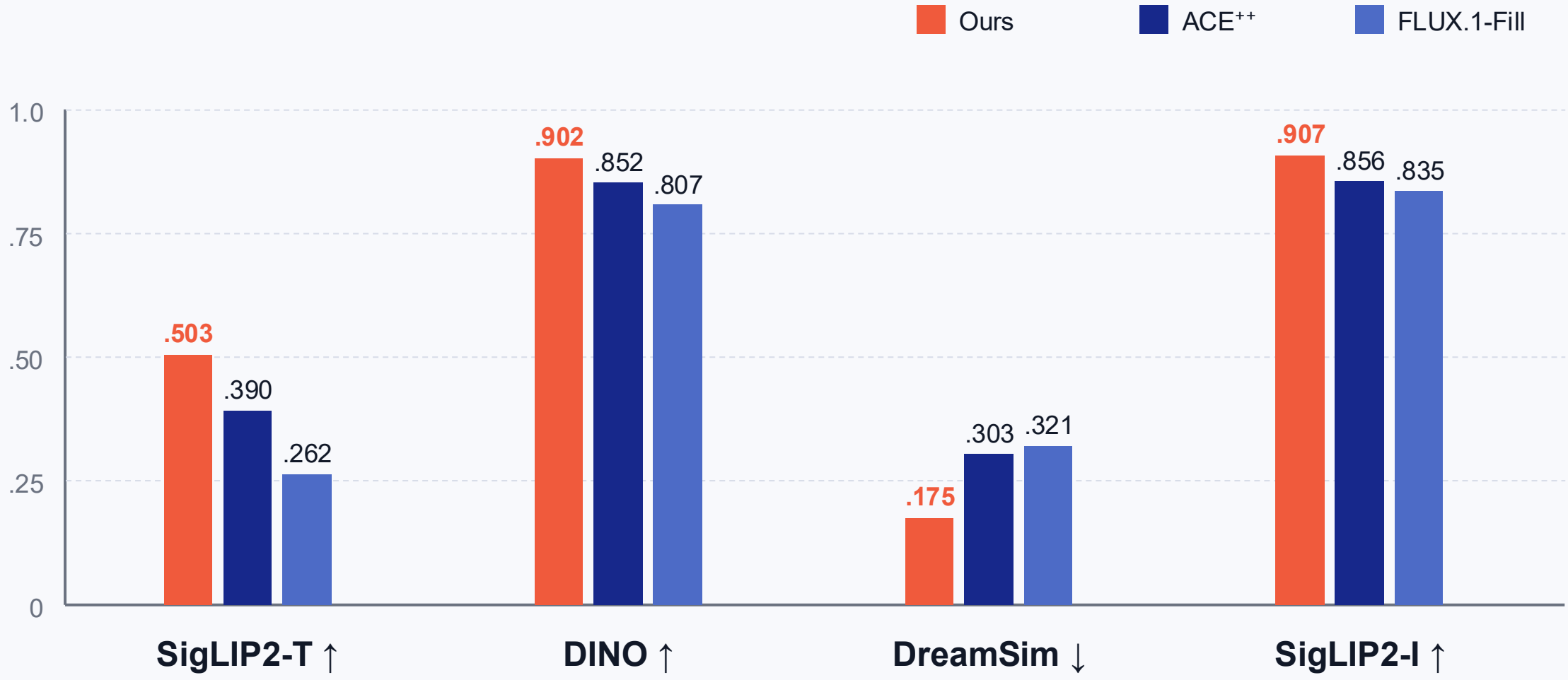
GLOBAL QUALITY · OURS-BENCH

DINO 0.988 **DreamSim 0.054** **SigLIP2-I 0.965** **SigLIP2-T 0.776**

Local metrics are computed inside the edit bounding box.

leads all four local metrics

Instruction alignment and perceptual quality improve inside the edit box.



⚠ Public systems are compared end to end; they were not retrained under one matched recipe.

The gains transfer

Three protocol-specific transfer evaluations are reported separately.

MagicBrush · Final Turn

First-turn input → final target

Method	L1↓	CLIP-I↑	CLIP-T↑
ACE++	0.081	0.891	0.307
FLUX.1-Fill	**0.075**	0.898	0.297
Q-Control	0.078	0.898	0.322
Ours	**0.076**	**0.907**	**0.324**

MagicBrush · Multi-Turn

Each turn uses the previous-turn GT

Method	L1↓	CLIP-I↑	CLIP-T↑
ACE++	0.047	0.936	0.306
FLUX.1-Fill	0.044	**0.947**	0.300
Q-Control	0.044	0.943	**0.315**
Ours	**0.043**	**0.946**	**0.315**

ICE-Bench · modified

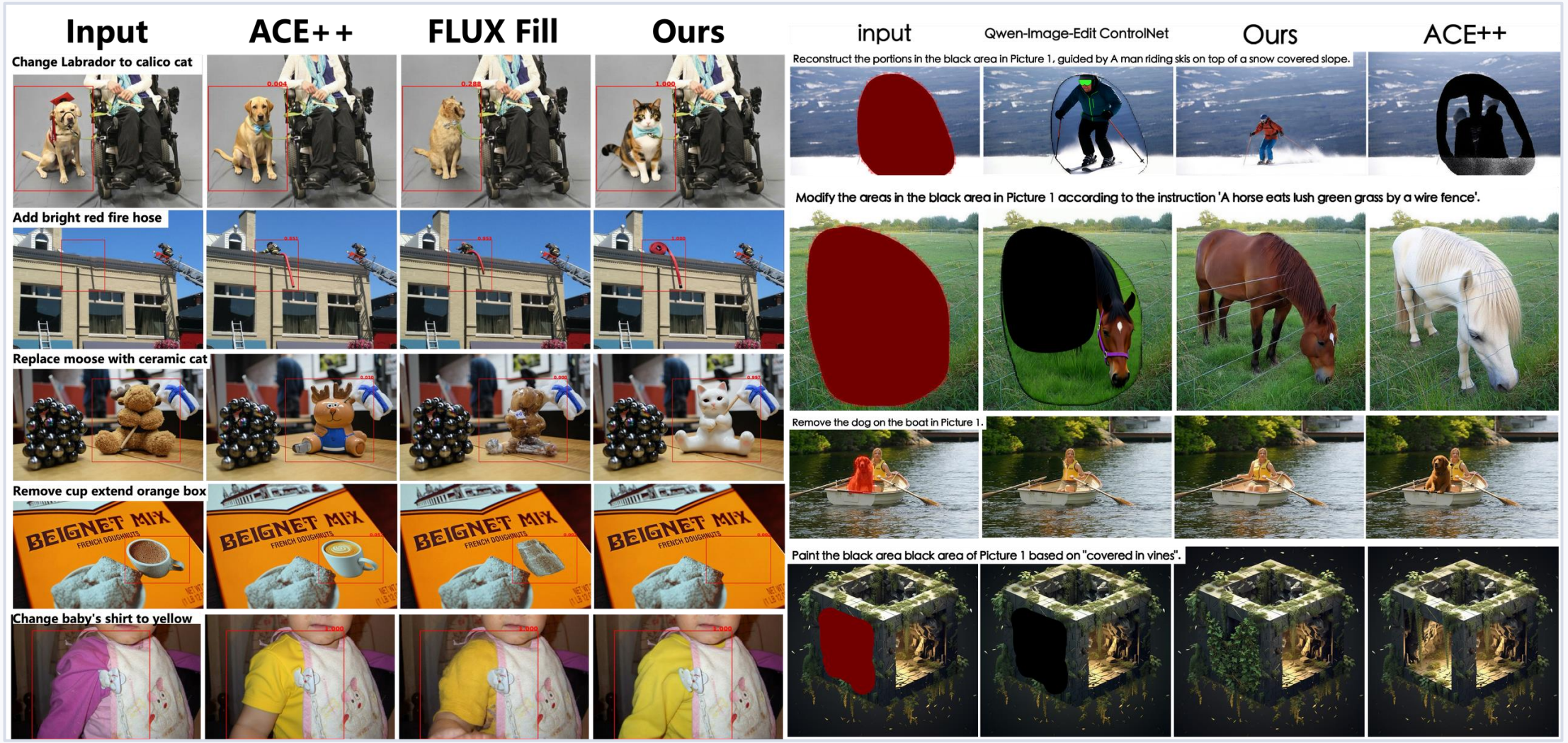
Modified prompt protocol

Method	AES↑	PF↑	SRC↑
ACE++	0.480	0.575	0.925
Q-Control	0.472	**0.588**	0.937
Ours	**0.494**	**0.582**	**0.942**

Final: first-turn input → final target · Multi: previous-turn GT input · Protocols are reported separately.

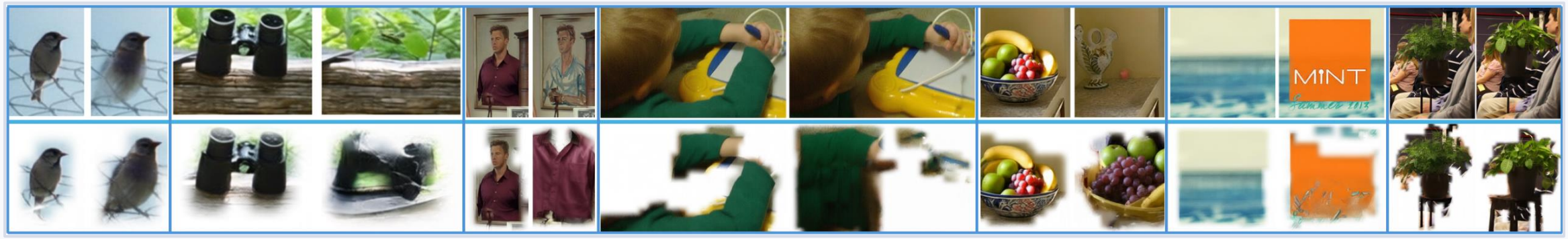
The subject follows the prompt

Not the mask contour.



Adaptive routing drives the local gains

The extra path alone is not enough.





Matched backbone, rank-512 LoRA, optimizer, data, 5×96GB, and 30K micro-steps. The clearest benefit is local precision.

Compact control, real runtime cost

The gate is lightweight; BridgePath is not free.

COMPACT CONTROL

13.31M

GateBlock parameters across 60 layers

≈1.13B reference branch

ControlNet-style copied block comparison

Rank-512 LoRA is fused before inference.

RUNTIME COST

25% edit region

latency and peak memory

Method	Latency	Peak VRAM
Ours	**111.649 s** (+21.38%)	**57.379 GiB** (+0.04%)
Q-Control	112.311 s (+22.10%)	59.482 GiB (+3.71%)

⚠ 13.31M describes the GateBlocks—not the complete trainable checkpoint.

THANK YOU

- For more details in potser：**OS1A-05**
- https://github.com/XIONGPEILIN/BRIDGE

A binary gate makes routing explicit

Each Subject token chooses scene fusion or geometric freedom.

1 **Predict**
token-level gate probability

$$p_i^l = \sigma\left(h_\phi^l(H^l)_i\right), \quad i \in \mathcal{S}_l$$

2 **Commit**
round to a hard routing decision

$$G_i^l = \mathrm{round}(p_i^l)$$

3 **Route**
select the effective PE

$$PE_{\mathrm{eff},i}^l = G_i^l\, PE_{\mathrm{base},i} + (1 - G_i^l)\, PE_{\mathrm{swap},i}, \quad i \in \mathcal{S}_l$$

G = 1
subject-centric · geometry freedom

G = 0
background-anchored · scene fusion