

Generative Editing

Images, art direction, and video

Harry Yang / 8 September 2026

Opening



One object, three creative decisions

Image editing

How can the material change while identity survives?

Art and moving images

How can a visual choice carry an idea through time?

The reflection is part of the edit

SOURCE



MATERIAL EDIT



Glass becomes ceramic. What should happen to the reflection?

Necessary change or accidental drift?

Visible difference	Interpretation	What to inspect
Amber glow disappears	Expected for opaque ceramic	Light transmission
Blue stem changes shape	Possible identity drift	Silhouette
Reflection changes	May be necessary	Material and geometry

The edit contract determines which differences count as errors.

Tonight's route

90 minutes

Spoken teaching across 123 slides

Part 1

How image editors work

Part 2

Art direction and iteration

Topic 3

Video editing and visual transformation

In addition

Presentations, break, and discussion pauses

Three decisions you can defend

- 01** Which control addresses this failure?
- 02** Which poster belongs in this exhibition?
- 03** Which moment would make you reject the clip?

PART 1

How image editors work

What changes? What survives?



Before Part 1: image editing

- 01** When glass becomes ceramic, what should stay the same?
- 02** Which parts of the scene must change with the material?
- 03** What could a model misunderstand in that request?

The edit contract



Change

The gallery becomes a forest.

Preserve

Pear identity, pedestal, framing.

Allow

Reflections and ambient light to adapt.

An invariant can be semantic or exact

Semantic preservation

The same pear remains recognizable.
Its appearance can adapt to new light.

Exact preservation

Specified pixels remain identical.
Copying or compositing can enforce this.

Editing as conditional generation

$$y \sim p_{\theta}(y \mid x, c, r)$$

x source image c edit instruction

r optional references y edited result

A plausible result can still violate the edit contract.

A condition is not a hard constraint

$$y \sim p_{\theta}(y \mid x, c, r)$$

accept(y) = contract checks pass

Many outputs can satisfy the same instruction.

A probability model does not certify preservation.

A smaller workspace: image latents

1024 × 1024 × 3

3,145,728 RGB values

64 × 64 × 64

262,144 latent values

12× fewer values

For this f16c64 example

A variational autoencoder (VAE) compresses images into latent representations.

Is the detail lost before the edit?

- 01** Encode the source with the VAE
- 02** Decode without requesting a change
- 03** Inspect tiny text, fine texture, and edges

Why token count matters

Illustrative tokenization	Token count N	Attention pairs N ²
64 × 64 latent locations	4,096	16,777,216
Group 2 × 2 locations	1,024	1,048,576

4× fewer tokens gives 16× fewer pair scores in dense self-attention.

Double the resolution: count the cost

2× each spatial axis

4× as many image tokens

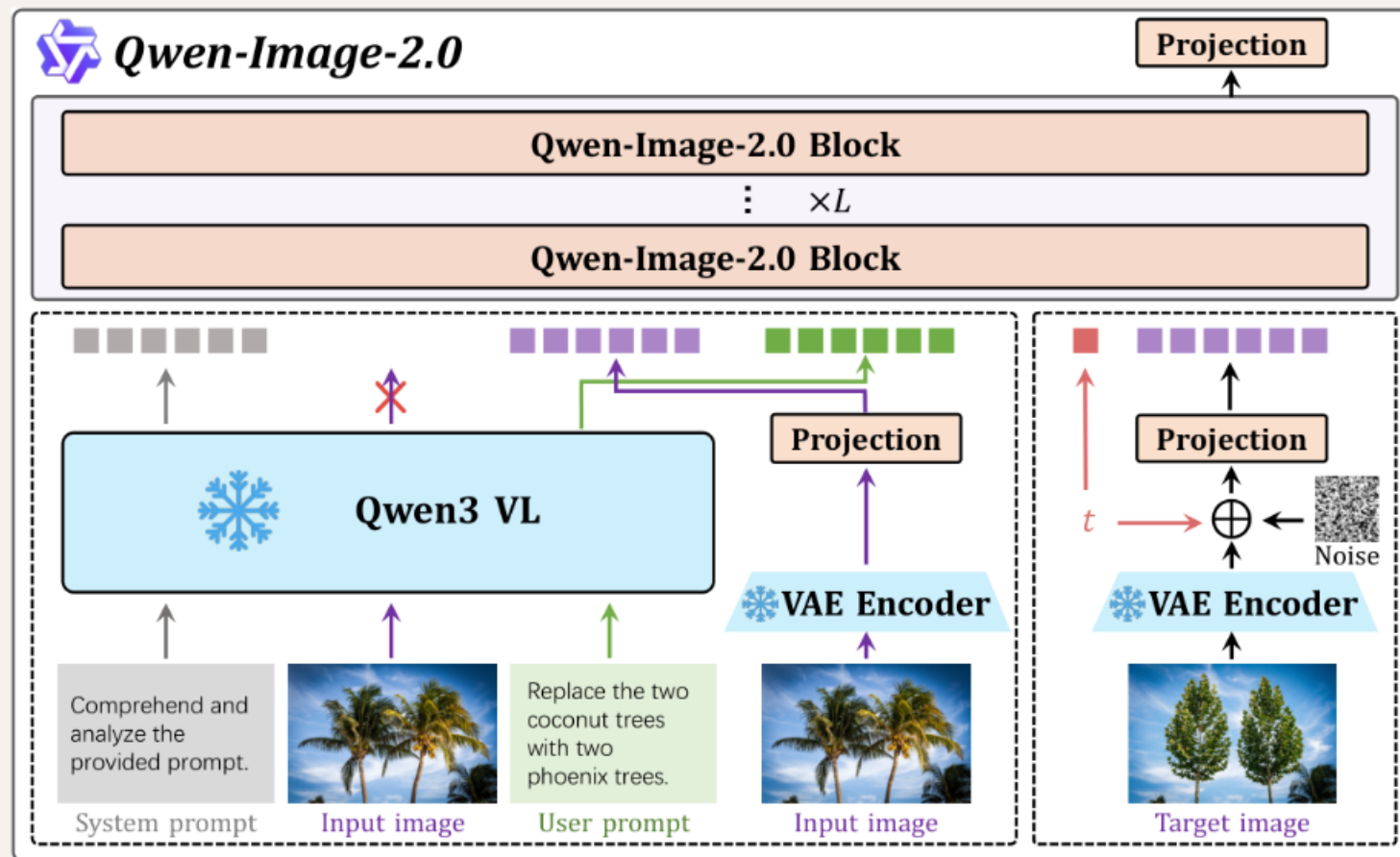
4× image tokens

16× dense attention pair scores

A component calculation

Measure whole-system runtime.

Inside a current image editor



The missing target at inference

Training example

Source + request + edited target
A known target supplies supervision.

Inference request

Source + request + references
The edited target must be generated.

Two representations of the source

Semantic features

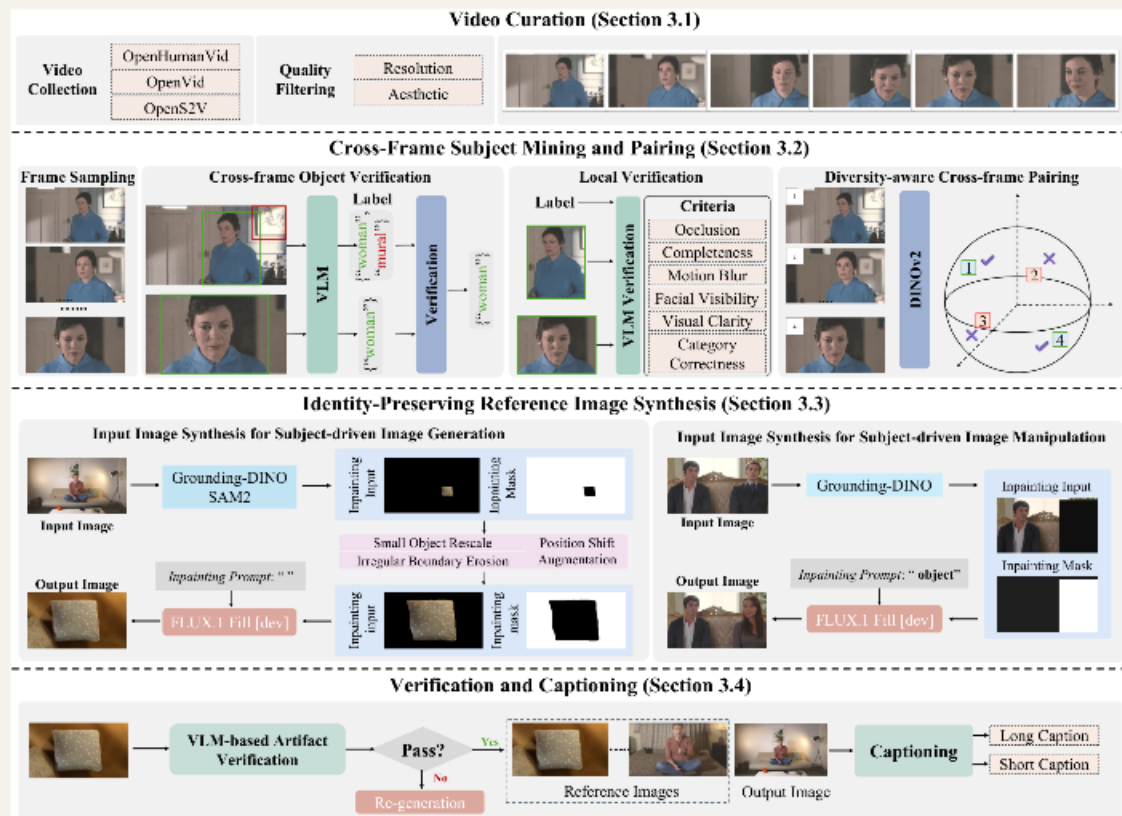
What is present?
Which object does “it” refer to?
Which change does the prompt request?

Visual latents

What does it look like?
Where are its shapes and textures?
Which details help retain identity?

Understanding the request and retaining appearance are related tasks.

OpenSubject: identity across new scenes



Same identity,
varied viewpoints

Synthesize new
reference contexts

Verify identity
and edit fidelity

Video pairs teach what can change while identity persists.

Flow matching: the training target

$$\mathbf{z}_t = (1 - t) \boldsymbol{\varepsilon} + t \mathbf{z}_1$$

$$\mathbf{v}^* = \mathbf{z}_1 - \boldsymbol{\varepsilon}$$

$\boldsymbol{\varepsilon}$ is noise. $\mathbf{z}_1$ is the target image latent.

The model learns a velocity at intermediate t .

Teaching convention: $t = 0$ is noise, $t = 1$ is data.

The interpolation endpoints

$$\mathbf{z}_0 = (1 - 0) \boldsymbol{\varepsilon} + 0 \mathbf{z}_1 = \boldsymbol{\varepsilon}$$

$$\mathbf{z}_1 = (1 - 1) \boldsymbol{\varepsilon} + 1 \mathbf{z}_1$$

Substitute the endpoints before reasoning about the path.

Predict the next sampling step

$$\mathbf{z}_{t+\Delta t} = \mathbf{z}_t + \Delta t \cdot v\theta(\mathbf{z}_t, t, \mathbf{x}, \mathbf{c})$$

$$0.20 + 0.10 \times 0.60 = ?$$

A scalar example of an Euler update.

Repeat updates, then decode the final latent into pixels.

A better solver cannot clarify the brief

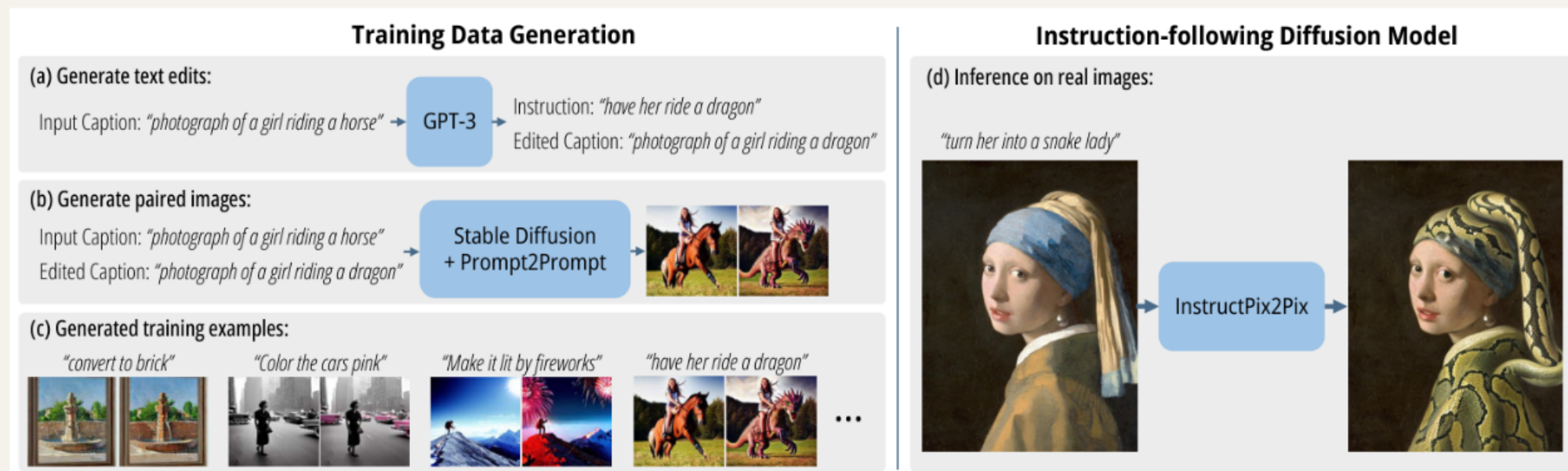
Numerical question

Does the solver follow
the learned field accurately?

Editing question

Does that learned field
produce the intended change?

Where edit supervision comes from



A training example pairs a source, an instruction, and an edited target.

InstructPix2Pix used synthetic edit data to teach this relationship.

What a mismatched training pair teaches

Instruction	Unwanted target change	Possible learned association
Make it ceramic	Camera moves closer	Material implies reframing
Replace the background	Subject becomes larger	Setting implies scale change
Change the color	Object identity changes	Color implies replacement

Illustrative failure cases, not measured dataset statistics.

Classifier-free guidance (CFG)

$$\hat{\mathbf{v}} = \mathbf{v}_{\text{base}} + s (\mathbf{v}_{\text{cond}} - \mathbf{v}_{\text{base}})$$

$\mathbf{v}_{\text{base}}$ baseline prediction $\mathbf{v}_{\text{cond}}$ conditioned prediction

$s = 1$ recovers the conditioned prediction.

Larger s extrapolates. It can also amplify artifacts.

Guidance goes beyond the prediction

$$0.2 + 2 \times (0.5 - 0.2) = 0.8$$

$v_{\text{base}} = 0.2$, $v_{\text{cond}} = 0.5$, $s = 2$

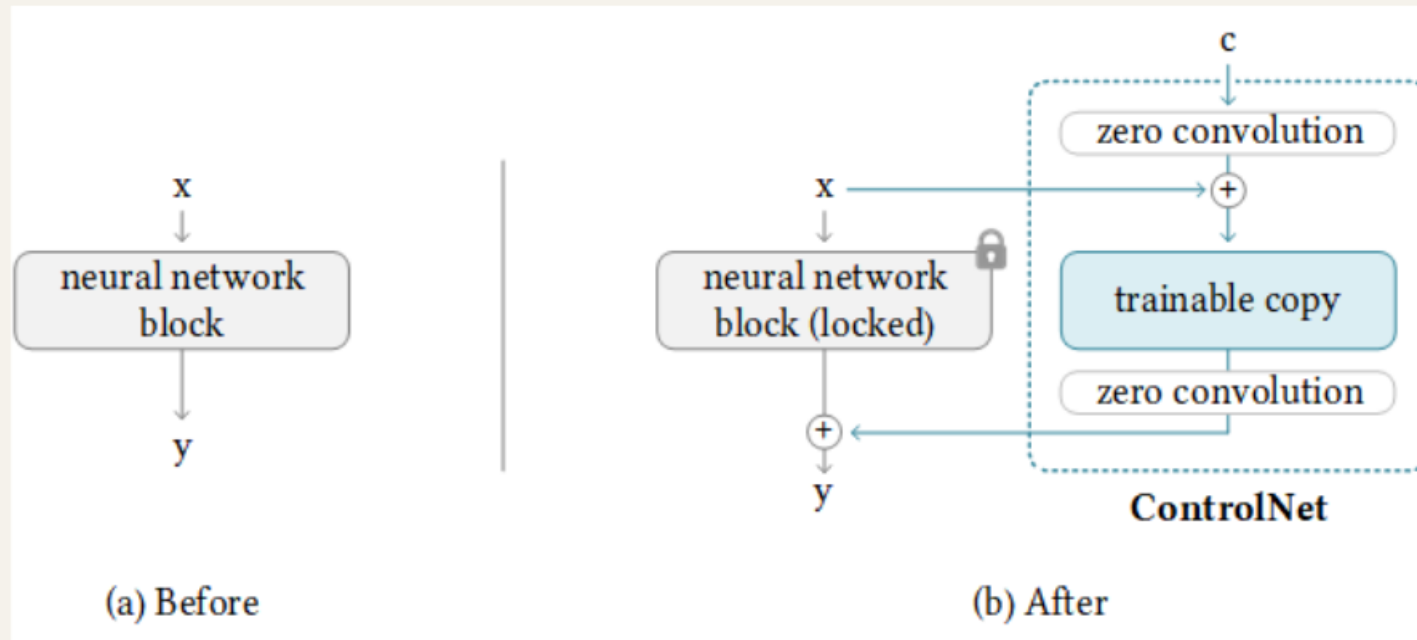
Extrapolation can strengthen useful changes and errors.

Different controls carry different information

Control	What it supplies	Useful when...
Instruction	Requested semantic change	The object or action is clear
Mask	Region eligible for editing	Locality matters
ControlNet	Spatial cues: edges, depth, pose	Pose or layout must survive
Reference image	Appearance or identity cues	Words are insufficient

Availability and exact behavior depend on the editor.

ControlNet: add spatial conditioning



Freeze the pretrained diffusion model.

Train a copied branch with a spatial condition.

Zero-initialized 1×1 convolutions connect the paths.

Multiple references need distinct identities

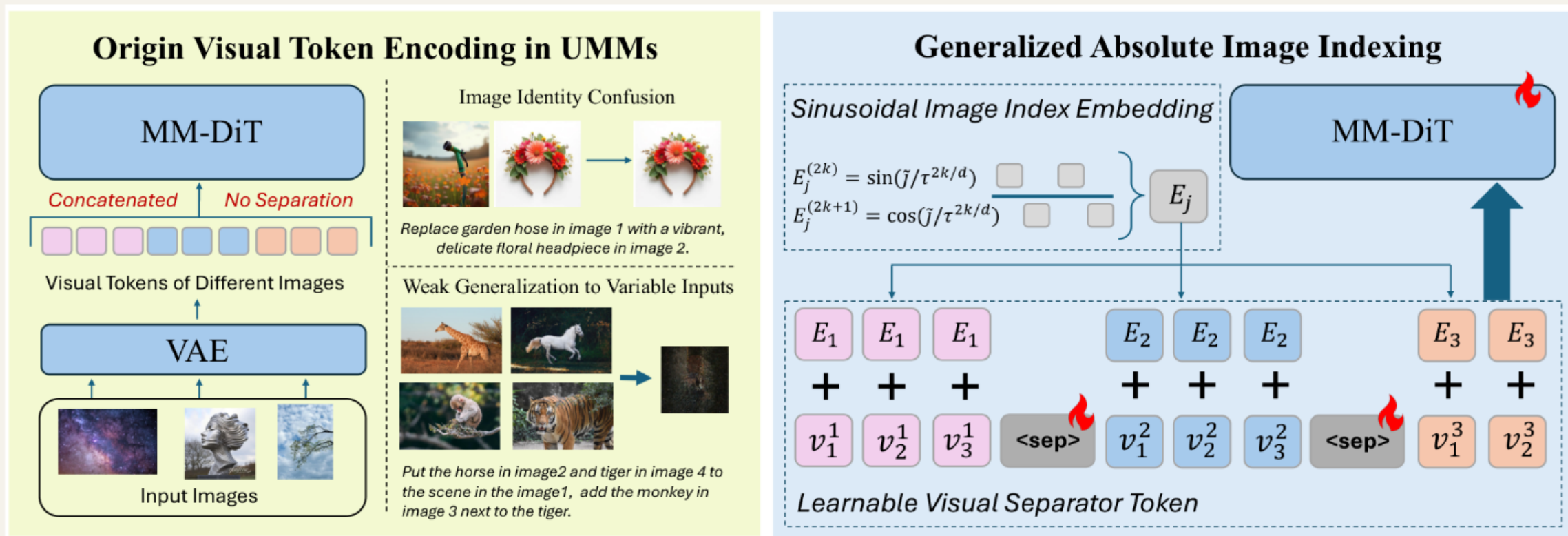


Image index: which reference? Separator: where does it end?

A reference-role failure is visible in the output

Input	Instruction	Ours	DreamOmni2	Qwen-Edit	OmniGen2
	<i>Make the surface of the car in image 1 look like the reflective glass texture in image 2, Change the vegetation style in image 1 into the desert cactus landscape style in image 3</i>				
	<i>Make the fabric of the curtains in image 1 look like the shimmering silk in image 2, Change the artistic style of the portrait in image 1 into the watercolor painting style from image 3</i>				

Which reference supplied the subject? Which supplied the requested change?

Attention as weighted information gathering

Illustrative query: “pear surface”	Weight	Value
Material-reference token	0.8	0.9
Background-reference token	0.2	0.1

Weighted value = $0.8 \times 0.9 + 0.2 \times 0.1 = 0.74$

Softmax weights sum to one

$$0.8 + 0.2 = 1$$

$$0.8 \times 0.9 + 0.2 \times 0.1 = 0.74$$

This scalar example illustrates weighted combination.

Attention weights alone do not establish causality.

LoRA: low-rank adaptation

$$W' = W + B A$$

4,096² versus 2 × 4,096 × 16

A rank-16 update uses 131,072 parameters here.

The full square matrix has 16,777,216 parameters.

Reference conditioning supplies input. LoRA learns a parameter update.

Did it learn the subject or the setting?

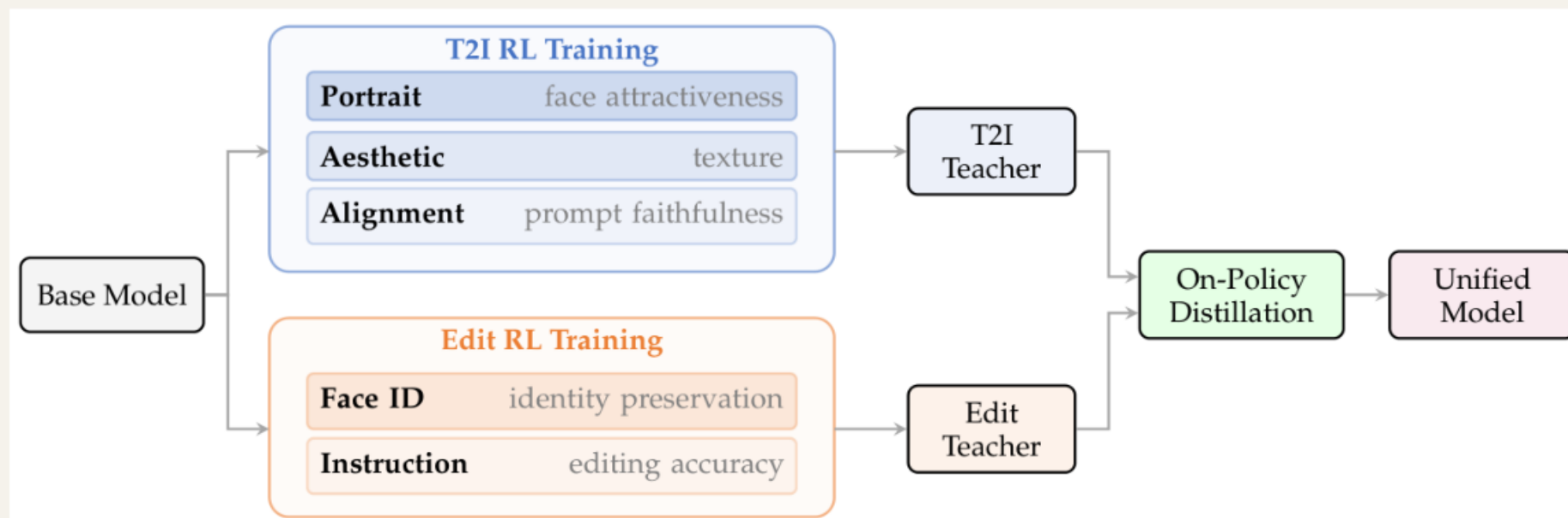
Training success

The adapted model reproduces familiar views and backgrounds.

Useful generalization

The concept survives a new view, setting, or composition.

What does the reward favor?



Qwen-Image-2.0-RL separates task-specific reward training, then distills.

Aesthetic preference and identity preservation can pull in different directions.

One reward cannot stand in for every intention

Request: a studio portrait preserving the reference subject's facial features.



Separate edit fidelity from the preference signal used to rank outputs.

Attractiveness and fidelity are separate tests

Beautiful, but wrong

The image is attractive.
The pear changes identity.
The exhibition's premise disappears.

Faithful, but unconvincing

The right object changes.
The reflection contradicts it.
The material feels pasted on.

Evaluate the requested change and the retained world separately.

A useful diagnosis changes the next action

Observed failure	First hypothesis	Candidate revision
Wrong object changes	Ambiguous reference	Clarify object or region
Identity drifts	Weak appearance grounding	Improve source/reference
Reflection contradicts material	Dependent change omitted	Revise light or composite

Choose the next step from evidence in the failed result.

Discuss: the ceramic edit

- 01 Would you preserve the reflection or let it change? Why?
- 02 When would a mask help more than a longer prompt?
- 03 What would edges or depth control still leave uncertain?

A worked ceramic-edit contract

Preserve the blue stem and pear silhouette.

Allow transmission, highlights, and reflections to adapt.

Inspect the boundary and water before accepting the edit.

The brief, the model, and the review

- 01** **The brief: define the intended change.**
- 02** The model: propose an edited result.
- 03** The review: check the change and what survived.

After Part 1: explain the control

- 01 How does ControlNet differ from a mask or a text prompt?
- 02 Why can stronger guidance make an edit worse?
- 03 What evidence would show that an edit preserved identity?

Break

10 minutes

After the break: what makes a set of images belong to one artwork?



One exhibition, different visual worlds

“ What makes two very different images belong to the same artwork?

PART 2

Art direction

Reference, composition, and iteration

Before Part 2: artistic intention

- 01 What makes an image feel fragile rather than merely attractive?
- 02 Can very different images belong to the same artwork? Why?
- 03 Which artistic decisions would you keep for yourself?

AFTER RAIN

A fictional exhibition about fragile objects in changing environments.

Deliverables: a poster and a five-second moving image.

Keep the amber pear and blue stem recognizable.



Making fragility visible

Artistic intention	Possible visual decision	Question for critique
Vulnerability	Small object, large surroundings	Does scale create exposure?
Stillness	Quiet light, restrained movement	Does time feel suspended?
Instability	Reflection before direct view	Does the reveal change meaning?

Possible interpretations of a fictional brief, not a fixed recipe.

Hokusai: scale, rhythm, and tension



Scale

A small distant mountain.

Rhythm

Repeated curves and crests.

Tension

A wave hangs over the boats.

Reference as a relationship, not a label

Name the scale relationship or visual rhythm you need.

Translate that relationship to your own subject.

Judge the translation against the new work's intention.

Where does your eye arrive?

A small object in a large, quiet space.

A reflection extends its presence.

Where can the title live?



Negative space has a job

In the composition

It changes the apparent scale and isolation of the object.

In the finished poster

It provides room for a title and controls the reading order.

A visual language has rules

Torn edges replace smooth contours.

Flat layers replace optical depth.

Amber and indigo remain the anchors.



When a reflection breaks the collage

Coherent treatment

Edges, depth, and reflections follow the chosen paper language.

Unintended exception

A photographic reflection contradicts the flat material system.

Reference roles

Reference	Borrow	Do not borrow
Subject image	Pear silhouette and blue stem	Its original background
Composition study	Scale and negative space	Unrelated objects
Material study	Edge quality and surface behavior	A conflicting light direction

A short role statement reduces ambiguity before generation.

What did the reference contribute?

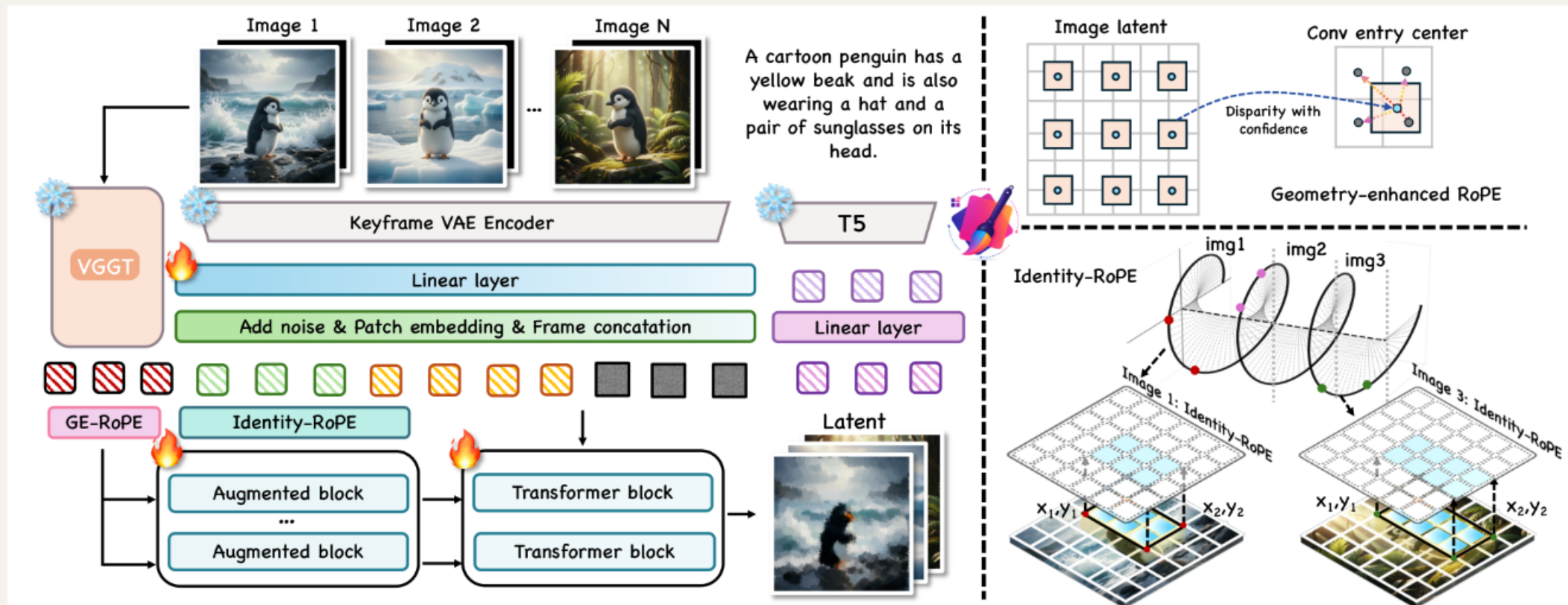
- 01** State what the reference should contribute
- 02** Compare outputs with and without it
- 03** Inspect the intended quality and unwanted transfers

An art-direction prompt

“ Use the pear reference for its silhouette and blue stem. Place it low on the right in a quiet flooded gallery. Leave the upper left pale and open for a title. Use soft daylight and restrained amber reflections. Keep the mood fragile and still.

Subject identity + composition + light + purpose

Group Editing: one decision, many views



A coherent series requires correspondence across images.

Iteration as a controlled comparison

Version	One deliberate change	Inspect	Decision
A	Baseline composition	Eye path	Keep framing
B	Softer light	Material legibility	Compare
C	More open title area	Hierarchy	Compare

Illustrative iteration plan. Keep a record of unsuccessful versions.

A lucky image or a reliable direction?

A single comparison

One attractive output may reflect a favorable random sample.

A stronger comparison

Several outputs per condition reveal stability, drift, and recurring failures.

Typography is part of the image

A clear first read.
Space for the artwork.
Exact, editable type.

AFTER RAIN

Studies in fragile light

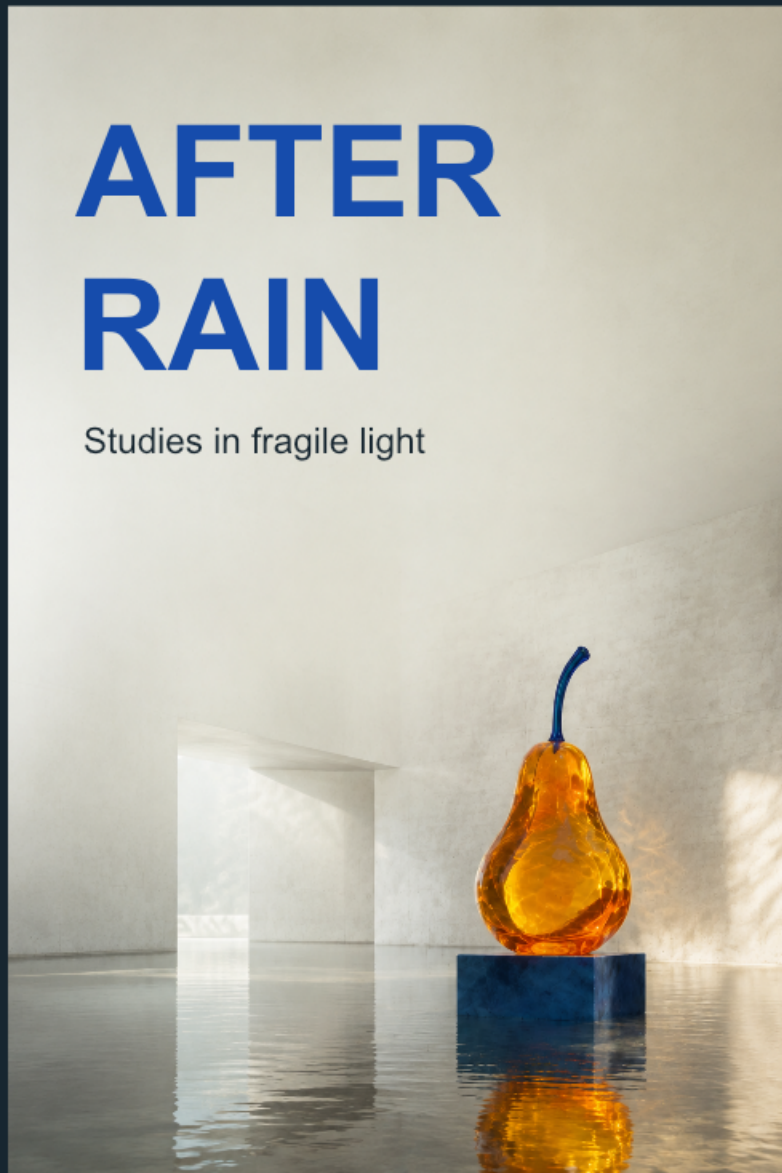


Typography needs a reading order

- 01** What should the viewer read first?
- 02** Where does the eye move next?
- 03** Does the type leave room for the artwork?

AFTER RAIN

Studies in fragile light

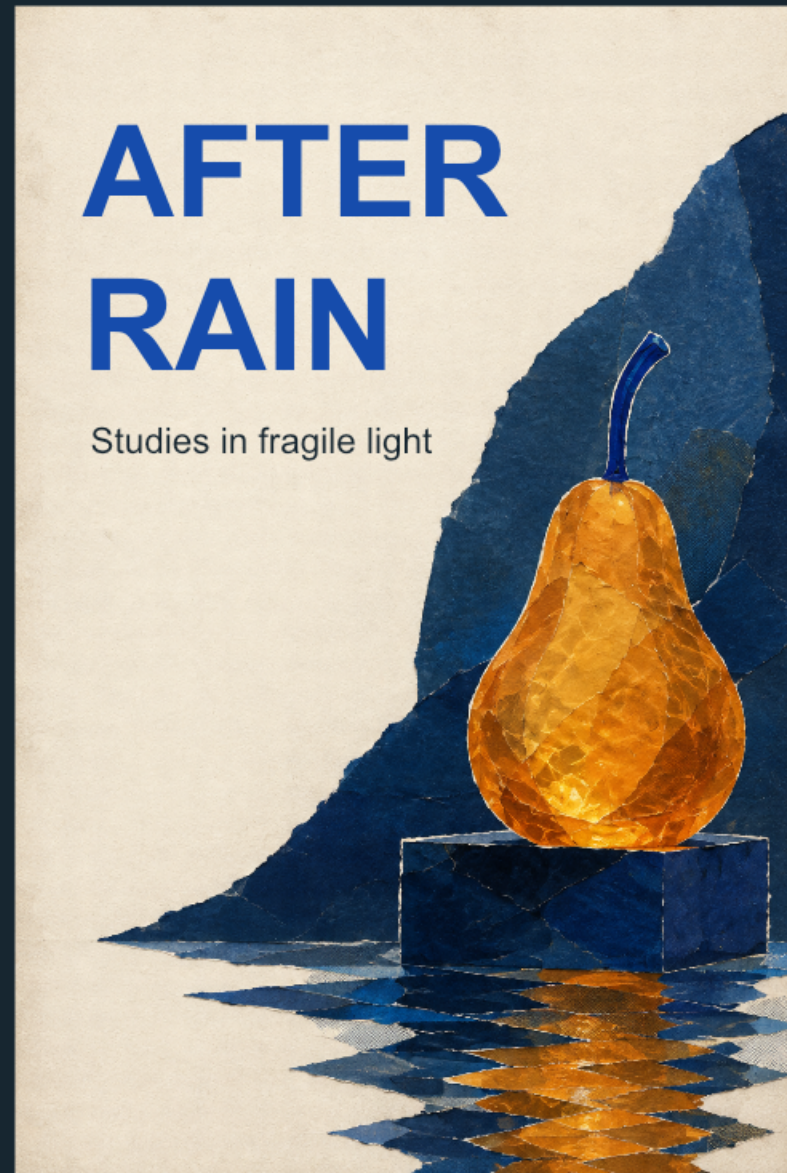


QUIET / SPATIAL

Part 2 / Art direction

AFTER RAIN

Studies in fragile light



TACTILE / GRAPHIC

Why would you exhibit this one?

Name the visual decision that carries fragility.

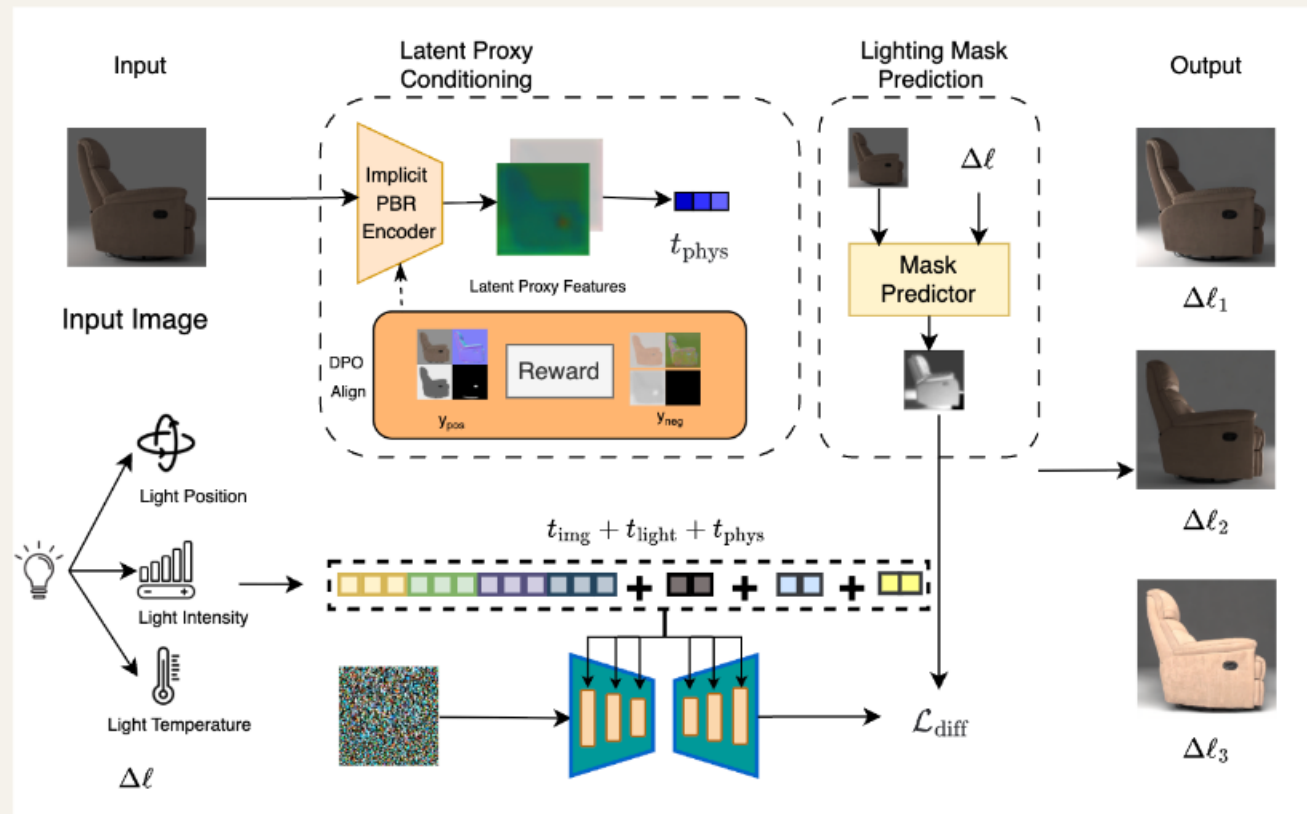
Explain why the rejected direction serves it less well.

Keep an unexpected failure if it becomes an intentional choice.

A critique with observable evidence

Question	Evidence to point at
Intent	The image communicates fragility through scale or material
Coherence	Lighting, edges, and spatial logic support one language
Control	The subject remains recognizable through variation
Hierarchy	The first and second points of attention are clear

LightCtrl: lighting as an artistic choice



Direction, intensity, and color change how a material reads.

Discuss: two artistic directions

- 01 Which poster communicates fragility more clearly? Why?
- 02 What does the rejected direction reveal about your choice?
- 03 Which single revision would most change the audience's reading?

A process record makes critique more precise

Record	Why keep it?
Intention and invariants	Defines what the attempt was testing
Prompt and reference roles	Explains how the attempt was specified
Rejected result and observation	Makes the next revision informed

A small record can make an experiment reproducible and teachable.

A gallery conversation

- 01** Which visible decision carries the idea?
- 02** What changed without a clear artistic reason?
- 03** Which revision would most change the audience's reading?

After Part 2: defend a decision

- 01 How can a technically correct image fail as an artwork?
- 02 Which visual rule should remain across a series?
- 03 What evidence makes a critique useful for the next revision?

TOPIC 3

Generative video editing

Visual transformation across time



Before Topic 3: images in motion

- 01 What new failures become possible when an image moves?
- 02 What should happen when the pear disappears behind a column?
- 03 Can four convincing frames prove that a video works?

A video edit must survive the next frame



Identity persists while pose, visibility, and viewpoint change.

Consistency is not stillness

Expected change

Pose, perspective, illumination, and visibility evolve with the shot.

Unwanted instability

Identity, texture, or style jumps without support from scene motion.

A moving image can reveal an idea

REFLECTION



PARTIAL REVEAL



FULL OBJECT



Let the audience infer the object before seeing it.

Five seconds can have a structure

0.0–1.5 seconds

Reflection suggests the object

1.5–3.0 seconds

A partial view tests the inference

3.0–5.0 seconds

The wide view changes its context

An image editor can read a contact sheet

INPUT CONTACT SHEET



EDITED CONTACT SHEET



Instruction: edit the skier into a robot. / Published Figure 2

Four good frames can hide a bad video

Sampled frames show selected appearances.

They omit transitions and some visibility events.

Full playback is needed to test continuity.

The grid can live in latent space

SOURCE



EDITED



A virtual grid is a representation choice

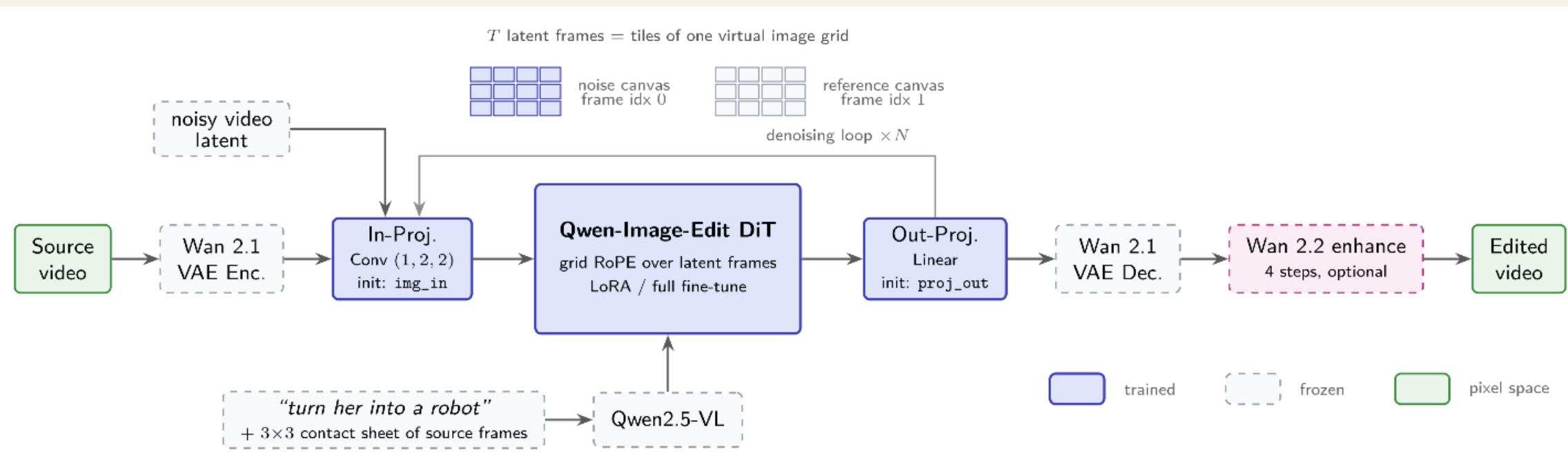
What it provides

A shared arrangement of tokens for the image-editing transformer.

What it does not guarantee

Physical motion, correct occlusion, or persistent identity.

Repurposing an image model for video



Video VAE latents enter an image-editing transformer.

Learned projections connect the representations.

Adaptation connects incompatible representations

- 01** Encode the clip with a video VAE
- 02** Project video latents into the editing model
- 03** Adapt the model and decode the edited result

45 frames become 12: why?

45 pixel frames

$$4 \times 11 + 1$$

12 latent frames

$$11 + 1$$

One virtual 3×4 grid

For this example

Wan 2.1 VAE keeps the first frame special, then compresses time by groups of four.

Temporal compression preserves a special first frame

$$4k + 1 = 45$$

$$k = (45 - 1) / 4 = 11$$

For 45 pixel frames, $k = 11$.

The latent sequence has $k + 1 = 12$ frames.

Guidance in a published video ablation

SOURCE



WITHOUT CFG



WITH CFG



Request: a cyberpunk workshop with holographic documents.

What this comparison actually shows

Evidence shown

In this selected example, guidance makes the edit more complete.

Evidence still needed

Broader clips, failure cases, and a quality–cost comparison.

Local editing across frames

SOURCE



EDITED



An attribute changes while the moving subject should remain recognizable.

Follow the identity marker

- 01** Choose a small recognizable feature
- 02** Track it through pose and partial visibility
- 03** Inspect its reappearance, shape, and attachment

Global stylization across frames

SOURCE



EDITED



A shared visual language must survive the whole shot.

A style can flicker while motion stays correct

Line density changes without a scene explanation.

Textures appear to crawl across stable surfaces.

Review the visual language as well as object motion.

Why raw frame difference is misleading

$$E = \text{mean } |I_{t+1} - \text{warp}(I_t, F_t)|$$

Align corresponding visible content before comparing.

Use a visibility mask to exclude occlusions.

Low error can also reward an overly static video.

The frozen video wins the wrong test

A misleading shortcut

Repeated frames can reduce a naive difference score.

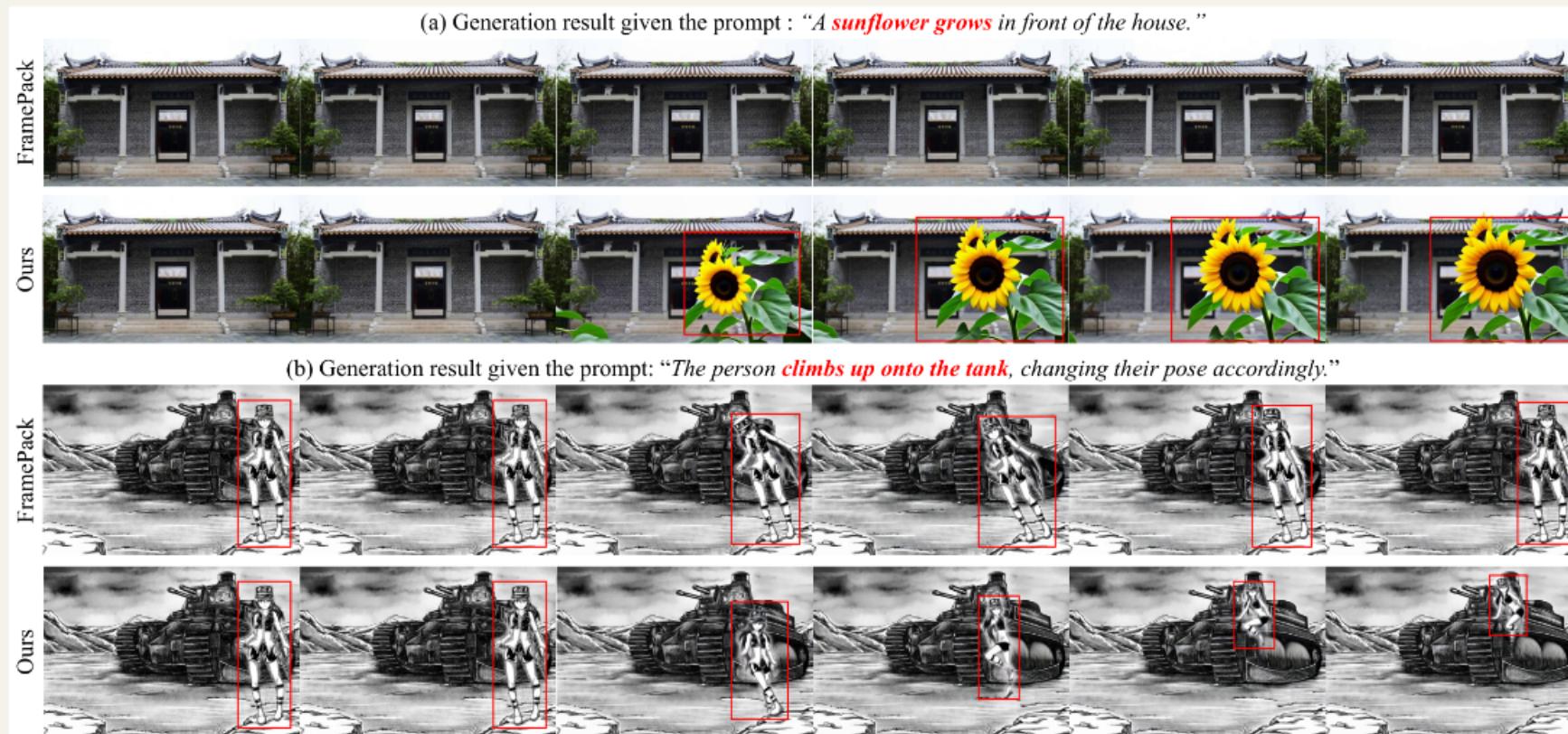
The intended result

The requested motion must persist while appearance remains coherent.

A keyframe can anchor the edit

- 1 Choose a representative source frame
- 2 Edit its appearance and inspect the result
- 3 Propagate the approved look through the clip
- 4 Review motion, occlusion, and the final frame

AlignVid: when the image overrules the text



Selective attention scaling can help the requested event happen.

A five-second video edit brief

“ Change the pear’s body to ivory ceramic throughout the shot. Keep the blue stem, camera movement, and object position. Let reflections match the new material. The pear must remain the same object when it reappears from behind the column.

Transformation + motion invariants + visibility event

The moment the pear returns

- 01** Inspect identity before the column
- 02** Allow partial or complete invisibility
- 03** Compare identity and material on reappearance

A small addition creates new relationships

SOURCE



EDITED



Inspect placement, scale, contact with the scene, and motion.

An addition must belong to camera motion

Single-frame appearance

Texture, shape, and local lighting may look convincing.

Scene relationship

Scale, placement, and trajectory must remain coherent as the camera moves.

The shot review

Viewing mode	What to inspect
Full speed	Does the transformation distract from the intended action?
Frame by frame	Do edges, markers, reflections, or textures jump?
Occlusion / re-entry	Does the same object return?
First / last frame	Does the shot support the next cut?

Fast or complex motion remains a reported challenge for the research system.

A failed shot suggests a different workflow

Failure	Possible next step
Drift after long occlusion	Try a shorter segment or a useful keyframe
Boundary artifacts	Repair with a compositing pass
Transformation overwhelms motion	Reduce the edit and compare again

A workflow change can be more useful than another unchanged prompt.

Discuss: the difficult moment

- 01 Which moment in this shot would you inspect most closely?
- 02 What should remain recognizable after the column?
- 03 What failure would make you reject an attractive clip?

A five-second test plan

The request

Ceramic body; blue stem retained.
Camera and object motion preserved.

The evidence

Full playback plus inspection
before and after the column.

Three decisions without a model

Why does a material edit affect a reflection?

How does a reference differ from LoRA?

Why can four good frames still make a bad video?

Match the problem to the intervention

Choose the intervention

Optical consequences → edit contract
This request → reference conditioning
Learned adaptation → LoRA

Choose the evidence

Exact regions → pixel comparison
Identity → distinctive features
Continuity → full playback

Explain why the intervention fits the observed failure.

The mechanisms behind your decisions

Material changes alter how light interacts with the object.

A reference conditions an output. LoRA changes learned weights.

Errors can occur between sampled frames or during occlusion.

Three levels of checking an edit

- 01** Did the requested change happen?
- 02** Did the necessary invariants survive?
- 03** Does the result serve the artwork's intention?

Who decides?

Models propose images.

Artists choose what matters.

The readings question both roles.



After Topic 3: judge the whole shot

- 01 What can an approved keyframe establish—and what can it not?
- 02 Why might a low frame-difference score reward a bad video?
- 03 Which evidence would convince you that identity survived motion?

AI goals and human direction

An Alien Mind

Jakub Pachocki / OpenAI / 6 September 2026

Which creative decisions would you delegate to AI, and what evidence would make you trust the result?

How to fix your entire life in 1 day

Dan Koe / 23 December 2025

What do you want your creative practice to become, and what small experiment could you begin tomorrow?

Reading 3: image-guided video editing

Introducing Aleph 2.0 and Edit Studio

Runway / 21 May 2026

Before generating the clip

An edited frame establishes the desired appearance.

After generating the clip

Test persistence through motion and occlusion.

Optional technical reading

InstructPix2Pix (2023)

Edit supervision and image/text guidance

Flow Matching (2023)

Learning velocity fields for generation

Qwen-Image-2.0 (May 2026)

Unified generation and editing architecture

Qwen-Video-Edit (August 2026)

Repurposing an image editor for video

Three source types, three kinds of evidence

Source	Read for	Ask
Research perspective	Argument about alignment	Which claims are forecasts?
Reflective essay	Framework for creative intention	Where might it fail?
Vendor announcement	Documented workflow	What does the demo leave untested?

The reading pack links to all three originals.

Separate image and text guidance

$$\hat{\epsilon} = \epsilon_{\emptyset} + s_i (\epsilon_i - \epsilon_{\emptyset}) + s_t (\epsilon_{i,t} - \epsilon_i)$$

$\epsilon_{\emptyset}$ no image or text condition

ϵ_i image only $\epsilon_{i,t}$ image and text

InstructPix2Pix formulation. This predicts noise rather than flow velocity.

What remains when the terms cancel?

$$\varepsilon_{\emptyset} + (\varepsilon_i - \varepsilon_{\emptyset}) + (\varepsilon_{i,t} - \varepsilon_i)$$

$$= \varepsilon_{i,t}$$

At image scale = text scale = 1, intermediate terms cancel.

Training: the target supplies the lesson

```
source, instruction, target = sample_edit_triplet()
```

```
condition = encode(source, instruction)
```

```
target_latent = vae.encode(target)
```

```
noise, t = sample_noise_and_time()
```

```
z_t = (1 - t) * noise + t * target_latent
```

```
prediction = model(z_t, t, condition)
```

```
loss = mean_square(prediction - (target_latent - noise))
```

```
update_trainable_weights(loss)
```

Inference: where did the target go?

```
source, instruction = receive_edit_request()
```

```
condition = encode(source, instruction)
```

```
z = sample_noise()
```

```
for t, step in sampling_schedule:
```

```
    velocity = model(z, t, condition)
```

```
    z = z + step * velocity
```

```
edited_image = vae.decode(z)
```

Choose the approach by the requirement

Main need	Candidate approach	First failure to inspect
Exact untouched pixels	Mask plus compositing	Boundary and lighting seam
New visual language	Reference-guided editing	Subject drift
Reusable learned concept	Adaptation such as LoRA	Overfitting
Transformation in motion	Video editor / keyframe workflow	Occlusion and flicker

An Alien Mind

Pachocki distinguishes achieving a goal from generalizing human values. Trust requires evidence.

- 01** How does achieving an assigned goal differ from respecting human values?
- 02** Which claim in the reading is a forecast, and what evidence would you seek?
- 03** What would you keep under human control in an AI-assisted creative project?

How to fix your entire life in 1 day

Koe proposes examining identity and habits to choose a direction. The promise invites critique.

- 01 Which idea could help an artist choose what to make, and why?
- 02 Which claim seems useful or overstated? What experience supports your view?
- 03 How could easy AI generation support a creative intention or distract from it?

Introducing Aleph 2.0 and Edit Studio

Runway proposes approving an edited image, then applying the change through a video.

- 01 What can an edited keyframe establish, and what must the full clip show?
- 02 Which demonstration or preservation claim would you test more carefully?
- 03 How could this workflow serve AFTER RAIN's intention rather than dictate it?

Your final judgment

- 01** Where should an artist place the boundary between control and surprise?
- 02** How would you judge an edit: technical success, artistic purpose, or both?
- 03** Which decision would you delegate to AI, and what evidence would you require?