Generative editing

Lecture 02 · Full teaching transcript

123 slides · 12,666 spoken words · approximately 90 minutes at 140.7 words per minute. Timestamps are pacing estimates, not audio timecodes. Presentations, breaks, and discussion pauses are additional.

001 / 00:00 — Generative Editing

Imagine you are the art director. Your exhibition opens tomorrow. You love this glass pear, but you want to see it in ivory ceramic. You give the editor a tiny request: change the material, keep everything else. Then you notice the reflection. Should it still look like glass? Suddenly, a three-word edit contains a whole argument about the world.

That is our starting point. We will follow this fictional artwork through image editing, exhibition design, and a moving shot. The exhibition is called AFTER RAIN. The pear is our recurring character. By the end, you should be able to explain a model's mechanism, defend a visual choice, and catch a failure that a beautiful preview can hide. First, look carefully at what you think must survive.

002 / 00:54 — One object, three creative decisions

We are going to give the same object three different jobs. First it is a technical puzzle: can its material change while its identity survives? Then it becomes an artwork: can its placement and lighting make us feel that it is fragile? Finally it becomes a character in time: can it disappear behind something and return as the same object?

Those jobs demand different judgments. A technically clean image may say very little. An expressive poster may contain an intentional impossibility. A wonderful still may belong to a broken video. As its job changes, you may find yourself approving a change that you would have rejected five minutes earlier.

003 / 01:40 — The reflection is part of the edit

For a few seconds, ignore the pear. Look at the water beneath it. Now look back at the body. If the body becomes opaque ceramic, what happens to the light that used to pass through the glass? The answer cannot live entirely inside the object's outline.

This comparison is a prepared classroom illustration. Use it to locate the consequences of the request: transmission, highlights, and the reflection below. We still want the blue stem and recognizable silhouette. But preserving every surrounding pixel could preserve the wrong optics. a successful local edit may require a carefully justified change somewhere else.

004 / 02:22 — Necessary change or accidental drift?

Let us make the opening comparison more disciplined. The disappearance of transmitted amber light is consistent with changing transparent glass into opaque ceramic. A changed blue stem, by contrast, would need a separate justification because the instruction did not request that transformation.

The reflection is more interesting. A material change can require its appearance to change, while its placement should still agree with the object and the water. We cannot classify every difference using a simple rule that change is bad. We need a model of the intended scene. That is why the edit contract includes allowed consequences as well as invariants.

005 / 03:05 — Tonight’s route

Here is the route through that puzzle. We begin inside an image editor: its representations, sampling process, and controls. Then we take the art director's seat and decide what those controls should accomplish. In the final technical section, the image starts moving and the preservation problem becomes harder.

The spoken script is planned for about ninety minutes. The marked discussions, the break, and student presentations need additional class time. You will see the same pear return in different situations. Each return asks us to revise our judgment, rather than learn a completely unrelated example. At the end, three readings let us question who sets the goal and who decides whether it was achieved.

006 / 03:53 — Three decisions you can defend

Here are three decisions I want you to be able to defend by the end. When an edit fails, which control would you change, and why? When two posters both look good, which belongs in this exhibition? When a video looks convincing, which moment would you inspect before accepting it?

You do need a link between mechanism and consequence. A mask tells us where; a reference can show us what; an artistic brief tells us why the result matters. Today we will practice making those links aloud. The aim is to leave with reasons you can use when the next tool has a different interface.

007 / 04:38 — How image editors work

The pear gives us a concrete technical problem: change its material while retaining the information that makes it recognizable. We can describe this as conditional generation under preservation constraints. The inputs specify the change; the contract specifies what should survive. Throughout this section, connect each mechanism to one part of that problem. We begin by making our expectations explicit.

008 / 05:03 — Before Part 1: image editing

Before we open the machinery, decide what success should look like. When glass becomes ceramic, what should stay the same? Which parts of the scene must change with the material? What could a model misunderstand in that request?

Discuss these three questions and point to visible evidence for your choices. If your partner wants to preserve a detail that you want to change, identify the artistic intention behind each answer. Pause the video here. We will return to these expectations when we write the worked ceramic-edit contract.

009 / 05:40 — The edit contract

The gallery has become a forest. Let us write the edit contract. The intended change is the environment. The pear's identity, its pedestal, and the framing should remain recognizable. The ambient light and reflections are allowed to adapt to the forest.

Why include that last category? Because preserving every visible relationship would contradict the new setting. Imagine retaining a bright gallery-window reflection inside a dark forest. The model might preserve the source faithfully and still produce an implausible scene. Before generating, separate the requested change, the invariants, and the consequences that should adapt. Afterwards, inspect those categories separately.

010 / 06:22 — An invariant can be semantic or exact

Suppose a curator says, keep the pear exactly the same. There are at least two meanings hiding in that sentence. One means that a visitor should recognize the same sculpture in a new room. Its highlights may change with the room. The other means that selected pixel values must remain identical.

Those are different requirements. Generative conditioning can encourage recognizable identity. Copying source pixels or compositing can enforce exact preservation in a specified region. Neither requirement automatically makes the whole image coherent: a copied region may now have the wrong lighting. Before choosing a tool, settle the meaning of same.

011 / 07:05 — Editing as conditional generation

Read the expression as a distribution of possible edited results, given the source, instruction, and optional references. The symbol theta represents the model's learned parameters. The vertical bar means 'given.' We are not asking for a unique answer in every case. Two different forest scenes might both satisfy the same brief.

However, a source image is a condition rather than a promise that every unmentioned pixel will be copied. To judge success, we need a more precise contract than 'it looks plausible.'

012 / 07:40 — A condition is not a hard constraint

The probability expression says that the system produces a candidate given conditions. It does not contain a certificate that the candidate passes our checks. The acceptance step is a separate decision that we impose on the result.

Imagine two forest outputs. Both look plausible, but one changes the pear's stem and one retains it. The model can assign probability to both; our contract can reject one. In practice, acceptance may involve human inspection, image comparisons, or explicit production constraints.

013 / 08:14 — A smaller workspace: image latents

The RGB image has 1024 by 1024 pixels and three channels: just over three million scalar values. With spatial reduction by sixteen and sixty-four latent channels, the representation becomes 64 by 64 by 64: 262,144 values.

Notice the trap. Sixteen times smaller along each spatial axis does not mean sixteen times fewer values overall, because the number of channels changes too. Here the scalar count falls by a factor of twelve. That is specific to this configuration. The model works in a learned representation, and its decoder must recover visible detail. Our next question is what that compression already loses before we request any edit.

014 / 08:59 — Is the detail lost before the edit?

Suppose the small exhibition caption is already blurry after an edit. You might spend an hour rewriting the prompt to say, preserve every letter. There is a quicker diagnostic: encode the original image and decode it again without requesting a change.

If the letters are already damaged in that round trip, part of the problem lies in representation and reconstruction. A more emphatic instruction cannot restore a detail the editing path never retained faithfully. Look at fine text, thin edges, and small textures first. If reconstruction is sound but the edit damages them, investigate the editing stage. It also explains why exact exhibition typography deserves its own editable layer.

015 / 09:45 — Why token count matters

Now consider the number of tokens a transformer processes. In our toy example, one token per position on a 64 by 64 grid gives 4,096 tokens. Grouping each two-by-two region gives 1,024 tokens, a reduction by four.

Dense self-attention compares token pairs. Squaring the two counts gives 16,777,216 and 1,048,576 pair scores. That is a reduction by sixteen. This is why representation choices matter so much for cost. It is not a claim that the whole system runs sixteen times faster. Text tokens, reference tokens, other layers, and implementation details also matter.

016 / 10:24 — Double the resolution: count the cost

Before reading the second number, make a prediction. We double the image width and double its height, keeping the patch scheme fixed. How many image tokens do we get? Four times as many. In dense self-attention, each token can compare with every other token. Four times four gives sixteen times as many pair scores.

This calculation describes the image-token attention component, not a promise that the entire system becomes sixteen times slower. Other operations and implementations matter. The useful habit is to ask what grew: pixels, tokens, comparisons, or measured runtime. They are related quantities, but they are not interchangeable.

017 / 11:07 — Inside a current image editor

Let us walk through this architecture from the input side. The source image contributes semantic information through a vision-language component and visual information through a VAE. The target stream begins with an evolving noisy representation. The transformer processes that stream together with the conditions.

During training, we possess an example of the desired edited target, so we can measure what the model should learn. During inference, we have the source and request but not the desired target. The model must generate it. That difference is essential. The architecture does not quietly receive the answer when you ask it to edit your photograph.

018 / 11:50 — The missing target at inference

There is a piece of information on the left that we do not possess on the right: the desired edited image. During supervised training, a source, an instruction, and a target show the model what a successful change looks like. During use, we provide the request because that target does not yet exist.

This distinction explains a surprisingly common confusion. Showing a reference to an editor is not automatically teaching new weights. It may simply be conditioning this particular generation. Adaptation, which we will reach with LoRA, changes trainable parameters. It is available when we construct the learning signal, and absent when we ask the trained model to make our ceramic pear.

019 / 12:38 — Two representations of the source

Think of these two representations using our pear. Semantic features help with questions such as: which object is the pear, what does ceramic mean, and what does 'it' refer to in the instruction? Visual latents provide appearance information, including shapes and textures that help keep the particular source recognizable.

This is a conceptual distinction, not a claim that the branches have perfectly isolated responsibilities. Learned representations can overlap in what they encode. If an edit fails, ask whether the model misunderstood the request or lost important visual information.

020 / 13:16 — OpenSubject: identity across new scenes

How do you teach a system the difference between a pear and this particular pear? That question connects to OpenSubject, research I coauthored with Yexin Liu and our collaborators. A video offers something valuable: repeated observations of a subject as the view changes. The shared identity is a learning opportunity.

Follow the main path in the figure. We curate clips, verify subjects across frames, and select diverse pairs. Inpainting or outpainting helps synthesize reference inputs, followed by verification. The corpus contains 2.5 million samples; the contribution is training data and a benchmark. For our exhibition, imagine the same sculpture photographed in several rooms. We want freedom to change the setting without losing its distinctive features. That is a classroom application of the identity problem, rather than a claim that our fictional pear was tested in the paper.

021 / 14:14 — Flow matching: the training target

The editor needs a way to learn how to move from noise toward an image. In this simple flow-matching setup, we construct an intermediate latent by mixing noise, epsilon, with a known target latent, z one. The mixing variable t tells us where we are along that training path.

For this straight interpolation, the target velocity is the target latent minus the noise. The model sees an intermediate state and learns to predict that direction, given the source and instruction. There is also an important clock distinction: t here is the generative process's time. Later, our video will have another time axis—the seconds that pass in the scene.

022 / 15:00 — The interpolation endpoints

Before trusting an equation, check its easiest cases. At t equals zero, the coefficient on the target becomes zero, so we recover the noise. At t equals one, the coefficient on noise vanishes, so we recover the target latent.

Along this simple straight path, the difference between target and noise gives the direction we train the model to predict. You do not need to imagine a recognizable half-finished image at every intermediate point. The calculation takes place in a learned representation, and this path is a teaching construction rather than the only possible design.

023 / 15:40 — Predict the next sampling step

Here is the smallest possible sampling calculation. One coordinate currently equals 0.20. Its predicted velocity is 0.60, and the next step has size 0.10. Before looking at the result, say what operation should happen. We add a small amount of motion to the current state: 0.20 plus 0.10 times 0.60, giving 0.26.

In a real latent, many coordinates update together. These numbers are invented to show an Euler step, and practical samplers can use more elaborate solvers. But the sequence is now less mysterious: predict a direction, take a step, repeat, then decode.

024 / 16:20 — A better solver cannot clarify the brief

Imagine following a beautifully accurate set of directions to the wrong gallery. Taking smaller steps will not repair the destination. The same distinction helps with generative sampling. A better numerical solver can follow a learned field more accurately. That alone does not settle an ambiguous editing request.

If the pear's boundary is unstable across sampling settings, numerical or model behavior may matter. If we never specified whether the glass reflection should become ceramic, we also have a brief problem. Ask whether the system struggled to realize a clear instruction, or whether we have not agreed on what success means. Those situations call for different next actions.

025 / 17:05 — Where edit supervision comes from

Where does the ability to follow an edit instruction come from? A training example can contain a source image, an instruction, and the corresponding edited target. InstructPix2Pix is an early example that used synthetic editing data to teach this relationship.

Suppose a training pair says 'make the object ceramic,' but its target also moves the camera and replaces the background. The learning signal no longer cleanly identifies the requested transformation. The model may learn unwanted associations. This gives us an important data-quality question: did the target accomplish the requested change while preserving what should remain? Attractive targets alone are not enough to teach dependable editing.

026 / 17:50 — What a mismatched training pair teaches

A training pair is a lesson, and lessons can accidentally teach the wrong thing. Imagine a source glass pear and a target ceramic pear that also has a different background. The written request says only to change the material. Which visible changes should the model associate with that request?

If this mismatch recurs in the training data, the model can learn an unwanted association between a material change and a scene change. Compare the instruction with the actual difference between source and target. Ask what supervision is rewarding, including changes nobody meant to label. The preservation contract begins in the examples used to teach the editor.

027 / 18:35 — Classifier-free guidance (CFG)

Sometimes a conditioned prediction moves in the right direction but too weakly. Classifier-free guidance uses the difference between a conditioned prediction and a baseline prediction to steer the result. Read the formula as baseline, plus a scaled change in direction.

When s equals one, the baseline terms cancel and we recover the conditioned prediction. Above one, we extrapolate beyond it. That can strengthen the requested change, but it can strengthen errors as well. In an editing system, the baseline may still retain image information; baseline does not always mean no information at all. Think of guidance as a specific operation on predictions, rather than a universal quality dial.

028 / 19:21 — Guidance goes beyond the prediction

The baseline predicts 0.2 and the conditioned version predicts 0.5. With guidance scale two, do we get a value somewhere between them? No. We take 0.2 plus twice the difference of 0.3, which gives 0.8. We have moved beyond the conditioned prediction.

If the useful direction includes a slight mistake, we can amplify both. For the pear, the new material might become clearer while the stem or boundary becomes less faithful. Compare the requested change and preservation separately. One slider can move those judgments in opposite directions; a single overall impression can hide the tradeoff.

029 / 20:02 — Different controls carry different information

Different controls communicate different kinds of information. Text describes a requested change. A mask specifies a region. A spatial condition can describe pose, depth, or edges. An image reference can supply appearance information that would be difficult to express precisely in words.

These are not interchangeable knobs, and every product does not expose all of them. ControlNet and IP-Adapter are examples of distinct architectural approaches. If exact untouched pixels are essential, a generation mask alone may be insufficient; explicit copying or compositing can enforce that requirement. Choose a control by asking what information is missing, then check whether the resulting image actually respected it. A mask indicates a permitted region; it does not by itself solve seams or the reflection of a changed material. The next slide names an architecture that learns to use a spatial condition.

030 / 21:00 — ControlNet: add spatial conditioning

Suppose your sentence is understood, but the pear keeps changing shape. You can describe its outline with more adjectives, or supply spatial evidence. ControlNet gives an edge map, depth map, or pose a learned route into a compatible diffusion model.

Trace the two paths in this original architecture. The pretrained path stays frozen. A trainable copy processes the added condition, and zero-initialized one-by-one convolutions connect its features to the main path. Initially those connections contribute zero; training learns their contribution. The copied branch itself starts from pretrained weights, not all zeros.

Now return to our request. Edges can help specify the outline. They cannot, by themselves, tell us whether the reflected ceramic looks convincing or the blue stem retains its identity.

031 / 21:52 — Multiple references need distinct identities

With several references, the system must know which reference contributes which information. Imagine asking for the sculpture from image one and the atmosphere from image two. If the references become confused, you may get the wrong object with the right lighting.

The illustrated research approach uses separators and image-index information to distinguish inputs. In an art brief, name the role of each image rather than presenting a pile of vaguely related inspiration. Then inspect for leakage: did a composition reference accidentally replace the subject? This figure describes one proposed mechanism, not a universal design used by every editor.

032 / 22:34 — A reference-role failure is visible in the output

Look at these published multi-image examples with reference roles in mind. Before judging the output, identify what each input was supposed to contribute. Then trace the subject and the requested transformation into the result.

A reference-role failure can look superficially attractive. The system may borrow the wrong object's appearance or import a background that was never intended. We are examining qualitative examples from a particular paper, not conducting a broad comparison between products. The figure helps us practice an inspection method: follow the intended contribution of each input and look for unintended transfers between them.

033 / 23:15 — Attention as weighted information gathering

Attention lets a token gather information from other tokens with different weights. Our simplified scalar example gives weight 0.8 to a value of 0.9, and weight 0.2 to a value of 0.1. The weighted result is 0.74.

The real mechanism operates on learned vectors, so this is arithmetic intuition rather than a literal description of artistic decision-making. The important structure is selective combination: every available source need not contribute equally. In a multi-reference edit, the system also needs to distinguish where information came from. Otherwise, gathering information successfully can still produce the wrong mixture of subject identity and visual style.

034 / 23:58 — Softmax weights sum to one

The weights in this simplified attention example sum to one. That makes the result a weighted combination of the values. A larger weight makes the associated value contribute more to this particular calculation.

Now be cautious about jumping from the calculation to an explanation of the final image. Real networks contain many layers, heads, and transformations. A high weight at one point may not tell us why a visible feature ultimately appeared. For practical reference editing, the useful question remains whether the intended information survives in the output, regardless of how compelling an attention visualization looks.

035 / 24:39 — LoRA: low-rank adaptation

What if a useful concept must recur across many requests? A reference can condition a generation, while LoRA adapts selected learned weights using a low-rank update. The update is a product of two smaller matrices. Instead of learning every entry of a large update matrix independently, we express that update as a product of two smaller matrices.

For a 4096 by 4096 weight matrix and rank sixteen, the full matrix has 16,777,216 entries. The two factors together have 131,072, a factor of 128 fewer for this update. That is not a claim that the entire model shrinks by 128. The base weights still exist. Low rank makes adaptation economical, but does not certify that the learned subject survives a new pose. That question requires examples the adaptation did not already see.

036 / 25:35 — Did it learn the subject or the setting?

An adapted model reproduces your favorite training portrait perfectly. Is that enough to use the character in a new film? Consider what else it may have learned: the familiar camera angle, the background, even the lighting that always accompanied the subject.

A held-out check deliberately changes those circumstances. Ask for a new view or a different setting and inspect the distinctive features. We want a reusable concept, rather than a narrow ability to reproduce familiar combinations. For the pear, keep the blue stem recognizable while moving the exhibition outdoors. If identity collapses there, more praise for the training examples will not solve the problem.

037 / 26:19 — What does the reward favor?

A reward tells a learning process what kinds of outputs to favor. The figure shows a recent approach that separates task-specific reward training and then distills what is learned. Look at the categories: editing quality involves more than a single judgment of attractiveness.

Now imagine an artwork whose purpose is to feel awkward or disturbing. A generic preference for polished images could work against that purpose. Human preference signals are useful, but they do not define artistic merit for every project. This is the bridge to the next section: technical optimization can help produce candidates, while an artist still needs to decide which properties serve the work.

038 / 27:05 — One reward cannot stand in for every intention

These edit examples invite a question about evaluation. Which properties would you reward separately? You might ask whether the instruction was carried out, whether identity survived, and whether the image is visually convincing.

Those judgments can disagree. A highly polished output might erase an awkward feature that is central to the artwork. An unusual composition might serve the brief while attracting a lower generic preference score. We should understand what a reward encourages before treating a high score as artistic approval. The published examples illustrate the research setting; our classroom task is to articulate the intention against which we would judge a particular result.

039 / 27:49 — Attractiveness and fidelity are separate tests

Which failure would you notice first? On one side, the output is beautiful, but it has quietly replaced our particular sculpture with a generic decorative pear. On the other, the right sculpture is present, but its reflection still behaves like the old material.

The first may win a quick aesthetic vote. The second may pass a checklist that only asks whether the requested object changed. Neither satisfies the full brief. Has the editor lost identity, broken the scene's optical logic, or missed the work's intention? Naming the mismatch gives us a next action. Saying only that an output looks a little wrong leaves the diagnosis unfinished.

040 / 28:34 — A useful diagnosis changes the next action

A diagnosis becomes useful when it changes the next action. If the wrong object changes, first suspect an ambiguous reference or region specification. If the correct object becomes another instance, appearance grounding may be the issue. If the material looks disconnected from its reflection, the dependent change may be missing.

These are candidate explanations, not automatic conclusions. Choose a small revision that tests one of them. If it does not help, reconsider the hypothesis. This is more informative than changing the prompt, seed, reference, and model all at once, because then a successful result would leave us unsure which intervention mattered.

041 / 29:17 — Discuss: the ceramic edit

We now have several ways to communicate an edit. Would you preserve the reflection or let it change, and why? When would a mask help more than a longer prompt? What would edges or depth control still leave uncertain?

Discuss these three questions using the pear already on screen. For each control you favor, name the failure it addresses and something it cannot settle. Pause the video here. The next slide offers one possible contract; compare it with your reasoning rather than treating it as the only artistic answer.

042 / 29:55 — A worked ceramic-edit contract

Here is one defensible answer to the ceramic puzzle. Preserve the blue stem and recognizable silhouette. Allow transmission, highlights, and the reflection to change with the new material. Then inspect the boundary and water, because that is where the request reaches beyond the object.

Notice the wording: allow a justified consequence, rather than permit arbitrary drift. We have not given the model permission to redesign the gallery. We have named the changes necessary to make this material transformation coherent. A mask could help localize work, while compositing might protect a region that truly must remain exact. The contract tells us how to use those tools.

043 / 30:40 — The brief, the model, and the review

We can now explain our opening puzzle at three levels. The brief decides what should change and what should survive. The model uses representations, conditioning, and sampling to propose a result. Our review checks whether that proposal actually satisfies the brief.

Keep those levels separate when you diagnose a failure. Stronger guidance will not choose the exhibition's purpose. A more eloquent artistic statement will not enforce identical pixels. A beautiful preview will not prove preservation. The useful skill is connecting the right intervention to the observed problem. If your partner can tell when to use it and what it leaves uncertain, you have understood more than its name.

044 / 31:26 — After Part 1: explain the control

We can now explain our controls rather than just name them. How does ControlNet differ from a mask or a text prompt? Why can stronger guidance make an edit worse? What evidence would show that an edit preserved identity?

Discuss these three questions. Choose a concrete requirement—a silhouette, an untouched region, or a distinctive stem—and connect your explanation to it. Then consider whether the control guarantees that requirement or only helps express it. Pause the video here before the break.

045 / 32:00 — Break

We will take ten minutes and resume when the class is ready. When you come back, think about a series of images you would recognize as belonging to one artwork. What makes them belong together: the subject, the palette, the treatment of space, or something else?

Leave that question open for now. We will use it to move from individual editing operations to a coherent visual language.

046 / 32:29 — One exhibition, different visual worlds

Look again at the artwork we have been using as a technical test. After the break, it has a different job. We are no longer asking only whether the edit obeys a request. We are asking what a visitor might feel, and which visual decisions create that feeling.

Two images can have different surfaces and still belong to one exhibition. Two images can share a palette and feel unrelated. The difference is worth arguing about. In the next section, I will show prepared alternatives for AFTER RAIN. Choose a direction in your mind, and be ready to explain a visible reason. Your neighbor may choose the other one.

047 / 33:15 — Art direction

Now take the art director’s seat. The model can produce many plausible outputs; we decide which differences matter for AFTER RAIN. As we compare the prepared images, name the visible relationship that carries the idea. That could be scale, light, or the treatment of space. Your explanation will be more useful than simply calling an image impressive.

048 / 33:39 — Before Part 2: artistic intention

Before seeing the poster alternatives, consider what you want a visitor to experience. What makes an image feel fragile rather than merely attractive? Can very different images belong to the same artwork, and why? Which artistic decisions would you keep for yourself?

Discuss these three questions. You might disagree about the effect of empty space or the meaning of a material. Locate the visual evidence behind that disagreement. Pause the video here. Keep your starting position in mind when we compare the prepared directions.

049 / 34:15 — The exhibition brief

Here is our commission: AFTER RAIN, a fictional exhibition about fragile objects in changing environments. Imagine a visitor seeing its poster across a corridor before encountering a five-second moving image inside. What should that visitor expect to feel?

The prepared examples keep the amber pear and blue stem recognizable. We will examine photographic and collage directions, then consider how a moving reveal changes the experience. Nobody needs to generate an image during this lecture. Fragile is a useful beginning, but it is not yet an art direction. We need to translate that word into something visible enough to compare and precise enough to revise.

050 / 34:59 — Making fragility visible

Try replacing fragile with expensive in the brief. You might still choose glass, but would you use the same composition? A large centered object and assertive lighting might suggest a luxury product. A small object surrounded by quiet space might instead seem exposed.

Scale can make the pear feel vulnerable. Restrained light can make us look closely. A reflection can suggest a world less stable than the object itself. These are artistic hypotheses to test against an audience's reading, not a formula that makes every image fragile. Point to the feature doing the work. If nobody can locate it in the image, the intention may still be living only in the prompt.

051 / 35:47 — Hokusai: scale, rhythm, and tension

In Hokusai's Great Wave, look first for Mount Fuji. It is small and distant. Now follow the curves of the wave and the boats beneath it. Scale and rhythm create a tension we can discuss without turning the artwork into a style label.

For AFTER RAIN, we might borrow a relationship: a small, vulnerable form facing a much larger environment. We do not need to reproduce the wave or ask for a generic imitation of the artist. The reference becomes useful when we can say which visual decision we are studying. Then we must test its translation. Our quiet flooded gallery has a different subject and emotional register.

052 / 36:33 — Reference as a relationship, not a label

Instead of using a reference as a label, describe a relationship you can observe. You might want a small stable form beneath a large dynamic curve, or a rhythm that moves the eye through the composition.

Then translate that relationship into the new subject and brief. The result should be judged as its own artwork, not as a contest to resemble the reference. This approach makes references more useful to collaborators because they can understand what you are borrowing. It also makes iteration more focused: if the intended tension is missing, you can revise scale or rhythm rather than vaguely asking for more influence from the source.

053 / 37:19 — Where does your eye arrive?

Before admiring the surface detail, decide where your eye arrives. Is it the pear, the reflection, or the space waiting above them? A poster has to organize that first encounter. If every region is equally busy, the audience has to invent a hierarchy we have not provided.

Here, the open area can make the object feel small and give the title somewhere to live. We can ruin the sense of quiet by filling every available gap with decorative detail, even if each addition is attractive on its own. When revising this image, I would first ask whether the composition expresses the intended scale and silence. More texture comes later, if the work needs it.

054 / 38:08 — Negative space has a job

Negative space does two jobs in this poster. It changes how large or isolated the object feels, and it creates a place for typography. These are connected design decisions rather than separate finishing steps.

If you fill the pale area with dramatic detail, the image may become more visually active but leave no calm route for reading the title. If you reserve too much space, the subject might lose the presence you wanted. Evaluate the space in relation to the finished use. A generative background is part of a larger composition when it must carry text, branding, or other exact elements.

055 / 38:51 — A visual language has rules

The collage direction changes the rules of the world. Torn edges replace smooth contours. Flat layers replace optical depth. Amber and indigo keep a connection to our recurring subject. Look at how those decisions affect the pear's apparent weight and vulnerability.

If the body looks like paper but the water behaves like a photograph, do we accept that collision? We might, if it is deliberate and supports the work. We might reject it as an unresolved mixture. The answer depends on the intended visual language. The next comparison asks you to locate precisely where that language holds together or breaks.

056 / 39:34 — When a reflection breaks the collage

A photographic reflection inside a paper collage can be a mistake, or the most interesting decision in the image. We need to know whether the mismatch is an intentional disruption and whether it produces the intended effect.

Suppose the exhibition explores unstable memories. An impossibly photographic reflection could make sense. Suppose the brief calls for a coherent world assembled from torn paper. The same reflection might weaken it. This does not mean every accident deserves an explanation after the fact. Make the intention specific, examine what the viewer can actually see, and compare alternatives. A strong critique can distinguish a productive contradiction from an excuse for an unresolved result.

057 / 40:20 — Reference roles

A reference becomes easier to use when we assign it a role. One image might define the subject, another the composition, and another the material treatment. Name the property you want from each.

The model may not isolate those roles perfectly. That is why you should look for unwanted transfers, such as importing a reference's background when you only wanted its palette. Try removing one reference and observing what disappears. Start with the smallest useful set. If five references contradict one another, adding a sixth may make the problem harder to diagnose rather than solve it.

058 / 41:01 — What did the reference contribute?

We have added a reference because we think it helps. How would we discover whether it actually does? Remove it and compare. First state its intended role: perhaps the pear's silhouette, perhaps the composition's sense of scale. Then look for that quality in outputs with and without the reference.

Also inspect what came along uninvited. A useful subject reference can carry a background or lighting scheme we did not want. This is an ablation: change one component to understand its contribution. Random generation makes one lucky comparison weak evidence, so several samples can help. If we cannot explain what a reference contributes, we may be making the request more complicated without making the direction clearer.

059 / 41:50 — An art-direction prompt

Listen to how this prompt distributes responsibility. The pear reference supplies silhouette and the blue stem. Placement low on the right organizes the composition. A pale upper-left area reserves space for the title. Soft daylight and restrained reflections support the mood.

Every phrase points toward something we could inspect in an output. Compare that with asking for a stunning masterpiece with beautiful lighting. This prompt is still a proposal, not a binding contract enforced by the model. If an output fails, we can now name the failed relationship: the title area is crowded, the light is too assertive, or the reference's background has leaked into the scene.

060 / 42:36 — Group Editing: one decision, many views

A single poster can look coherent by itself. A series exposes a harder problem: does the same material decision survive across viewpoints? Our Group Editing collaboration studies related images that should be edited consistently. Treating each image independently can produce slightly different costumes or materials.

The method arranges related images as pseudo-video frames to use a video model's consistency prior. VGGT provides geometric correspondences. Geometry-enhanced rotary positional embeddings connect geometry features with image latents, while Identity-RoPE supports identity preservation. Follow the penguin examples through the figure before reading every label.

Reliable correspondence is part of the technical problem: when views cannot be matched well, repeating the same instruction alone does not establish a coherent series.

061 / 43:25 — Iteration as a controlled comparison

Iteration becomes informative when we know what changed between attempts. Suppose we compare two lighting treatments. Keep the subject and composition instructions stable, and record the references and model version. If a seed control exists, holding it fixed can help the first comparison.

A shared seed still does not guarantee identical composition after a prompt change. If results vary widely, examine several outputs for each condition. Otherwise, one lucky sample may decide the whole direction. The table is a proposed experiment, not measured results. Its purpose is to make each generation answer a question about the artwork.

062 / 44:06 — A lucky image or a reliable direction?

Imagine you compare two prompts once, and the second gives a wonderful poster. Did its wording cause the improvement, or did it receive a favorable random sample? From one output, it can be difficult to tell. Several outputs per condition reveal whether the direction is stable or merely fortunate.

For art direction, we can ask two different questions. Would I exhibit this particular image? Could I reliably develop a series using this direction? One exceptional output may answer the first while leaving the second unresolved. Repeating generations is helpful when it reveals a pattern relevant to the brief; it becomes a distraction when we keep browsing alternatives to avoid deciding what we value.

063 / 44:54 — Typography is part of the image

Now the image becomes a poster. The title is an editable typographic layer, which lets us choose its wording, line breaks, and placement exactly. That matters when a work has to carry an exhibition name rather than merely resemble a poster in a generated preview.

Look at how the title occupies the area we deliberately left open. The image and type were designed to cooperate. We can adjust hierarchy without asking the model to regenerate the sculpture and risk changing it. This is a useful division of labor: generation develops the visual material, and direct layout controls exact communication. The audience sees one composition. It does not need to know which parts came from which tool.

064 / 45:43 — Typography needs a reading order

Typography establishes a sequence of attention. Ask what the viewer should read first, what comes next, and where their eye returns to the artwork. In our poster, reserved space allows the title to be clear without covering the sculpture.

This is also a production decision. Exact text can remain editable, while the image carries the material and atmosphere. If the title feels too dominant, change size, placement, or contrast and inspect the whole composition again. Do not judge the text in isolation from the image. The poster is the relationship between them, including the empty space that lets each element do its job.

065 / 46:27 — Two directions for the same brief

Take a few seconds with both directions before I describe them. Which would you put outside AFTER RAIN? Choose privately first, so your answer is not just a response to mine. Then identify the visible feature that made you choose.

The photographic direction can invite attention to material and atmosphere. The collage direction can make fragility feel constructed through paper edges and layers. Both can serve the brief, but they make different promises to a visitor. A useful defense goes beyond realism versus abstraction. Tell us what the audience is likely to notice or feel, and how the composition produces that reading. We will use the disagreement to decide what to revise, rather than vote for a universally better image.

066 / 47:18 — Why would you exhibit this one?

Selection is an artistic decision. Once generation gives us many plausible alternatives, choosing one determines what the work becomes. The reason should connect to the brief: perhaps the small object feels more exposed, or the torn edge makes fragility physically legible.

The rejected direction is useful evidence too. Explain what it does well and why you are choosing something else for this exhibition. An unexpected model output can also change the direction, if you choose to develop it deliberately. The key question is whether you can now articulate the intention and carry it through subsequent decisions. Surprise can begin a work; it does not finish the artist's judgment.

067 / 48:04 — A critique with observable evidence

A useful critique names a visible feature, explains its effect, and proposes a revision. For example: the crisp reflection makes the collage feel photographic, so I would simplify its shape to match the flatter layers.

Compare that with 'I do not like the reflection.' The first comment gives the artist a relationship to inspect and a possible next move. You can disagree about the intended effect, but make the disagreement specific. This is a classroom critique framework rather than an official grading rubric. Use it to connect evidence in the image to the idea the work is trying to communicate.

068 / 48:47 — LightCtrl: lighting as an artistic choice

A curator asks, can the same sculpture feel more vulnerable without changing its shape? Lighting is one way to answer. Our LightCtrl collaboration studies controllable relighting from a single image, including light direction, intensity, and color temperature.

Follow the chair through the figure. The latent proxy encoder extracts compact physical cues. A lighting-aware mask guides the denoiser toward regions affected by the change, and preference optimization in the proxy branch supports physical consistency. The method connects an interpretable lighting request with image generation.

For our pear, look for consequences rather than the word dramatic: where do highlights move, what happens to shadow, and how does the material read? Inferring geometry and material from one photograph is ambiguous, so the result still needs inspection. The exhibition example is our application of the idea, not an additional result reported by the paper.

069 / 49:47 — Discuss: two artistic directions

Choose between the two prepared directions for AFTER RAIN. Which poster communicates fragility more clearly, and why? What does the rejected direction reveal about your choice? Which single revision would most change the audience’s reading?

Discuss these three questions with a specific visual feature in view. You can prefer quiet space or the instability of torn paper, but explain how that choice serves the exhibition. Pause the video here. Listen for a persuasive reason to choose the direction you initially rejected.

070 / 50:22 — A process record makes critique more precise

Consider what a process record lets us discuss. An artist can record the intention and invariants, the prompt and reference roles, and one observation about a rejected result. That record explains what an attempt was testing.

Without that context, a folder of attractive images can be difficult to interpret. We may not know which input changed or why a version was rejected. A few precise notes make a comparison more informative. In a critique, this lets us ask about the artist’s decisions and the evidence behind them rather than guessing only from the final image. It also helps distinguish an intentional departure from an accidental change.

071 / 51:07 — A gallery conversation

Imagine we are standing between the two finished posters in a gallery review. I would begin with a visible decision that carries the idea, then locate an unintended change, then propose one revision worth discussing. The order matters: the critique begins by understanding the work before prescribing a repair.

We can hear two contrasting readings of the same poster. One person may find the empty space quiet; another may find it emotionally distant. Ask which details support each reading. There is no image-making task here. We are practicing how to make feedback useful to the next decision. A good proposal names what would change and what effect we expect, so that a later version could confirm or challenge the reasoning. Discuss these three questions, then pause the video for the gallery conversation.

072 / 52:03 — After Part 2: defend a decision

Before the artwork begins to move, defend a decision about the images. How can a technically correct image fail as an artwork? Which visual rule should remain across a series? What evidence makes a critique useful for the next revision?

Discuss these three questions through one of the prepared posters. Make words such as coherent or expressive concrete: locate the feature, describe its effect, and explain what changing it would do. Pause the video here. We will carry those artistic rules into the video section.

073 / 52:39 — Generative video editing

Our image now has to survive time. The subject can move, disappear, and return. We will examine selected frames from published research, follow the technical representations, and use a prepared AFTER RAIN scenario to decide what a convincing edit requires. Keep the preservation contract, but add an event that could break it.

074 / 53:01 — Before Topic 3: images in motion

A moving image creates new ways to break our contract. What new failures become possible when an image moves? What should happen when the pear disappears behind a column? Can four convincing frames prove that a video works?

Discuss these three questions. Separate a correct change in visibility from an unwanted change in identity. Name a moment you would need to see between the selected frames before accepting the shot. Pause the video here. Your prediction will give us a concrete test for the methods that follow.

075 / 53:38 — A video edit must survive the next frame

A still image lets us choose a flattering instant. A video makes the object keep its promises in the next frame. Look at the source sequence. As viewpoint and visibility change, we continue to recognize the subject. An edit has to preserve that relationship while introducing its requested transformation.

Remember our two clocks. The sampling steps describe how a model generates a result. The frames here describe time passing in the depicted scene. A system may use many sampling steps to produce a short clip, but those steps are not extra seconds of action. Now the preservation contract must hold across an event, not just inside a frame.

076 / 54:24 — Consistency is not stillness

Would a video be perfectly consistent if every frame were identical? Only if the intended scene were perfectly still. In our moving shot, pose, viewpoint, illumination, and visibility should change. Consistency means that those changes remain coherent with the scene.

The blue stem may become hidden during a turn. That is different from its color changing without a lighting explanation. A texture should move with its surface, rather than crawl independently across it. If there is no convincing explanation, you may have found instability. This is why the goal cannot simply be to minimize change from one frame to the next.

077 / 55:07 — A moving image can reveal an idea

Here is a prepared storyboard, not a generated video result. In the first view, the reflection suggests an object we have not fully seen. A partial view then gives us enough evidence to make a guess. The wide view finally changes our understanding of the setting.

The audience is doing something during those five seconds: forming an expectation, testing it, and revising it. Compare this sequence with revealing everything in the first frame. The same sculpture could be present, but the experience would differ. For AFTER RAIN, the timing of information can carry fragility as strongly as material or lighting. We should decide that structure before asking a tool to fill in motion.

078 / 55:55 — Five seconds can have a structure

Five seconds is short, but it can still have a structure. Our proposed first interval shows a reflection, giving the viewer clues. The second offers a partial view. The final interval reveals the wider setting and changes how the object is understood.

Try another allocation and predict the effect. A longer reflection might create uncertainty; a quick reveal might make the piece feel like a product shot. These timings are artistic proposals, not measured outputs from a video model. By deciding the experience first, we can evaluate whether the generated motion and cuts support it. A technically smooth sequence can still have the wrong rhythm.

079 / 56:40 — An image editor can read a contact sheet

The research begins with a surprisingly simple experiment: arrange video frames into a contact sheet and give that image to an image editor. This asks whether an existing image-editing capability can transfer across multiple views presented together.

Treat the result as evidence that motivates a research direction. A grid can make frames available in a shared context, but a successful-looking selection does not prove reliable video editing. There may be flicker between the sampled frames. The next steps investigate how to represent video more effectively and adapt the model, rather than assuming a contact sheet alone solves time.

080 / 57:22 — Four good frames can hide a bad video

But the missing intervals are where some of the most revealing failures live. A feature can jump, disappear, and return between the frames on this page.

Use the contact sheet for what it does well: compare appearance across selected moments and locate regions to inspect. Then review full playback, with closer attention around turns and occlusion. Think of a film review based only on publicity stills. You might judge the costume and lighting, but you have not yet seen the performance. In generative video, that missing performance includes whether the same object continues to exist convincingly from moment to moment.

081 / 58:05 — The grid can live in latent space

The grid can also be constructed in latent space. Instead of only combining visible pixel images, the method encodes frames and arranges their tokens into a virtual image grid. The published example changes a bus into a graphics card.

The key idea is shared treatment of positions within that constructed representation. Do not confuse this step with using a video VAE; they are different design decisions. Putting frames near one another in a representation can help the model exchange information, but it does not impose a guarantee of physical continuity. We still need to inspect what survives across changing views.

082 / 58:48 — A virtual grid is a representation choice

A virtual grid provides a way to arrange information for a model trained around image-like structure. It creates shared context and positional relationships among frame representations. That can be a useful bridge when reusing an image editor.

But arrangement alone does not enforce the laws of motion or the persistence of a hidden object. Those abilities depend on the model, adaptation, data, and other parts of the workflow. Separate the representation choice from the capability claim. A diagram can explain how frames enter a system without proving that the system handles every difficult event. That proof would require appropriate outputs and evaluation.

083 / 59:31 — Repurposing an image model for video

This architecture asks whether an image editor's abilities can be reused for video. The video VAE encodes a clip into a compressed latent representation. Learned projections connect those video latents with the image-editing transformer, so the model can operate on a compatible arrangement of information.

Trace that main route before reading the smaller branches. Adaptation connects them; the paper explores LoRA and full training choices, with an optional enhancement stage. The attractive idea is reuse of editing knowledge. The test is whether that reuse preserves the temporal relationships our shot needs. Architecture explains where information flows; the edited clip shows whether the intended event survives.

084 / 60:16 — Adaptation connects incompatible representations

The image model and video VAE do not necessarily speak the same representational language. Learned projections help connect them. Adaptation then allows the editing transformer to operate usefully with the video representation.

This is a common engineering pattern: reuse a capable component while learning the interface and behavior needed for another task. The benefit is not automatic. A projection must preserve useful information, and the adapted model must learn how the new structure relates to edits. In the published pipeline, decoding returns the edited representation to video. Follow the information through every stage rather than treating the name of the reused model as an explanation by itself.

085 / 61:02 — 45 frames become 12: why?

Here is a counting puzzle with a small trap. In this example, forty-five pixel frames are compressed with a special first frame and a factor of four for the remaining temporal groups. Forty-five is four times eleven, plus one. The latent sequence therefore has eleven plus one, or twelve frames.

Those twelve latent frames can be arranged in the illustrated three-by-four virtual grid. Simply dividing forty-five by four would miss the first-frame convention. This arithmetic belongs to the representation used in this example; it is not a rule for every video model. A small detail in temporal compression changes what the editing model actually receives.

086 / 61:47 — Temporal compression preserves a special first frame

Solve the frame-count relationship step by step. Start with four k plus one equals forty-five. Subtract one to obtain forty-four, then divide by four to get k equals eleven. The latent sequence contains k plus one frames, so its length is twelve.

This arithmetic encodes the special handling of the first frame in the stated convention. It is different from simply dividing the total number of pixel frames by four. Understanding them can prevent mistakes when arranging latent frames into a virtual grid or comparing the representation with the original clip.

087 / 62:26 — Guidance in a published video ablation

The requested edit turns the scene into a cyberpunk workshop with holographic documents. Compare the results before focusing on the setting labels. Which version carries out the requested transformation more completely, and what visible evidence supports your answer?

The paper presents this case as an example where classifier-free guidance improves edit completeness. Its baseline retains source latents, and its implementation includes rescaling. This does not establish one universally best guidance value. Guidance also has a computation cost when it requires another model pass. Relate the example back to our equation: changing a prediction combination affects how strongly the result follows the condition.

088 / 63:09 — What this comparison actually shows

An ablation asks what changes when one component is removed. In the displayed example, guidance makes more of the requested transformation visible. That is useful evidence about this comparison. It is not yet a measurement of reliability across the kinds of shots we might produce for an exhibition.

As a reader, separate the observation from the next question. We can observe a more complete edit here. We would still want to know what happens across other clips, how often preservation suffers, and what the computation costs. You can learn a mechanism from a selected example while designing a stronger test for the decision you actually need to make.

089 / 63:55 — Local editing across frames

This is a local edit: the sheep's face changes and receives a white star-shaped patch. Look at the patch across poses. Does it remain attached to the same facial region, with a plausible change in apparent shape as the head turns?

The most attractive frame is not enough. An identity marker can slide, disappear, or change shape at another moment. Selected frames let us ask the right questions, but the full clip is needed to check continuity between them. For our discussion, identify a small recognizable feature that will make identity drift easier to notice.

090 / 64:35 — Follow the identity marker

Choose a distinctive feature and follow it through the shot: its attachment to the subject, its shape, and its reappearance after partial visibility.

For our pear, the blue stem is a useful witness. It may become hidden as the camera moves; we should allow that. When it returns, it should still belong to the same object. A marker that slides across the surface or reappears in a new shape tells us something a general impression of smooth motion might miss. We are using the marker as a diagnostic aid, while still judging the whole object's identity and the shot's intended motion.

091 / 65:18 — Global stylization across frames

Here the instruction changes the whole sequence into a minimal monochrome sketch. Local object replacement and global stylization allow different degrees of visual freedom. Even with a large style change, motion and scene structure should remain readable.

Look for rules across frames: line density, silhouette treatment, and the handling of depth. Do they feel like one visual language? This connects directly to our collage discussion. A style is more useful to an art director when described through operations that can persist through time. If every frame reinvents those operations, the result may feel unstable even when each still is appealing.

092 / 66:01 — A style can flicker while motion stays correct

A video can preserve object motion while its style flickers. Line density may jump, a paper texture may crawl, or shading may switch between flat and volumetric treatments without a scene explanation.

Review style as a set of temporal rules, just as we reviewed it across surfaces in the collage. Some variation is appropriate when the camera or light changes. The question is whether the material language remains coherent. If your artwork depends on a particular drawing or collage treatment, this check is as important as whether the subject stays in the same place. Motion correctness alone does not establish visual consistency.

093 / 66:44 — Why raw frame difference is misleading

If we subtract one frame from the next, a perfectly correct camera movement can create a large difference. The same surface has simply moved to another pixel location. A more meaningful comparison first aligns corresponding visible content, then measures the remaining appearance difference.

The teaching equation uses a motion warp for that alignment and a visibility mask to exclude occluded regions. We should not demand agreement for content that is hidden or newly revealed. This is an illustrative metric, not a claim about the exact evaluation used by the paper. A low difference can also reward a video that barely moves. A metric must be checked against the task it is supposed to represent.

094 / 67:33 — The frozen video wins the wrong test

Imagine two candidates for our five-second shot. One shows a coherent camera move around the pear with modest frame differences. The other repeats a single beautiful frame. A naive difference score might prefer the frozen version. Our audience would immediately notice that the intended reveal never happened.

That is a useful example of optimizing the measurement while missing the purpose. We need evidence for both coherent appearance and requested motion. Neither can substitute for the other. Each time, one easy-to-observe quality tempted us to stand in for the whole brief. Good evaluation keeps the intended result in view, especially when a convenient score seems reassuring.

095 / 68:18 — A keyframe can anchor the edit

An edited keyframe can make an art direction easier to approve. Instead of describing every material and lighting choice in words, we can point to an image and say, this is the appearance we want. The workflow then carries that approved look through the source clip.

Follow the four stages: select a representative frame, edit and inspect its appearance, propagate the look, and review the resulting motion. Will the material persist during a turn? Will the identity survive a column passing in front of it? This distinction is central to the Runway reading later: an appealing interaction pattern still needs a shot review suited to the artwork.

096 / 69:04 — AlignVid: when the image overrules the text

What if the approved image is so influential that the requested event never happens? Our AlignVid collaboration studies that tension in image-to-video generation. In the published examples, a baseline omits a sunflower or leaves a person standing; the corresponding AlignVid results implement more of the requested event.

The intervention scales queries or keys in selected attention blocks and denoising steps without retraining. In a scalar form, scaling Q by gamma changes the weights to softmax of gamma times Q K transpose over square root d. It changes attention concentration. Classifier-free guidance instead combines predictions with different conditioning; these are distinct operations.

For an artist, the interesting failure is a faithful-looking image that refuses to do what the scene requires. The published frames illustrate that tension; a finished shot still needs temporal review.

097 / 70:00 — A five-second video edit brief

Our five-second brief changes the pear's body to ivory ceramic while retaining the blue stem, camera movement, and position. Reflections may adapt to the new material. The pear must remain the same object after passing behind a column.

Which clause is hardest to verify? The reappearance is a strong candidate, because the system must maintain identity through a period of invisibility. This is more demanding than transferring a visible color from frame to frame. Begin with one clear transformation in a short clip. Then make the critical visibility event part of the review, rather than discovering it only after choosing a favorite result.

098 / 70:44 — The moment the pear returns

The column is the moment of truth. Before the pear disappears, we can inspect its silhouette, material, and stem. During occlusion, there may be nothing visible to compare. When it returns, the model has to make the same object convincing again.

Do not reject correct invisibility as a failure. Instead, compare the identity before and after the event, including how the object emerges at the boundary. A smooth-looking clip can still reveal a subtly redesigned pear on the other side. We are not merely watching for anything strange; we are testing the preservation requirement at the point where the shot makes it hardest to satisfy.

099 / 71:29 — A small addition creates new relationships

Adding a small object creates several new relationships. In this published example, the edit adds a drone. Even if the drone looks convincing by itself, its scale, placement, and movement must fit the scene.

Imagine a camera moving forward while the added object changes size in the wrong direction. The object might be beautifully rendered, but its relationship with the camera would expose the edit. Or it might hover at an unintended location relative to other objects. When you review an addition, evaluate those relationships across time. A close-up of the inserted object cannot answer every question about whether it belongs.

100 / 72:12 — An addition must belong to camera motion

An added object must fit the camera as well as the scene. A convincing texture and shape in one still do not establish a coherent trajectory. If the camera approaches, the apparent size and position of the addition should evolve in a compatible way.

The exact expectation depends on whether the object is stationary or moving independently. That is why the brief should specify the intended relationship. Inspect the addition relative to nearby objects and the background, not only in a crop. A production-quality edit is a collection of relationships that survive over time, rather than a new object that looks impressive in isolation.

101 / 72:56 — The shot review

Review the whole shot at normal speed, then inspect moments where failure is most likely. Pay attention to identity, intended motion, occlusion, boundaries, and the ending. High-motion content is a limitation the research authors specifically identify.

If a clip fails, choose a revision based on the failure. You might reduce the transformation, shorten the segment, add a reference frame where supported, or repair part of the result through conventional compositing. Another generation is useful when it tests a reasoned hypothesis. Do not let repeated attempts distract you from whether the piece still communicates the artistic intention you began with.

102 / 73:38 — A failed shot suggests a different workflow

A failed shot is information about the workflow. If the appearance is wrong from the beginning, revising the keyframe or the material direction may help. If the first frame is convincing but identity breaks after occlusion, the problem calls for temporal review and a more suitable propagation or editing approach.

If a region must remain exact, direct compositing may be part of the solution. We can also split a complicated transformation into stages, provided the transitions remain coherent. The important move is to connect the observed failure to the next intervention. Repeating a vague request with more enthusiasm does not use what the failed shot has taught us.

103 / 74:24 — Discuss: the difficult moment

Return to the proposed five-second ceramic-pear shot. Which moment would you inspect most closely? What should remain recognizable after the column? What failure would make you reject an attractive clip?

Discuss these three questions by describing an event and the evidence you would seek. If your answer is temporal consistency, make it visible: what changes, when, and why would that violate the intended shot? Pause the video here. Compare your acceptance criteria with another person’s before we examine the compact review plan.

104 / 74:59 — A five-second test plan

Here is a compact test plan. The request is an ivory ceramic body with the blue stem retained. Camera motion and object motion should follow the source. The clip includes a column so that we can inspect a difficult reappearance.

The evidence includes normal-speed playback and a closer look before and after that event. We also inspect reflections because a material change has optical consequences. This plan does not require a long benchmark suite. It simply makes the request, the likely failure, and the acceptance evidence explicit. You can use the same structure to test a different subject or visual transformation.

105 / 75:42 — Three decisions without a model

Let us check whether the mechanisms are now useful without a model in front of us. Consider the three questions on screen. A material change reaches the reflection: explain why. A subject reference and LoRA both help represent a concept: explain the difference. Four frames look excellent: explain why the clip can still fail.

Take a short silent beat before answering. The next slide names the decisions hiding inside the questions. If you get stuck, locate the type of problem first: the physical scene, the model's information, or evidence across time. That is often enough to begin a clear answer.

106 / 76:25 — Match the problem to the intervention

Here are the decisions those questions conceal. If the reflection must change with the new material, we need to define the permitted region and consequences of the edit. If we want a concept for this request, a reference can condition generation; if we want learned adaptation, LoRA changes selected weights. If continuity matters, the evidence must include the intervals between our chosen stills.

Each decision rules out a tempting shortcut. Freezing every surrounding pixel may freeze the wrong reflection. Calling every supplied image training confuses conditioning with adaptation. Calling four stills a successful video omits time. Use this slide to repair your explanation, rather than memorize a preferred sentence.

107 / 77:11 — The mechanisms behind your decisions

The material changes how light interacts with the pear, so its visible consequences can extend into the reflection. A reference conditions an output; LoRA learns a low-rank update to selected model weights. Four good frames leave unobserved transitions, including events such as occlusion where identity can fail.

Those are compact answers, but each has an application. They help us write a better preservation contract, choose between two kinds of intervention, and design a more revealing review. If your answer used different words and preserved those distinctions, it works. Tomorrow's interface may rename its controls, but the difference between specifying a request, adapting a model, and checking its result will still matter.

108 / 77:58 — Three levels of checking an edit

We can check an edit at three levels. First, did the requested transformation occur? Second, did the necessary invariants survive? Third, does the result serve the artwork's intention?

A result can pass the first level and fail the second, as when the correct material appears on a different object. It can pass both technical levels and still fail the artistic one, as when the chosen effect undermines the intended mood. Keeping the levels separate makes critique more precise. It also explains why neither a generic aesthetic score nor a strict pixel comparison can answer every question we care about.

109 / 78:40 — Who decides?

We have spent the lecture making the editing request more precise. Now consider who gets to choose the request in the first place. A system can offer convincing alternatives, but somebody still decides which intention matters and which evidence counts as success.

The three readings let us examine that responsibility from different positions. Pachocki raises questions about capable AI systems and human values. Koe asks readers to reconsider their own goals and habits. Runway presents a workflow for turning an approved image into a video edit. Keep our pear in mind as we read them: we can delegate parts of its production while still arguing about what the artwork should become.

110 / 79:27 — After Topic 3: judge the whole shot

We have an approved appearance and a moving shot to judge. What can an approved keyframe establish, and what can it not? Why might a low frame-difference score reward a bad video? Which evidence would convince you that identity survived motion?

Discuss these three questions. Your answer should account for the requested event as well as the subject’s appearance. Use the column, a turn, or the sketch treatment as a concrete example. Pause the video here. We will next ask how the readings change our view of responsibility for those decisions.

111 / 80:06 — AI goals and human direction

These first two readings ask about direction from different sides. An Alien Mind considers the relationship between an AI system achieving a goal and respecting human values. Dan Koe's essay invites people to examine the goals and habits directing their own lives. We will use its proposal as material for critical discussion of creative practice.

Find a specific claim, explain your interpretation, and connect it to a decision from the lecture. A fictional artist is fine for discussing personal direction. The fifth quiz question asks you to reflect on one reading of your choice. Our class discussion can compare all three perspectives without requiring everyone to disclose personal experiences.

112 / 80:52 — Reading 3: image-guided video editing

The third reading moves from questions about direction to a concrete editing workflow. Runway's announcement of Aleph 2.0 and Edit Studio describes establishing an appearance in an edited image and applying that change through video. It connects directly to our keyframe discussion.

Read it with the ceramic pear and column in mind. Which decision does the interface make easier to express? Which result would you still need to watch before approval? The page is a vendor's account, so its demonstrations illustrate claimed capabilities rather than supply an independent comparative evaluation. We can learn from the interaction design and still propose an occlusion test that asks whether the workflow meets this exhibition's particular needs.

113 / 81:40 — Optional technical reading

These technical papers are optional companions to the discussion readings. Choose according to the question you want to investigate. InstructPix2Pix helps explain edit supervision; Flow Matching develops the generative training framework. Qwen-Image-2.0 and Qwen-Video-Edit provide recent architecture examples.

You do not need to read all of them before trying the class discussion. Start with a question, locate the part of the paper that addresses it, and distinguish the proposed method from the authors' evidence. The slide notes retain the other references on multi-image inputs, rewards, composition, and video editing.

114 / 82:18 — Three source types, three kinds of evidence

The three readings provide different kinds of material. A research leader's perspective develops an argument about alignment and future development. A reflective essay offers a way to examine personal direction. A vendor announcement presents a workflow and promotes its capabilities.

Read each according to what it can support. Identify forecasts, personal claims, and demonstrations rather than treating every sentence as the same kind of evidence. You may disagree with an author and still find a useful question. Our synthesis is practical: what do you want to make, what will you delegate, and what evidence will you use to decide whether the collaboration served that intention?

115 / 83:03 — Separate image and text guidance

This technical extension separates image guidance and text guidance in InstructPix2Pix. Begin with the prediction using neither condition. Add a scaled difference for including the image, then another scaled difference for adding the text alongside that image.

Set both scales to one and follow the cancellation: the intermediate terms disappear, leaving the fully conditioned noise prediction. This is a useful check that you understand the expression. Epsilon here denotes a noise prediction, unlike the velocity in our flow-matching slides. These controls belong to this formulation and should not be assumed to map directly onto every current editor's interface.

116 / 83:45 — What remains when the terms cancel?

Now set both guidance scales to one and follow the cancellation. The initial no-condition prediction cancels its negative copy in the first difference. The image-only prediction then cancels its negative copy in the second difference. The result is the prediction with both image and text conditions.

That gives us a reference point for understanding the two controls. If the text scale were zero while the image scale stayed one, we would instead recover the image-only prediction. It shows what information each difference adds. Keep epsilon's meaning explicit here: this InstructPix2Pix expression predicts noise, whereas the earlier flow-matching example predicted velocity.

117 / 84:28 — Training: the target supplies the lesson

Read the pseudocode in two groups. First we prepare an example: source, instruction, target, conditions, and target latent. Then we sample noise and time, construct the intermediate latent, predict velocity, and compare that prediction with the known training target.

The final update changes trainable weights based on the loss. At inference, there is no known edited target to supply in this way. We instead follow the learned generative predictions from a starting state. This pseudocode explains the conceptual loop; it omits practical engineering such as batches, precision, and scheduling. Its most important distinction is what information is available during learning versus use.

118 / 85:11 — Inference: where did the target go?

Look for the line that disappeared. The inference loop has no known edited target and no loss that updates weights. Instead, we encode the source and instruction, begin with noise, repeatedly predict a velocity and move the latent, then decode the result.

Training uses examples to adjust the learned model. In this simplified inference loop, we use that model to construct an output we do not yet possess. The code is conceptual; practical systems use their own representations and sampling schedules. But you can now explain why providing another reference changes the information available for a request without automatically becoming a model-training operation. It is the same distinction we used when choosing between references and LoRA.

119 / 86:00 — Choose the approach by the requirement

Before we turn to the readings' discussion questions, use this table as a compact decision aid. Exact untouched pixels suggest a role for masks and compositing. A new visual language may benefit from reference conditioning. A reusable learned concept may motivate adaptation. A transformation that must survive motion needs a video workflow and temporal evidence.

These approaches can cooperate in the same artwork. Choose according to the contract, then inspect the failure most likely to undermine it. That leaves us with a larger question for the readings: once production becomes easier, how do we choose a worthwhile direction and retain meaningful judgment over the result?

120 / 86:45 — An Alien Mind

An Alien Mind distinguishes achieving an assigned goal from generalizing human values in unfamiliar circumstances. Pachocki also raises questions about monitoring increasingly capable systems. Treat the essay as an argument containing claims and forecasts that we can examine.

For our discussion, imagine an exhibition team delegating production while retaining responsibility for what the work communicates. Discuss the three questions on screen: explain the goal-and-values distinction, identify a forecast and the evidence it would need, and defend a boundary for human control. Pause the video here. Our art-direction example is an analogy; it does not give an image editor the same agency or risk profile as an autonomous research system.

121 / 87:31 — How to fix your entire life in 1 day

Koe's title makes a dramatic promise: How to fix your entire life in one day. The essay proposes examining identity and goals, interrupting habitual behavior, and turning reflection into action. We can discuss the usefulness of that proposal without accepting the title as a guarantee.

Imagine an artist who can generate a hundred attractive images but cannot choose what to make. Does easy production help clarify a direction, or make avoiding the decision easier? Discuss the three questions on screen. Choose an idea worth using or challenging, explain the reason, and connect it to creative intention. Pause the video here. You can use a fictional artist or public example; no personal disclosure is needed.

122 / 88:20 — Introducing Aleph 2.0 and Edit Studio

Runway's reading proposes approving an edited image before applying its appearance through a video. It offers a concrete answer to a communication problem: an art director can point to the desired look, rather than describe every feature in words.

Now bring back the column. A convincing keyframe does not tell us what the pear will look like after it reappears. Discuss the three questions on screen: what the frame establishes, which claim deserves a harder test, and how the workflow serves AFTER RAIN's intention. Pause the video. The source is a vendor announcement; our job is to distinguish a useful demonstrated workflow from a reliability claim still needing evaluation.

123 / 89:06 — Your final judgment

Return to your first judgment of the glass pear. We began with a small request and discovered that it touched the scene's physics, the artwork's intention, and the audience's experience over time. You now have more precise ways to say what should change, what should survive, and how to judge the result.

Finish with the three questions on screen. Where would you place the boundary between control and surprise? How would you balance technical success and artistic purpose? What would you delegate, and what evidence would you require? Pause the video for the final discussion. Connect one mechanism or reading to a concrete artistic decision. Listen for an answer that makes you revise your own. That revision is a fitting last act for a class about editing.