

# AI art, images and video

Full English narration and Chinese companion explanations | 100 slides | 22 September 2026

## How to use this transcript

The English text is the complete narration script. The Chinese text is a concise companion explanation, not a word-for-word translation. Timestamps refer to the packaged 63:47 narration, including short transitions. Pause for discussion where useful. The audio uses a synthetic English voice.

## Reading sequence

### Survey of Video Diffusion Models: Foundations, Implementations, and Applications

Yimu Wang et al. | 2026 revision; TMLR

[Read the original paper](#)

### Diffusion-Based Visual Art Creation: A Survey and New Perspectives

Bingyuan Wang, Qifeng Chen and Zeyu Wang | ACM Computing Surveys, 2025; preprint 2024

[Read the original paper](#)

### Generative AI for Film Creation: A Survey of Recent Advances

Ruihan Zhang et al. | CVPR 2025 CVEU workshop

[Read the original paper](#)

### State of the Art on Diffusion Models for Visual Computing

Ryan Po et al. | Computer Graphics Forum, 2024

[Read the original paper](#)

### Advances in Artificial Intelligence: A Review for the Creative Industries

Nantheera Anantrasirichai, Fan Zhang and David Bull | Artificial Intelligence Review, 2026

[Read the original paper](#)

# 01 / Generative AI for art, images, and video

00:00:00 - 00:00:36

## Generative AI for art, images, and video

AMCC5160  
Lecture 04  
22 September 2026



Zhang et al., Film Creation, CVPR Workshop 2025, Fig. 6

01

Welcome to Lecture Four of AI-Driven Animation and Video Generation. Today we move from understanding an image generator to making deliberate choices across an artwork or a film. We will follow five survey papers. Surveys are useful because they organize many individual methods into a map of the field. Our purpose is to read that map and connect it to creative work. Throughout the lecture, ask two questions. What information does a method need? And what decision does it help an artist make? The figures come from the papers, while the studio example we develop together is a classroom exercise.

### 中文讲解

欢迎来到《AI驱动的视频与动画生成》第四讲。今天，我们从理解图像生成器，走向在艺术作品或影片中作出有意图的选择。课程依次讨论五篇综述。综述把许多方法组织成一张研究地图，我们的任务是把这张地图与创作联系起来。请始终思考两个问题：一种方法需要什么信息？它帮助艺术家决定什么？课程中的图表来自论文，而贯穿全讲的画室场景是课堂构想。我们不仅关心画面是否漂亮，也关心方法能否支持具体的创作决定。



## 02 / Five readings, one creative process

00:00:36 - 00:01:12

### Five readings, one creative process

- 01 Video models
- 02 Visual art
- 03 Filmmaking
- 04 Visual computing
- 05 Creative industries

Wang et al., Video Diffusion Survey, rev. 2026

02

Here is our route. We begin with video models and the problem of generating a sequence that holds together through time. We then consider visual art, where the intended experience determines which controls matter. The filmmaking survey shows how creators combine these tools in actual production workflows. Visual computing connects images and video to objects, scenes, and movement data. Finally, the creative-industries review places generation within a wider production process. You do not need to memorize every model name. By the end, you should be able to choose an appropriate kind of input, explain a likely limitation, and describe how you would judge the result.

#### 中文讲解

课程先讨论视频模型，理解如何生成在时间上连贯的序列；再讨论视觉艺术，看看预期体验如何决定所需控制。电影综述展示创作者如何在实际制作中组合工具。视觉计算将图像与视频连接到物体、场景和动作数据。最后，创意产业综述把生成放回完整的制作过程。无需记住所有模型名称。学完后，你应能选择合适的输入类型，解释可能的局限，并说明如何判断输出是否符合目标。

## 03 / What changes when an image moves?

00:01:12 - 00:01:50

### What changes when an image moves?



*"The ball is in blue fire and turns into a dragon in fire."*



*"A man rides on a huge fish, flying from the water into the sky."*

A still image leaves time open.  
A video chooses what happens next.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 13

03

What changes when an image starts to move? A still image suggests possibilities, but it leaves the next moment open. A video commits to a sequence of events. If a character raises a hand, that hand needs to belong to the same person as it moves. If the camera passes a chair, the chair should remain in a plausible position. These requirements involve appearance, movement, and spatial relationships. They also involve meaning. Holding on an empty room for ten seconds creates a different experience from showing it for one second. Time therefore adds both technical demands and artistic choices. A sequence can fail in either respect.

### 中文讲解

图像开始运动后，发生了什么变化？静态图像留下许多可能，而视频必须决定下一刻发生什么。人物抬手时，那只手在运动中仍应属于同一个人；镜头经过椅子时，椅子应保持合理位置。这涉及外观、动作和空间关系，也涉及意义。空房间停留十秒与一秒，会带来不同感受。因此，时间同时增加技术要求和艺术选择。一个序列可能在技术上出错，也可能技术上流畅却未能表达预期含义。

## 04 / Our running example

00:01:50 - 00:02:25

### Our running example

An empty studio.

A painting begins to move.

Someone returns.

Classroom exercise · inspired by the readings

04

We will use a simple studio scene throughout the lecture. The room is empty. A painting begins to move. Then someone returns. This is an invented teaching example, so there is no correct finished version to reproduce. You might imagine a frightening scene, a quiet memory, or a playful animation. Start by picturing what should stay stable. Perhaps the room layout and painting frame remain fixed while only the painted surface changes. Then consider what the returning person should notice. These decisions will help us compare methods. A tool becomes useful when we can say which part of this scene it should control.

#### 中文讲解

整讲使用一个简单场景：空画室里，一幅画开始运动，然后有人回来。这是教学构想，没有唯一正确的完成版本。你可以把它想象成恐怖场景、安静的回忆或有趣的动画。先考虑哪些东西应保持稳定，例如房间布局和画框不变，只改变画面内部。再考虑回来的人应注意到什么。这些决定能帮助比较方法。只有说清楚工具要控制场景的哪个部分，才能判断它是否有用。

# 01

## Video generation models

Wang et al.  
Survey of Video Diffusion Models  
2025, revised February 2026

Wang et al., Video Diffusion Survey, rev. 2026

05

Our opening reading is Wang and colleagues' Survey of Video Diffusion Models, first posted in 2025 and revised in February 2026. It covers foundations, implementations, and applications. We will follow its progression from the generation mechanism to inputs, control, editing, and connections with three-dimensional scenes. A survey figure often contains several methods at once. Read such a figure as a set of relationships, rather than as one model you must implement. Also pay attention to dates. A survey can contain important earlier examples alongside newer research. Its publication or revision date does not make every comparison a current ranking of available systems.

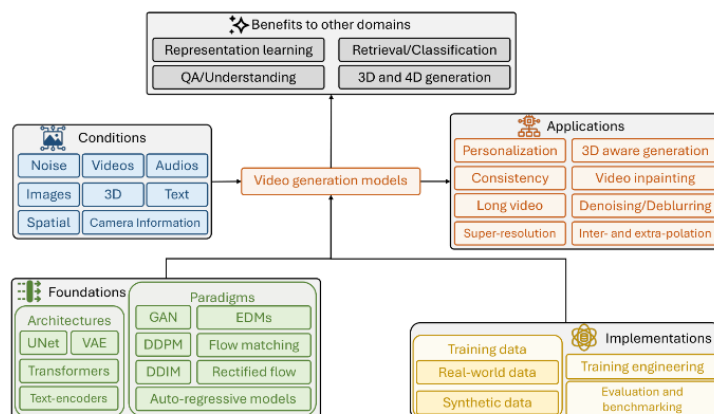
### 中文讲解

第一篇是Wang等人的视频扩散模型综述，2025年首次发布，2026年2月修订。它涵盖基础、实现和应用。我们从生成机制走向输入、控制、编辑以及三维场景。综述中的图可能同时组织多种方法，应把它看作关系图，而不是必须实现的单一模型。也要注意日期：重要的早期示例可以与较新的研究同时出现。论文的发表或修订时间，并不意味着其中每项比较都是当前可用系统的最新排名。

# 06 / The field at a glance

00:03:06 - 00:03:43

## The field at a glance



Inputs guide a model.  
The model produces a sequence.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 1

06

At a high level, video generation connects inputs to a model and then to a sequence. The inputs might include a sentence, an image, a source video, sound, or camera information. Each input supplies a different kind of constraint. A sentence can describe a mood or action, but it usually leaves many visible details unspecified. An image supplies more exact information about appearance. A source video can supply a performance or a movement pattern. When reading the overview, look for what enters the model and what comes out. This simple habit makes a complicated architecture easier to understand and helps you recognize what a demonstration actually controls.

### 中文讲解

从整体看，视频生成把输入送入模型，再得到序列。输入可以是句子、图像、已有视频、声音或相机信息。它们提供不同约束。句子能描述情绪或动作，却往往未指定许多可见细节。图像更具体地给出外观，源视频则能提供表演或运动模式。阅读总览时，先找出什么进入模型、什么离开模型。这个习惯能让复杂结构更易理解，也帮助我们判断演示究竟控制了什么。

## 07 / Three routes to a video

00:03:43 - 00:04:20

### Three routes to a video

Text describes a scene.

An image fixes a starting appearance.

A video supplies existing movement.

Wang et al., Video Diffusion Survey, rev. 2026

07

Consider three routes to the studio scene. With text-to-video, we describe the room and the moving painting, and the model invents much of their appearance. With image-to-video, we first choose a picture of the room, then ask for movement from that starting appearance. With video-to-video, we begin with an existing sequence and change selected properties, such as its visual style. None of these routes automatically solves every production problem. An initial image does not guarantee stable identity later in the shot. A source video does not guarantee exact motion preservation. Choose the route according to the information you already have and the properties you need to retain.

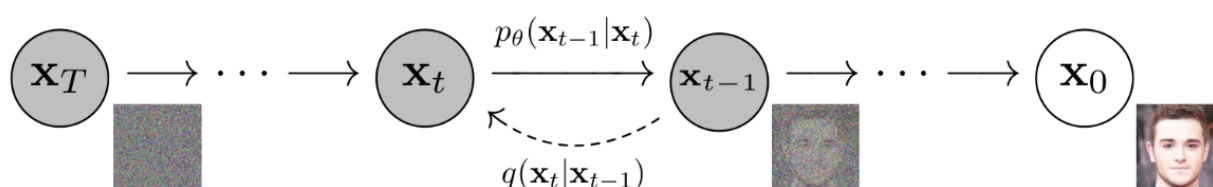
#### 中文讲解

生成画室视频有三条路线。文生视频用文字描述房间和运动的画，由模型补充大量外观。图生视频先选择房间图像，再从该外观出发产生运动。视频到视频则从已有序列出发，改变风格等属性。任何路线都不会自动解决所有制作问题：首帧不保证后续身份稳定，源视频也不保证动作完全保留。应根据已经拥有的信息，以及必须保留的属性来选择路线。

## 08 / Diffusion in one picture

00:04:20 - 00:04:58

### Diffusion in one picture



Training teaches noise removal.  
Generation refines a noisy sample.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 2

08

Diffusion is easiest to understand by separating training from generation. During training, a system encounters examples with added noise and learns how to predict information needed to remove that noise. During generation, it starts with a noisy sample and repeatedly refines it using the learned model. The diagram shows a succession of states, rather than a film of a face changing over time. The prompt or another condition influences the refinement. The model does not learn a new set of weights every time you generate an image. It uses what training has already established. This distinction helps explain why changing a prompt differs from adapting the model.

### 中文讲解

理解扩散，需要区分训练与生成。训练时，系统接触加入噪声的样本，学习预测去噪所需的信息。生成时，它从噪声出发，利用已经学到的模型反复细化。图中是计算状态的变化，不是人脸随影片时间变化。提示词或其他条件影响细化过程。每次生成图像时，模型通常不会重新学习一组权重，而是使用训练建立的知识。因此，修改提示词与适配模型是不同的操作。

## 09 / A video has two kinds of time

00:04:58 - 00:05:36

### A video has two kinds of time

Frame time: when an event happens.

Denoising time: how a sample improves.

Wang et al., Video Diffusion Survey, rev. 2026

09

Video diffusion has two different meanings of time. Frame time is the time inside the represented event: a person enters, walks across the studio, and looks at the painting. Denoising time is the sequence of computational refinement steps that produces the sample. A denoising step is not an extra frame in the finished video. Depending on the architecture, one step may update information for many frames together. Confusing these two timelines makes diagrams difficult to read. Whenever you see a time index, ask whether it identifies a moment in the scene or a stage in noise removal. The answer tells you what relationship the model is trying to represent.

#### 中文讲解

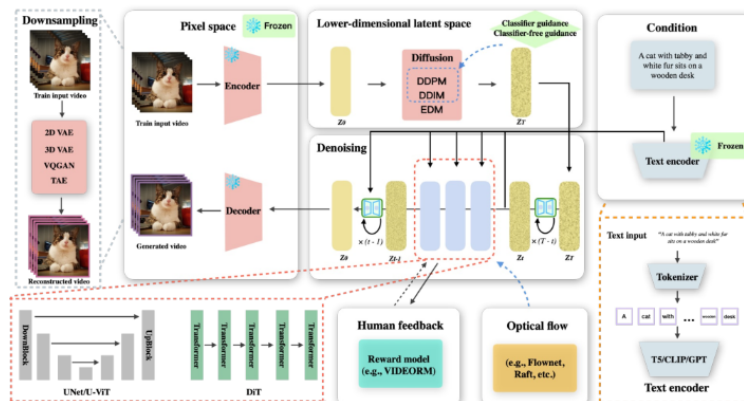
视频扩散包含两种时间。帧时间是场景里的时间：人物进入、穿过画室、看向画作。去噪时间是计算中的细化步骤。一次去噪步骤不是成片中的新增一帧；某些结构会同时更新许多帧的信息。混淆两种时间，会使模型图难以理解。看到时间下标时，应先问它代表事件发生的时刻，还是噪声去除的阶段。答案决定了模型正在表示哪一种关系。



# 10 / The generation pipeline

00:05:36 - 00:06:15

## The generation pipeline



A prompt guides refinement in a compressed space.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 4

10

The generation pipeline often works in a compressed space. First, a text encoder converts the prompt into a representation that the model can use. The generative network then refines a noisy representation under that condition. A decoder converts the result into visible frames. This division reduces the amount of information processed by the most expensive part of the system. It also means that the final image depends on several components. A problem with fine texture may involve compression or decoding as well as generation. For our studio scene, the prompt supplies an instruction, while the latent representation carries the evolving visual content during refinement.

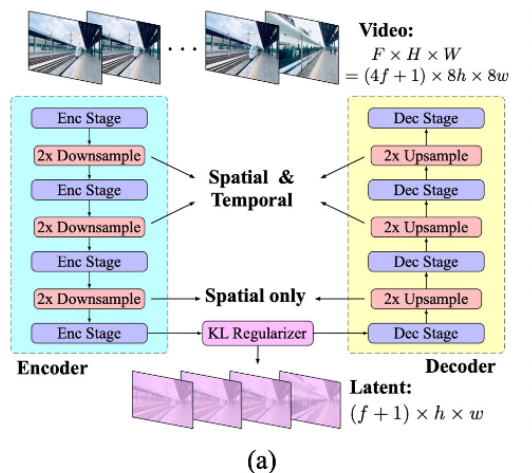
### 中文讲解

生成流程通常在压缩空间中工作。文字编码器先把提示词变成模型可用的表示，生成网络在条件引导下细化带噪表示，解码器再将结果转成可见帧。这样可以减少最昂贵阶段处理的信息量。但最终图像也因此依赖多个组件，细节问题可能与压缩或解码有关，不一定只出在生成网络。画室例子中，提示词提供指令，潜在表示则在细化过程中承载不断变化的视觉内容。

# 11 / A latent is a compressed representation

00:06:15 - 00:06:53

## A latent is a compressed representation



The model works with fewer values than raw pixels.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 7 (detail)

11

A latent is a compressed representation of the data. Imagine describing the essential structure of a room with fewer values than you would need to list every colored pixel in every frame. The encoder maps the original material into that representation, and the decoder maps it back. The figure shows compression across space and, in some stages, across time. Compression makes a large video more manageable, but it can also discard detail. A latent is not a miniature image that we can always interpret directly. It is a learned representation. Keep the practical tradeoff in mind: less computation can come with limits on small details and reconstruction fidelity.

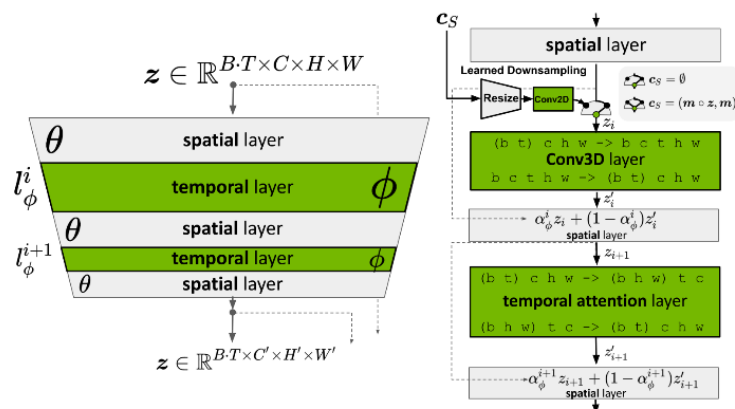
### 中文讲解

潜变量是数据的压缩表示。可以想象，我们用比逐个列出每帧像素更少的数值，表示房间的重要结构。编码器把原始素材映射进去，解码器再映射回来。图中有空间压缩，也有部分时间压缩。压缩使大型视频更易处理，但可能丢失细节。潜变量并不总是一张人能直接理解的小图片，而是学习得到的表示。减少计算量的同时，需要注意微小细节和重建保真度的限制。

## 12 / Spatial attention and temporal attention

00:06:53 - 00:07:30

### Spatial attention and temporal attention



Spatial: relationships within a frame.  
Temporal: relationships across frames.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 5

12

Spatial relationships occur within a frame. They include the position of a person relative to a door, or the relationship between a hand and the object it holds. Temporal relationships connect different frames. They help a model relate the hand now to the hand a moment later. The diagram separates spatial and temporal processing to make those roles visible. Attention is one way to connect relevant features, but it does not guarantee correct understanding of the scene. For the studio example, spatial processing helps organize the room at a particular instant. Temporal processing helps the painting and furniture maintain meaningful relationships as the sequence develops.

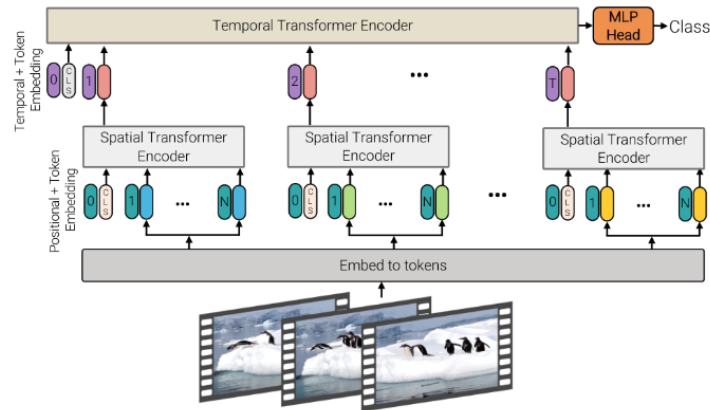
#### 中文讲解

空间关系发生在同一帧内，例如人相对于门的位置，或手与所持物体的关系。时间关系连接不同帧，例如此刻的手与稍后的手。图中分开显示空间和时间处理。注意力是连接相关特征的一种方法，但不保证模型真正正确理解场景。在画室例子中，空间处理组织某一时刻的房间，时间处理则帮助画作和家具在序列发展中保持有意义的关系。

# 13 / A second way to connect frames

00:07:30 - 00:08:10

## A second way to connect frames



A frame encoder handles appearance.  
A temporal encoder connects the sequence.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 6

13

This figure illustrates another way to connect frames: encode information within each frame, then connect those representations across the sequence. The pictured architecture is a video-understanding example with a classification output, rather than a complete video generator. We are using it to understand an architectural idea. The lower part handles appearance within individual frames. The upper part combines information through time. Separating these roles can make processing easier to organize. When you encounter a research diagram, check the output label before assuming its task. Similar building blocks can appear in recognition and generation systems, even though their training objectives and outputs differ substantially.

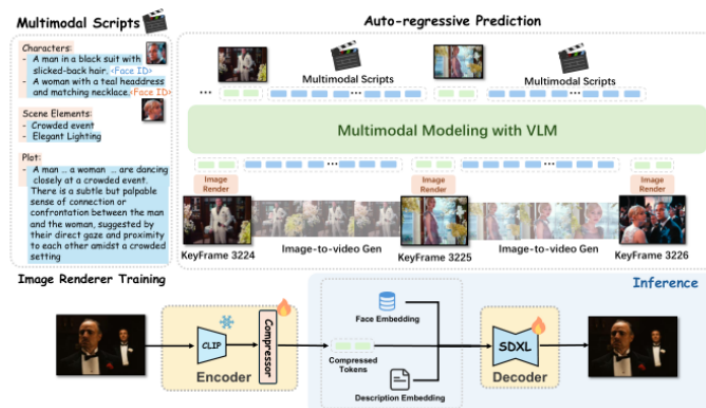
### 中文讲解

另一种连接帧的方法，是先编码每帧内部的信息，再在时间上连接这些表示。图中的具体结构用于视频理解，末端输出分类，并不是完整的视频生成器。这里借它理解结构思想：下层处理单帧外观，上层组合时间信息。阅读研究图时，先看输出标签，确认任务。识别和生成系统可能使用相似组件，但训练目标和输出不同。不能因为看到熟悉的模块，就把整张图当成生成流程。

# 14 / Predicting a sequence in parts

00:08:10 - 00:08:47

## Predicting a sequence in parts



Autoregressive methods build on earlier outputs.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 21

14

Some approaches build a sequence in parts, using earlier outputs to guide what comes next. This is the basic idea of autoregressive prediction. The diagram includes script information, keyframes, and a rendering process. You can think of the earlier material as context for the next decision. This can help extend a sequence, but errors in the context may also carry forward. If a character's coat changes in one generated part, later parts may preserve the new coat instead of the original one. For filmmaking, that makes checkpoints valuable. Inspect important keyframes and transitions before treating a longer generated sequence as a finished shot.

### 中文讲解

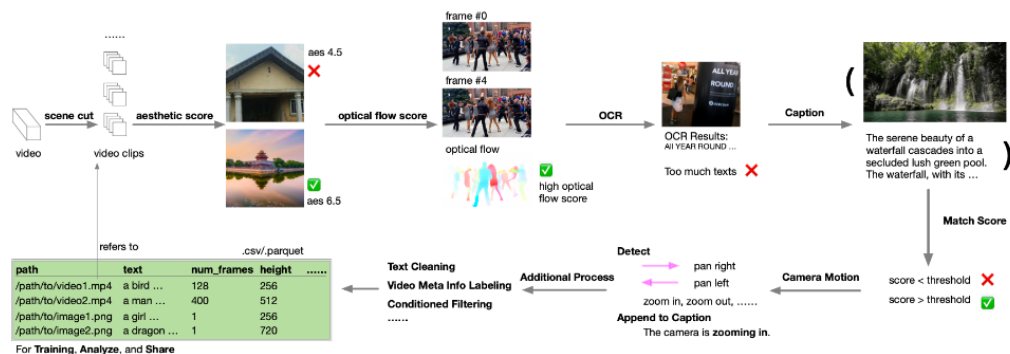
有些方法分段构建序列，以已有输出指导下一部分，这就是自回归预测的基本思想。图中包含剧本信息、关键帧和渲染流程。前面的内容成为后续决定的上下文，但其中的错误也可能延续。如果某段把角色外套改了颜色，后续可能保留错误版本。因此，制作中需要检查点。在把长序列当成完成镜头之前，先查看关键帧和重要转场，避免错误随着生成不断积累。

# 15 / Training data shapes the result

00:08:47 - 00:09:24

## Training data shapes the result

Open-Sora Data Processing Pipeline v1.1



Selection and captions affect what a model can learn.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 8

15

Training data shapes the kinds of results a model can produce. The figure shows a pipeline that selects clips, filters them, and attaches descriptions. Each stage changes what reaches training. A caption can describe an object, an action, or a camera movement. If the caption misses an important event, the model receives a weaker connection between language and that event. Filtering can remove unusable clips, but it also reflects decisions about what counts as desirable material. An artist should therefore avoid treating model behavior as a neutral view of all possible images. It reflects the material, labels, and objectives used to train the system.

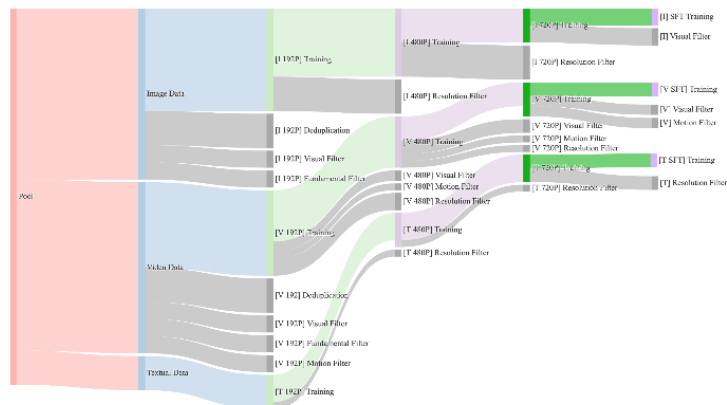
### 中文讲解

训练数据影响模型能够生成什么。图中展示片段选择、过滤和添加描述的流程，每个阶段都改变进入训练的材料。字幕可能描述物体、动作或镜头运动；若漏掉关键事件，语言与事件之间的联系就更弱。过滤能排除不可用片段，但也体现什么材料被认为值得保留。模型行为不是所有可能图像的中立总和，而是训练素材、标注和目标共同作用的结果。

# 16 / A closer look at data filtering

00:09:24 - 00:10:03

## A closer look at data filtering



Filtering trades coverage against usable training examples.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 9

16

The branching diagram makes filtering visible. A broad pool of material becomes smaller groups suited to particular training stages. Some examples may have adequate resolution but little movement. Others may have useful movement but poor captions or visual quality. The important point is that filtering trades coverage against usable examples. Removing difficult material may improve average training quality while leaving gaps in certain actions or styles. This matters when your project depends on an unusual visual language. If a model struggles with that language, repeatedly rewriting the prompt may not address the underlying limitation. The required examples or relationships may be poorly represented in training.

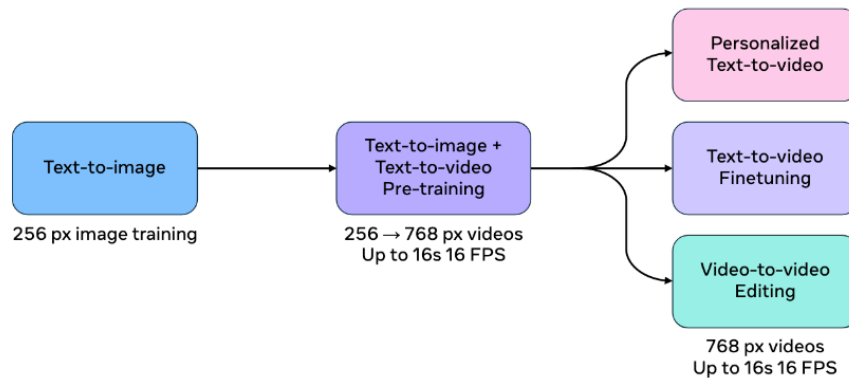
### 中文讲解

分支图把过滤过程可视化：大规模素材池被划分成适合不同训练阶段的群组。有些素材分辨率够高却缺少运动，有些运动有用但字幕或质量较差。过滤是在覆盖面与可用样本之间取舍。排除困难材料可能提高平均质量，也可能留下某些动作或风格的空白。如果模型不擅长你的特殊视觉语言，反复改写提示词未必能解决根本问题，因为训练中可能缺少所需的例子或关系。

# 17 / One foundation, several tasks

00:10:03 - 00:10:41

## One foundation, several tasks



Generation, personalization, and editing can share a base.

Wang et al., Video Diffusion Survey, rev. 2026

17

A single foundation can support several tasks. The figure moves from image training toward video training and then toward specialized applications such as personalization or editing. Shared foundations let researchers reuse learned visual knowledge instead of starting from nothing for every task. However, task support still depends on the model and the way it is adapted. Generating a new character and preserving a particular existing character are different requirements. For our studio scene, a general model might create an attractive room, while a specialized workflow might better preserve the same painting across shots. Shared origins do not make those capabilities interchangeable in practice.

### 中文讲解

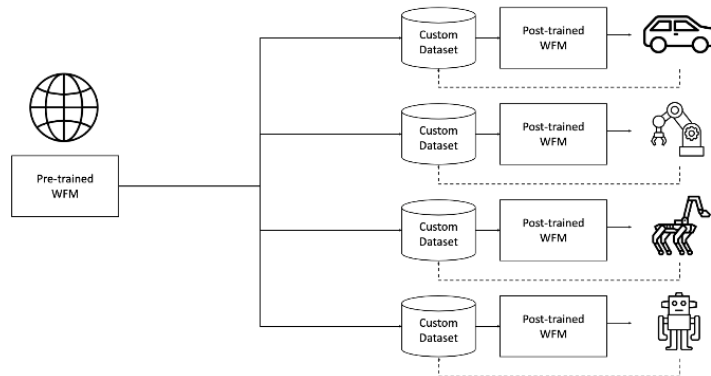
同一个基础模型可以支持多种任务。图中从图像训练走向视频训练，再走向个性化和编辑。共享基础让研究者复用视觉知识，但具体能力仍取决于模型和适配方式。生成一个新角色与保留某个已有角色，是不同要求。普通模型也许能生成漂亮的画室，专门流程可能更适合跨镜头保留同一幅画。共同的训练起点，并不意味着这些能力在实际制作中可以互换。



# 18 / A foundation model can adapt to a domain

00:10:41 - 00:11:21

## A foundation model can adapt to a domain



Further training specializes a general model.

Wang et al., Video Diffusion Survey, rev. 2026

18

This diagram from the Cosmos discussion shows a pretrained foundation model branching into domain-specific adaptations. The custom datasets concern physical settings such as vehicles and robots. Follow the diagram from the common base toward the separate applications. Further training helps specialize the model to the material and tasks of a domain. Adapting a model to an artistic style is a related idea, but this particular figure does not report an art experiment. The general lesson is that specialization requires suitable evidence. A small, carefully chosen dataset may teach a useful pattern, while an inconsistent collection may introduce unwanted associations or reduce the usefulness of the adaptation.

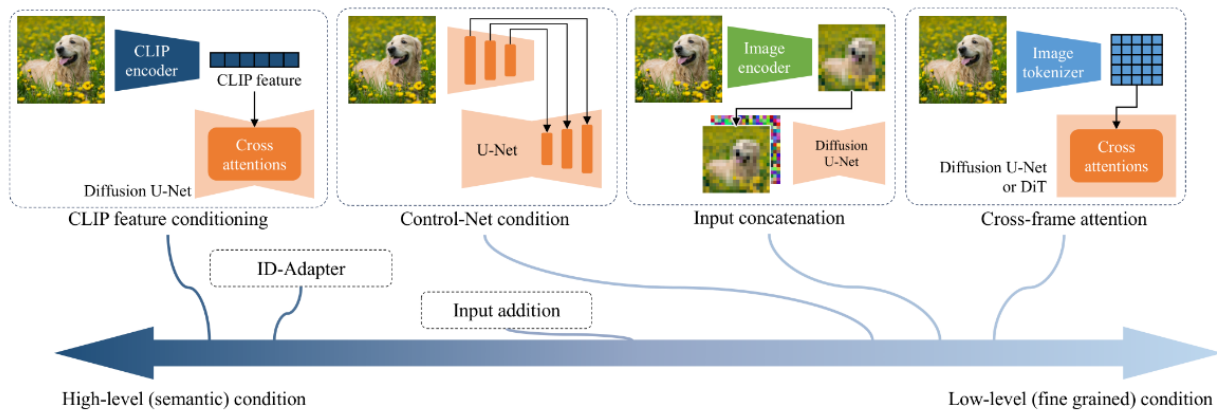
### 中文讲解

Cosmos相关图示展示基础模型如何分支适配不同领域，定制数据涉及车辆和机器人等物理环境。进一步训练使模型更适合某类任务和材料。适配艺术风格与此有相近思想，但这张图本身并未报告艺术实验。关键是适配需要合适的证据。经过选择的小型数据集可能教会有用模式，混杂的素材则可能引入不希望出现的关联，降低适配的实际价值。

## 19 / References carry different information

00:11:21 - 00:11:59

### References carry different information



A reference can suggest meaning, layout, or appearance.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 12

19

A reference image can communicate several kinds of information. It may suggest broad meaning, such as a dog in a garden. It may specify appearance, such as the exact markings of one dog. Or it may constrain structure, such as the location and outline of the subject. Different conditioning methods preserve these properties to different degrees. The figure organizes several ways to introduce image information. You do not need to memorize every connection. Instead, ask what you want the reference to contribute. For the studio scene, a mood image and a precise room-layout reference serve different purposes, even if both enter the workflow as pictures.

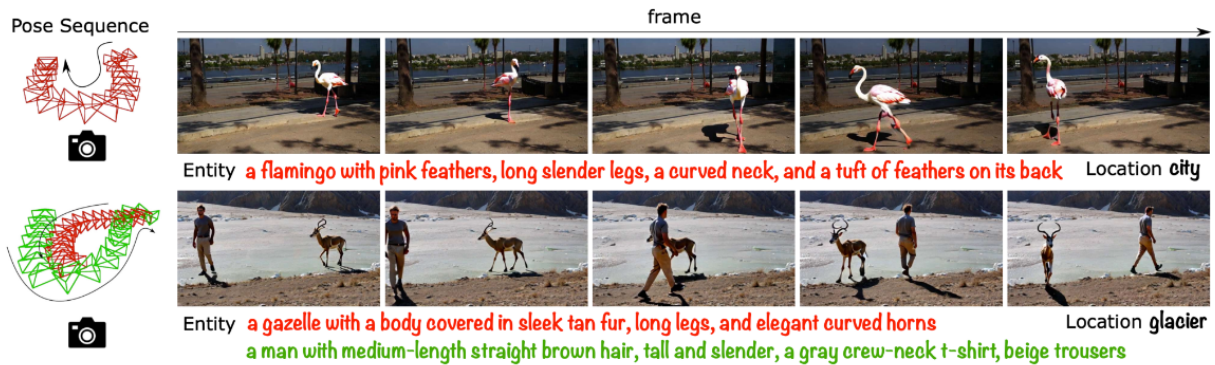
### 中文讲解

参考图可以提供不同信息：大致语义，例如花园里的狗；具体外观，例如某只狗的斑纹；或结构，例如主体的位置和轮廓。不同条件注入方法保留这些属性的能力不同。图中组织了若干引入图像信息的方法，无需记住每条连线。先问希望参考图贡献什么。画室的情绪参考与精确布局参考，即使都以图片形式输入，也承担不同任务。

## 20 / Motion and camera control

00:11:59 - 00:12:37

### Motion and camera control



An actor can move while the camera stays still.  
A camera can move around a still actor.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 14

20

Object motion and camera motion are distinct. A person can cross the studio while the camera remains still. The camera can also move around a person who does not move. In the resulting images, both situations produce changing pixel positions. A system therefore needs useful constraints if we want to control them separately. The figure combines scene descriptions with camera information. When planning a shot, specify the intended camera behavior independently from the action. Then inspect whether the output respects both. A visually energetic result may still fail if the camera moves when the scene requires a stable viewpoint, or if the subject freezes during the intended action.

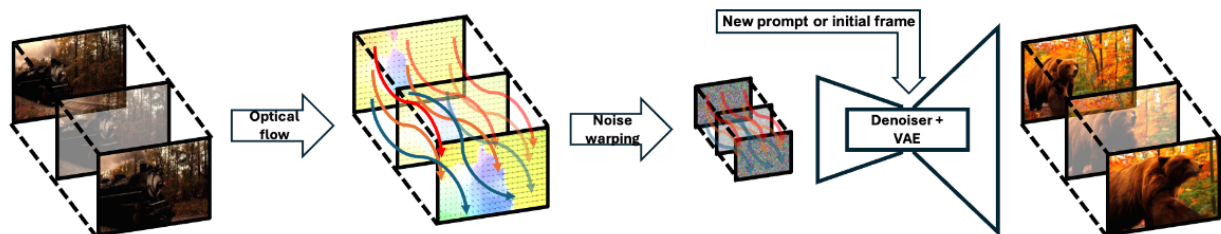
#### 中文讲解

物体运动与相机运动不同。人可以在固定镜头前穿过画室，镜头也可以绕着静止的人移动，两者都会造成像素位置变化。因此，要分别控制它们，就需要有用的约束。规划镜头时，应独立说明相机行为和场景动作，再检查输出是否同时遵守。画面动感十足仍可能不符合要求，例如应该固定的镜头乱动，或本应行动的主体保持静止。

## 21 / Motion as a field

00:12:37 - 00:13:16

### Motion as a field



Optical flow describes how image locations move.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 3

21

Optical flow describes how locations in one image correspond to locations in another. The colored paths in the figure represent a movement field. Such information can help transfer movement or guide the evolution of a generated sequence. Optical flow describes apparent image movement, so it can reflect both object motion and camera motion. It is not a complete reconstruction of the physical world. Occlusion creates additional difficulty because something visible in one frame may disappear behind another object. For creative work, a movement field can offer more specific guidance than a sentence such as move naturally. Its reliability still depends on the source and the scene.

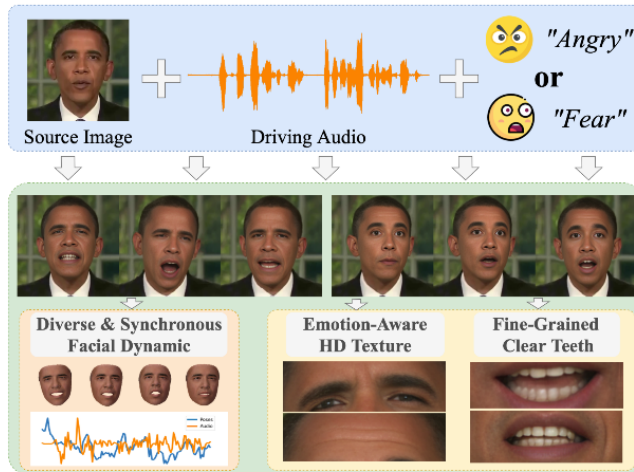
#### 中文讲解

光流描述一张图像中的位置如何对应到另一张图像。图中的彩色路径表示运动场，可用于迁移动作或指导生成序列。光流描述表观图像运动，可能同时来自主体与相机，并不是完整的物理世界重建。遮挡也增加困难，因为某个位置在下一帧可能不可见。相比“自然移动”这样的文字，运动场能提供更具体的引导，但可靠性仍取决于源素材和场景。

## 22 / Sound can guide a performance

00:13:16 - 00:13:53

### Sound can guide a performance



Audio can constrain mouth movement and timing.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 15

22

Sound can guide a visual performance. In this example, a source image and audio help determine a speaking face, with additional information related to expression. Audio supplies timing that a still image cannot provide. A voice has pauses, stressed syllables, and changing rhythm. Mouth movement should relate to those events if the result is meant to depict speaking. A convincing portrait alone does not establish accurate lip synchronization or a coherent performance. Watch transitions and listen at the same time. In a film workflow, this is also a reminder to plan sound early when it controls the timing of the visible action.

### 中文讲解

声音可以指导视觉表演。示例中，源图像和音频共同影响说话的人脸，还加入与表情有关的信息。音频提供静态图像没有的时间约束，例如停顿、重音和节奏变化。若要表现说话，嘴部动作应与这些事件对应。肖像逼真不等于口型同步准确，也不等于表演连贯。应边听边看转折。若声音决定画面动作的时机，制作流程中就应尽早规划声音。

## 23 / Editing a region through time

00:13:53 - 00:14:31

### Editing a region through time



The selected change must persist across frames.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 16

23

Editing a video region adds a temporal requirement to a familiar image-editing task. If we change the color of a car, that change should remain attached to the car across the sequence. The background should also behave consistently. The examples show local changes and reconstruction of selected areas. A mask identifies a region, but the region may move, change shape, or become partly hidden. This makes a single successful frame insufficient evidence. For the studio exercise, imagine changing only the painting while preserving the wall and frame. Inspect the boundary as the camera moves, because that is where small inconsistencies often become noticeable.

#### 中文讲解

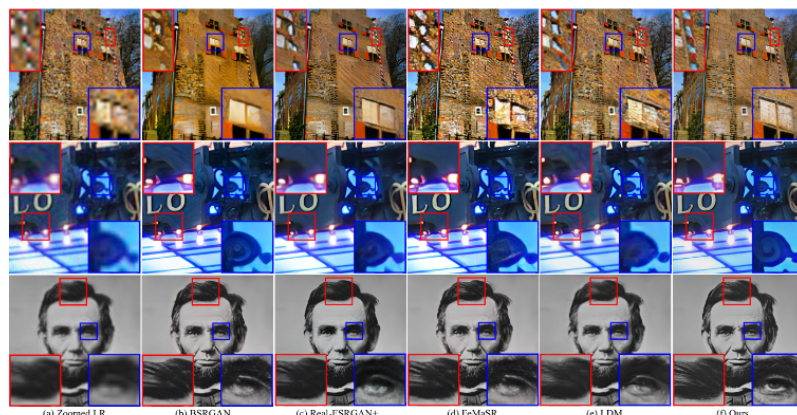
视频局部编辑为图像编辑增加了时间要求。改变汽车颜色后，新颜色应在整个序列中跟随汽车，背景也应保持连贯。遮罩标出区域，但区域可能移动、变形或被遮挡，所以单帧成功不足以证明整段成功。画室中只改画作内部、保留墙面和画框时，要检查镜头移动过程中的边界，因为微小不一致经常首先在那里暴露。



## 24 / Sharper detail can change what we see

00:14:31 - 00:15:10

### Sharper detail can change what we see



Plausible texture may differ from the original scene.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 18

24

Super-resolution methods can make details look sharper. These examples compare restoration methods on still images, even though the figure appears in a video survey. Read the columns as methods, not successive moments in time. Compare the brick texture and the facial details. A generated texture can be plausible without matching the original high-resolution scene. That difference matters if your aim is faithful restoration. It may matter differently if you are designing a new artwork. For video, an additional question remains: do the added details stay stable across frames? These still comparisons alone cannot answer that question. Sharpness and temporal consistency require separate inspection.

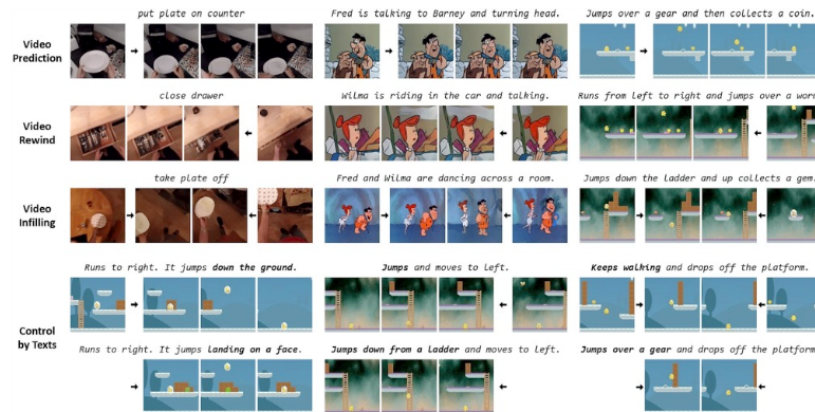
#### 中文讲解

超分辨率可以让细节更清晰。虽然图出现在视频综述中，这里比较的是静态图像恢复方法，列代表不同方法而非连续帧。观察砖墙和面部纹理：生成纹理可能合理，却不同于真实高分辨率场景。忠实修复与创作新作品对此有不同要求。对视频还需追问新增细节是否跨帧稳定。这些静态比较无法回答时间一致性问题；清晰度与稳定性需要分别检查。

# 25 / Before, between, and after

00:15:10 - 00:15:49

## Before, between, and after



Prediction extends a sequence.  
Interpolation fills a gap.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 17

25

Prediction and interpolation answer different questions. Prediction extends a sequence beyond what we already observe. Interpolation fills a gap between known moments. The examples also include generating earlier content, which asks what might have happened before the available sequence. These outputs are plausible constructions, rather than recovered facts about an unseen event. For our studio film, interpolation might help connect two chosen keyframes, while prediction might extend a shot after the painting starts moving. Both methods can produce unexpected intermediate actions. Check the whole path between important moments, because getting the beginning and ending right does not guarantee that the transition serves the scene.

### 中文讲解

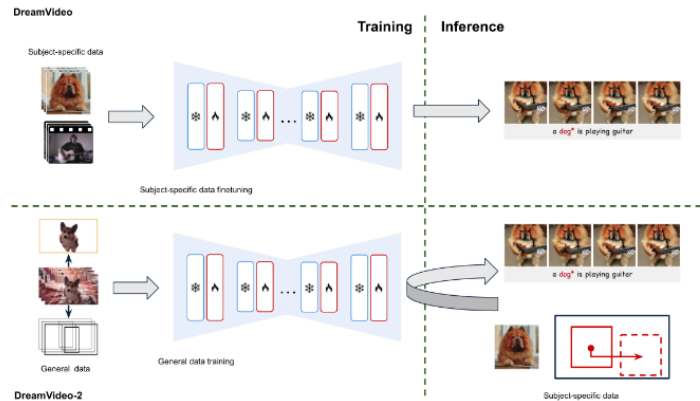
预测延长已知序列，插值填补已知时刻之间的空缺。示例也包含生成更早内容，即推测现有片段之前可能发生什么。这些都是合理构造，而不是恢复未被观察到的事实。画室影片中，插值可连接两个关键帧，预测可延长画作开始运动后的镜头。但中间动作可能出乎意料。开头与结尾正确，不代表整个过渡都符合场景意图，因此必须查看完整路径。



## 26 / A subject that stays recognizable

00:15:49 - 00:16:28

### A subject that stays recognizable



Identity needs to survive changes in pose and view.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 19

26

A recognizable subject needs to survive changes in pose, lighting, and viewpoint. The figure compares approaches that use subject-specific information during adaptation or inference. The technical details differ, but the creative requirement is easy to state. We want to recognize the same subject when the camera angle changes. A general description such as a small brown dog leaves many identifying features unspecified. References can provide those details more directly. For a film, evaluate identity across the shots you actually need, rather than only in a favorable close-up. A method may preserve a face from the front while struggling with a profile or an unusual expression.

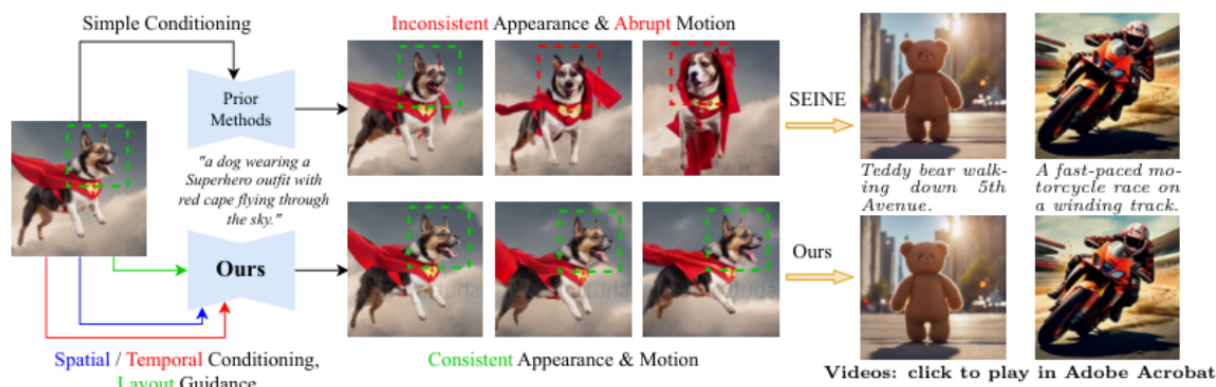
#### 中文讲解

可识别的主体需要经受姿态、光照和视角变化。图中比较在适配或推理时使用主体信息的方法。创作要求很简单：换一个角度，仍认得同一主体。“棕色小狗”只给出类别，参考图能补充具体特征。电影中应检查实际需要的镜头，而不仅是有利的正面特写。某种方法可能保留正脸，却在侧脸或特殊表情下失去身份特征。

## 27 / Continuity needs more than a good first frame

00:16:28 - 00:17:05

### Continuity needs more than a good first frame



Check appearance and movement across the whole shot.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 20

27

A good first frame is only the beginning of continuity. Look at appearance and movement throughout the shot. Does the subject keep its recognizable features? Does its motion develop smoothly and plausibly for the intended style? The figure presents examples intended to illustrate more consistent conditioning. Treat them as selected demonstrations, not proof that every output will remain stable. Also separate a deliberate transformation from an accidental drift. If our painting is meant to change, specify what may change and what must remain stable. The picture inside the frame might transform while the frame itself keeps the same shape and position.

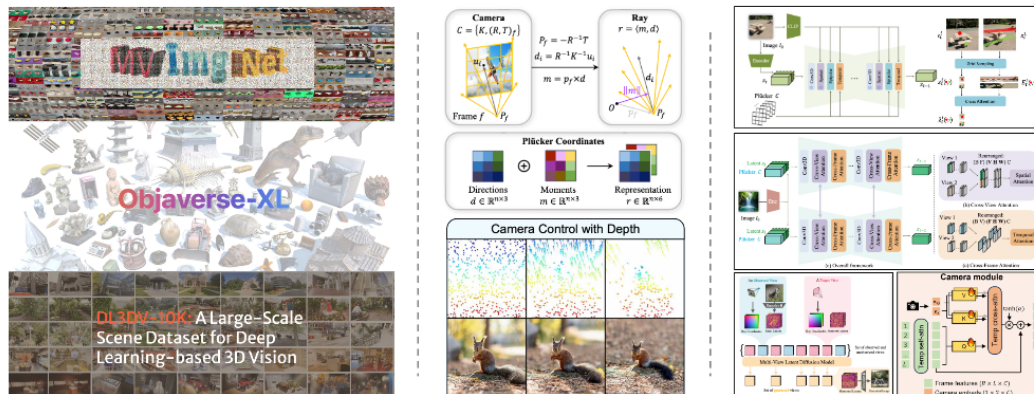
#### 中文讲解

好的首帧只是连续性的起点。要检查整段外观和运动：主体是否可识别，动作是否符合预期风格？图中是条件控制的精选演示，并不能证明每次输出都会稳定。也要区分有意转变与意外漂移。若画作本来就应变化，必须说清楚哪些属性可变、哪些必须稳定。例如画内内容可以变形，但画框的形状和位置不应随意改变。

## 28 / Video models can incorporate 3D information

00:17:05 - 00:17:45

### Video models can incorporate 3D information



Camera information links an image to a viewpoint.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 22

28

Video models can incorporate explicit three-dimensional information. The figure groups examples involving training datasets, camera representations, and architectural changes. The middle column connects a camera to rays through image locations. You can understand a ray as a direction from which the camera observes the scene. This provides more specific spatial information than a verbal instruction alone. The right column shows ways models incorporate such information. We are interested in the principle: viewpoint information can help connect images of the same scene. It still does not guarantee perfect geometry. Unseen surfaces and difficult camera paths remain important places to look for mistakes.

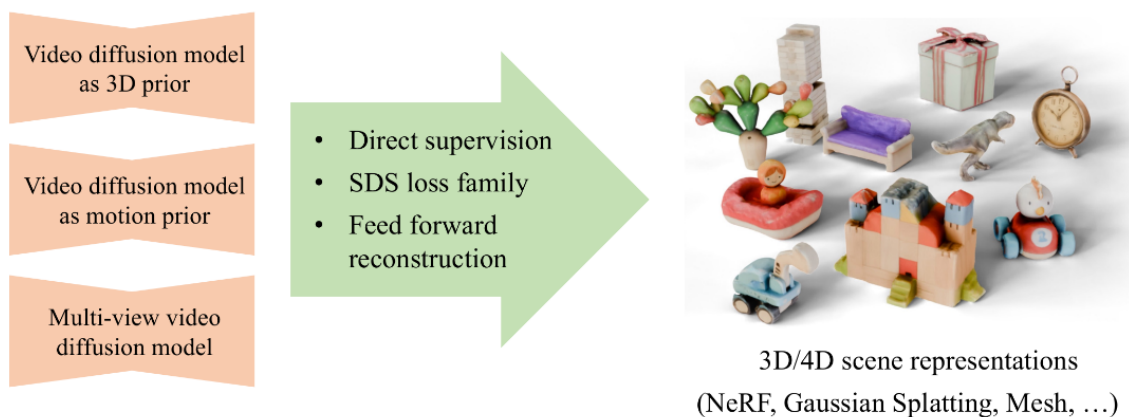
#### 中文讲解

视频模型可以纳入显式三维信息。图中依次组织数据集、相机表示和结构变化。中间部分把相机与穿过图像位置的射线联系起来；射线可理解为观察场景的方向，比一般文字镜头指令更具体。右侧展示利用这些信息的模型方式。核心思想是视角信息有助于连接同一场景的多张图像，但不保证几何完美。未见表面和复杂镜头路径仍需重点检查。

## 29 / Video knowledge can guide a scene representation

00:17:45 - 00:18:24

### Video knowledge can guide a scene representation



A generated sequence can help build a 3D scene.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 24

29

A video model can also provide knowledge for constructing a scene representation. The left side of the diagram identifies sources of visual or motion information. The middle describes ways to transfer that information. The right shows representations such as meshes, neural fields, and Gaussian splats. A video is a set of images over time. A scene representation aims to support rendering under chosen viewpoints, and sometimes at chosen moments. Moving from one to the other requires reconstruction or optimization. This matters to artists because a reusable scene can support later camera decisions, while a finished video usually commits you to the views already present in its frames.

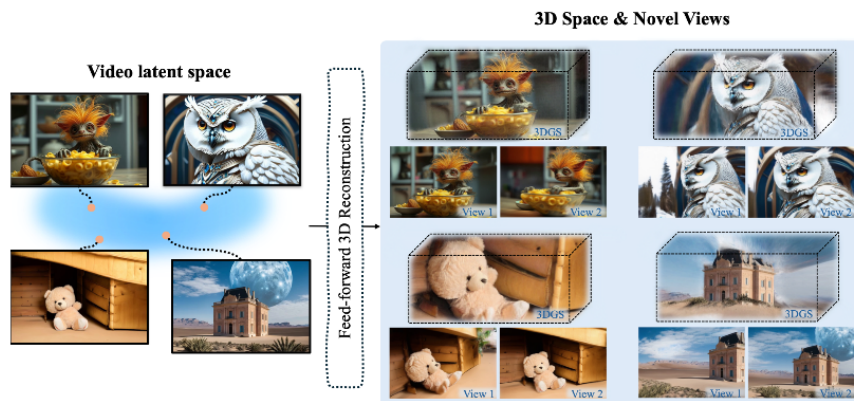
### 中文讲解

视频模型的知识也可以用于建立场景表示。图左是视觉或运动知识来源，中间是迁移方式，右边是网格、神经场和高斯表示。视频是一组随时间排列的图像，场景表示则希望支持指定视角，有时还包括指定时刻的渲染。两者之间需要重建或优化。这对电影制作很重要：可复用场景允许稍后调整镜头，而完成的视频通常已固定在原有帧的视角中。

## 30 / Wonderland connects video and reconstruction

00:18:24 - 00:19:03

### Wonderland connects video and reconstruction



A reconstruction model predicts a scene with new views.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 25

30

Wonderland illustrates a connection between video representations and scene reconstruction. Follow the diagram from the latent space on the left to the three-dimensional outputs on the right. The small paired views help show the purpose of reconstruction: we can render related views of a scene rather than simply display one image. The examples use Gaussian splats as a scene representation. We will revisit that idea later. For now, ask what additional freedom this creates for a filmmaker. You may gain some camera flexibility, but selected views do not demonstrate unrestricted movement or complete hidden geometry. Test the camera path your project actually requires.

#### 中文讲解

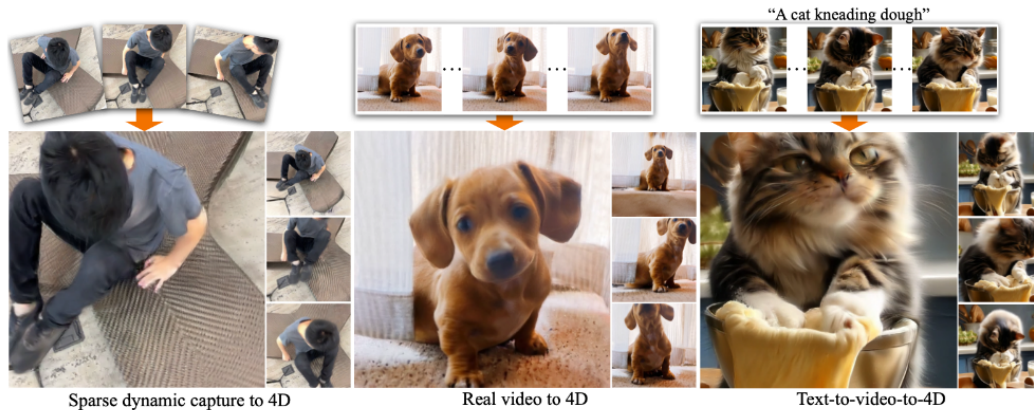
Wonderland连接视频表示与场景重建。沿着图从左侧潜在空间看向右侧三维结果，成对视图展示重建的目的：渲染相关的新视角，而非只显示单张图像。这里使用高斯泼溅表示场景。这可能为电影制作带来一定镜头灵活性，但精选视图不证明可以任意移动，也不证明隐藏几何完整。必须测试项目真正需要的镜头路径，而不是仅凭演示推断无限自由。



# 31 / A moving scene varies in viewpoint and time

00:19:03 - 00:19:42

## A moving scene varies in viewpoint and time



CAT4D models both dimensions.

Wang et al., Video Diffusion Survey, rev. 2026, Fig. 26

31

A dynamic scene changes in two ways: the viewpoint can change, and the action can advance through time. CAT4D addresses this combination. The examples include captured and generated source material connected to dynamic scene reconstruction. Imagine freezing a dancer at one moment and walking around them. Now imagine keeping the camera still while the dancer continues. These are different operations. A model that represents both needs to infer appearances across viewpoint and time. Much of that material may never have been observed directly. For creative work, this offers flexibility, but it also means that the quality of inferred regions and motions needs deliberate evaluation.

### 中文讲解

动态场景同时在视角和时间上变化。CAT4D研究这种组合，示例连接捕获或生成的素材与动态场景重建。想象先冻结舞者的某一瞬间并绕其观察，再固定相机让舞者继续动作，这是两个不同操作。兼顾两者的模型需要推断跨视角、跨时间的外观，许多内容从未直接被观察。它提供灵活性，也要求创作者认真检查推断出来的区域和动作。

## 32 / Video-model checkpoint

00:19:42 - 00:20:20

### Video-model checkpoint

What fixes appearance?

What controls movement?

What still needs inspection?

Wang et al., Video Diffusion Survey, rev. 2026

32

Let us pause and apply the video survey to our studio. What fixes the appearance of the room? An initial image or a scene representation might help. What controls the movement? A motion reference, audio signal, camera instruction, or other condition might supply useful information. What still needs inspection? The painting boundary, the person's identity, and the continuity of the room are possible answers. Choose one requirement and explain why a particular input helps. Then identify a failure that the input cannot rule out. This exercise is more useful than naming a fashionable model, because it connects the method to evidence you can actually inspect.

#### 中文讲解

把视频综述应用到画室：什么固定外观？初始图像或场景表示可能有帮助。什么控制运动？动作参考、声音、相机信息等可以提供约束。什么仍需检查？画作边界、人物身份和房间连续性都是答案。选择一个要求，解释某种输入为什么有用，再指出它无法排除的一种失败。这个练习把方法与可观察证据联系起来，比单纯说出一个热门模型名称更有价值。



# 02

## Visual art creation

Wang, Chen, and Wang  
Diffusion-Based Visual Art Creation  
ACM Computing Surveys, 2025

Wang, Chen & Wang, Visual Art Creation, 2025

33

Our next reading is Diffusion-Based Visual Art Creation by Wang, Chen, and Wang, published in ACM Computing Surveys in 2025. This paper shifts the emphasis from generating sequences to understanding artistic tasks and intentions. We will consider how technical choices relate to the kind of artwork someone wants to make. A system may generate a polished image while giving the artist too little control over its meaning or structure. Conversely, an imperfect result may become useful material for further work. Keep the studio example in mind. We now ask what the changing painting should express, and how that intention changes the way we judge its output.

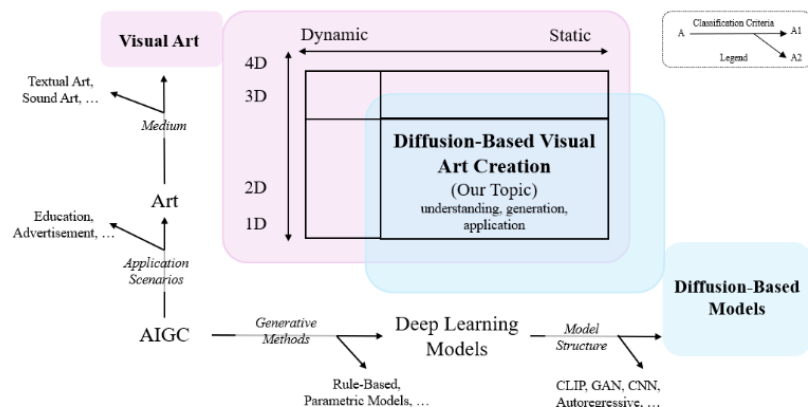
### 中文讲解

第二篇是Wang、Chen和Wang的《基于扩散的视觉艺术创作》，发表于2025年的ACM Computing Surveys。重点从生成序列转向艺术任务和创作意图。系统可以生成精致图像，却未必给予艺术家足够的意义或结构控制；不完美输出也可能成为进一步创作的材料。继续思考画室：运动的画究竟要表达什么？这个意图如何改变我们判断结果的方式？

## 34 / Artistic goals and technical methods

00:20:58 - 00:21:35

### Artistic goals and technical methods



The survey studies where these two perspectives meet.

Wang, Chen & Wang, Visual Art Creation, 2025, Fig. 1

34

The figure places diffusion-based visual art at the intersection of several categories. Art can use different media, serve different purposes, and exist as static or changing material. Technical methods also vary in what they represent and how they generate it. The survey studies where these perspectives meet. This is useful because a technical task label rarely describes the whole artistic problem. Text-to-image tells us something about inputs and outputs, but little about why the image matters. For our studio, an illustration, an installation, and a film might use related generated imagery while requiring different forms of control and different ways of evaluating the audience experience.

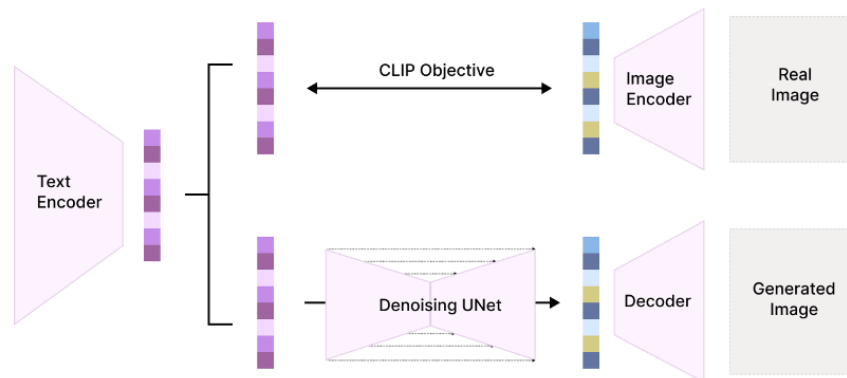
#### 中文讲解

图把扩散视觉艺术放在不同分类的交汇处。艺术可以使用不同媒介，服务不同目的，并呈现静态或动态材料；技术方法也有不同表示和生成方式。“文生图”说明输入输出，却没有解释图像为何重要。画室图像可以用于插画、装置或电影，各自需要不同控制，也需要不同的观众体验评价。技术任务名称不能代替对完整艺术问题的描述。

## 35 / Text becomes a condition for generation

00:21:35 - 00:22:16

### Text becomes a condition for generation



A text encoder turns words into a model representation.

Wang, Chen & Wang, Visual Art Creation, 2025, Fig. 2

35

Text becomes usable conditioning through an encoder. The diagram shows a conceptual connection between text representations, image representations, and a denoising network. CLIP is associated with learning relationships between image and text representations. In a diffusion workflow, text information can then influence the refinement of visual content. This simplified figure combines ideas and should not be treated as the exact architecture of every generator. The important point is that a sentence becomes a numerical representation, rather than a complete set of drawing instructions. A phrase such as *uneasy memory* therefore leaves many visible decisions open. References and editing can help make those decisions more specific.

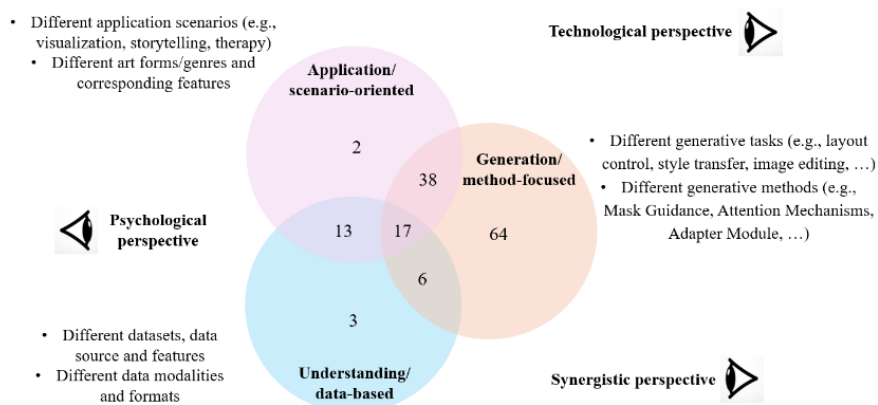
#### 中文讲解

文字通过编码器变成可用条件。图示概念性地连接文字表示、图像表示和去噪网络。CLIP学习图文表示之间的关系，文字信息可以进一步影响扩散生成。这里是简化示意，不是所有生成器的精确结构。句子被转成数值表示，并非一套完整绘画指令，因此“不安的回忆”仍留下许多视觉选择。参考图和编辑能帮助把这些选择具体化。

## 36 / Three research perspectives

00:22:16 - 00:22:54

### Three research perspectives



Creating images is one part of the picture.

Wang, Chen & Wang, Visual Art Creation, 2025, Fig. 3

36

The survey identifies three research perspectives: applications, generation methods, and understanding or data. The overlapping circles show that a paper may belong to more than one perspective. Work on an artistic application can also introduce a generation method or study how people interpret images. The numbers describe the authors' selected research corpus, not all possible work in the field. When reading a paper, identify which question it primarily answers. Does it introduce a tool, study an audience, or analyze material? This helps you avoid expecting a technical benchmark to answer an artistic question that the researchers did not set out to investigate.

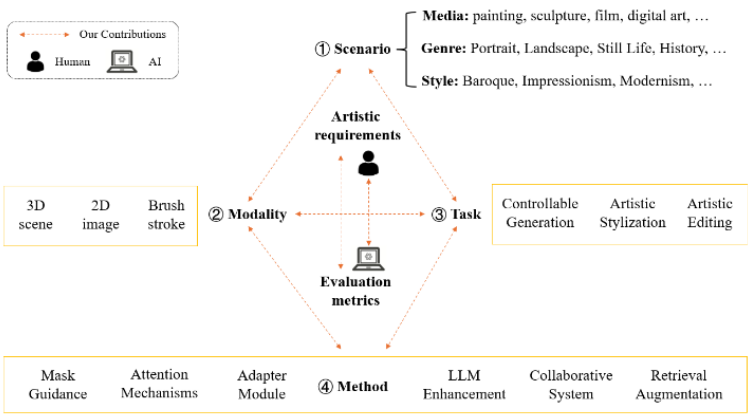
#### 中文讲解

综述区分应用、生成方法、理解与数据三个研究视角。重叠区域表示论文可能同时属于多个类别，数字仅描述作者选择的文献集合。阅读时应判断论文主要回答什么：介绍工具、研究观众，还是分析素材？这样可以避免要求技术基准回答研究者没有打算研究的艺术问题。不同证据支持不同结论，不能仅凭一个领域的结果跨越到另一个问题。

# 37 / The artist's intention comes first

00:22:54 - 00:23:31

## The artist's intention comes first



A feeling becomes a set of visible decisions.

Wang, Chen & Wang, Visual Art Creation, 2025, Fig. 4

37

An artistic intention becomes useful when we connect it to visible decisions. The survey's framework links scenario, modality, task, and method to artistic requirements and evaluation. Suppose our intention is an uneasy memory. We could use delayed movement, missing details, or a familiar room with one altered object. Each choice suggests different controls. A general request to make the scene more artistic does not tell us what to change or how to judge success. Start with the intended experience, identify a visible decision, and choose a method that gives you some control over it. The framework is a way to organize that reasoning.

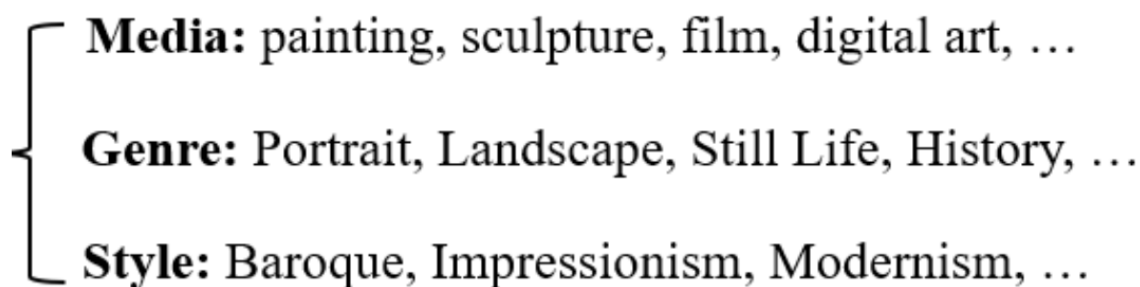
### 中文讲解

艺术意图需要连接到可见决定。框架把情境、模态、任务和方法与艺术要求及评价联系起来。若目标是不安的回忆，可以选择延迟运动、缺失细节或熟悉房间中的异常物体。每种选择需要不同控制。“更有艺术感”并未说明要改什么，也无法判断是否成功。先确定体验，再确定可见介入，最后选择能控制它的方法。这个框架帮助组织推理，而不是替你决定作品意义。

## 38 / A scenario gives the work a purpose

00:23:31 - 00:24:08

### A scenario gives the work a purpose



A portrait, an installation, and a film ask different things.

Wang, Chen & Wang, Visual Art Creation, 2025, Fig. 4 (detail)

38

A scenario gives the work a purpose. The enlarged figure distinguishes medium, genre, and style. A portrait concentrates attention differently from a landscape. A film unfolds over time, while an installation may respond to where a visitor stands or how long they remain. These distinctions affect both production and evaluation. The same generated image might be a finished print, a storyboard reference, or one frame in an animation. For the studio example, decide whether the audience watches a fixed sequence or discovers changes while moving through a space. That choice changes the kinds of continuity and control the work needs.

### 中文讲解

情境赋予作品目的。放大的图区分媒介、题材与风格。肖像和风景引导注意力的方式不同，电影随时间展开，装置则可能取决于观众位置和停留时间。同一张生成图可以是成品版画、分镜参考或动画的一帧。画室作品究竟让观众观看固定序列，还是在空间中自行发现变化？这一选择会改变连续性和控制方面的要求。

## 39 / A modality is the form of the material

00:24:08 - 00:24:45

### A modality is the form of the material



Pixels, brush strokes, and 3D scenes support different edits.

Wang, Chen & Wang, Visual Art Creation, 2025, Fig. 4 (detail)

39

A modality is the form of the material you work with. The figure names a three-dimensional scene, a two-dimensional image, and a brush stroke. These forms support different kinds of intervention. Editing pixels gives direct access to an image surface. Working with a scene may give access to camera position and spatial relationships. A stroke representation may preserve information about how a mark is constructed. None is universally best. Ask what you want to revise later. If you need to move the camera around the studio, a single finished image may be restrictive. If you need a fixed composition, its simplicity may be helpful.

#### 中文讲解

模态是材料的形式，例如三维场景、二维图像和笔触。像素编辑直接作用于图像表面；场景表示可能允许调整相机和空间关系；笔触表示可能保留痕迹的构造信息。没有一种形式普遍最好。先问以后要改什么。如果需要绕画室移动相机，单张完成图像可能限制很大；若只需要固定构图，简单的二维图像可能更适合。



# 40 / Four useful creative tasks

00:24:45 - 00:25:22

## Four useful creative tasks

Generate content

Control a result

Change a style

Edit selected content

Wang, Chen & Wang, Visual Art Creation, 2025

40

Four useful creative tasks are generating content, controlling a result, changing style, and editing selected content. These tasks often occur together, but distinguishing them helps diagnose a workflow. If the room is missing, you need content generation. If the room exists but the door is in the wrong place, you need control or editing. If the composition is suitable but the material should resemble charcoal, you may need stylization. Describe the problem before choosing the tool. Repeating full generation for every small correction can discard successful decisions that you would rather preserve. A local intervention may be a better fit for the actual need.

### 中文讲解

四类创作任务是生成内容、控制结果、改变风格和编辑局部。它们经常组合，但区分后更容易诊断流程。房间还不存在，需要生成；房间存在但门的位置错了，需要控制或编辑；构图合适但想改成炭笔质感，可能需要风格化。先描述问题再选工具。每次小修正都重新生成全图，可能丢掉已经成功的决定。局部介入往往更贴合实际需要。

## 41 / An edit needs a boundary

00:25:22 - 00:25:58

### An edit needs a boundary

A mask specifies where a change may occur.

The surrounding image provides context.

Wang, Chen & Wang, Visual Art Creation, 2025

41

A mask specifies a region where an edit may occur. The surrounding image provides context for making the change fit. In our studio, we might mask the inside of the painting while leaving its frame and the wall outside the selected region. This makes the request more precise than asking for a different painting in an entirely new room. However, a mask is not a guarantee that every unselected detail will remain identical. The behavior depends on the method and settings. Inspect the edge of the edit, the lighting, and nearby textures. A successful local change should also make sense within the larger composition.

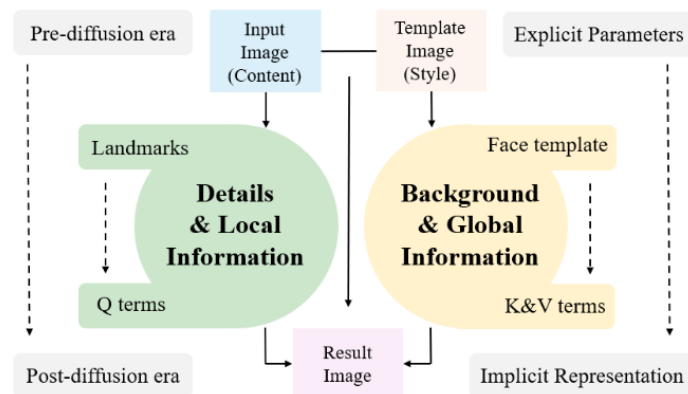
#### 中文讲解

遮罩指出允许编辑的区域，周边图像提供上下文。画室中可选中画作内部，保留画框和墙壁，比重新生成整间房更精确。但遮罩不保证未选区域完全不变，具体行为取决于方法和设置。应检查边缘、光照和邻近纹理。局部变化成功，还需要与整体构图相容，不能只看被修改的中心部分。

## 42 / Style includes structure

00:25:58 - 00:26:35

### Style includes structure



Shape and composition matter alongside color.

Wang, Chen & Wang, Visual Art Creation, 2025, Fig. 8

42

Style includes more than color or surface texture. The diagram connects local details with global information and shows how different parts of a method can influence a result. Think about the difference between a face painted with loose brushwork and a face whose proportions have also changed. Both may appear stylistic, but they alter different relationships. Composition, shape, and rhythm can be part of the visual language. When you request a style transfer, identify which of those properties may change. Otherwise, a method might produce the intended texture while removing a structural feature that was important to the identity or meaning of the original.

#### 中文讲解

风格不只有颜色或表面纹理。图把局部细节与整体信息联系起来。松散笔触的人脸与同时改变比例的人脸，都可能呈现风格变化，但改变的是不同关系。构图、形状和节奏也属于视觉语言。进行风格迁移时，应明确哪些属性允许变化，否则可能得到所需纹理，却丢失对身份或意义至关重要的结构特征。

## 43 / Content and style need separate decisions

00:26:35 - 00:27:14

### Content and style need separate decisions

Keep the subject recognizable.

Choose which aspects of appearance may change.

Wang, Chen & Wang, Visual Art Creation, 2025

43

Let us separate content and style as a practical exercise. Imagine that a character must remain recognizable while the image changes from photographic to painted. Which features should stay stable? You might choose the silhouette, facial proportions, or pose. Which features may vary? Brushwork and palette are possible answers. This separation is useful, but it is not universal. Some styles depend on changing proportions or spatial relationships. Explain your own boundary rather than assuming the software will infer it. For the studio painting, decide whether its subject remains recognizable during the transformation, or whether losing recognition is part of the intended experience.

#### 中文讲解

做一个区分内容与风格的练习：角色从摄影外观变为绘画外观时，哪些特征必须保留？可以是轮廓、脸部比例或姿态。哪些允许变化？可能是笔触和配色。这种区分有用，但不是普遍规律，因为某些风格本身就依赖比例变化。请明确自己的边界，不要假设软件会猜到。画室中的画在变形时，是应保持主题可辨认，还是失去辨认本身就是体验的一部分？

## 44 / A technical score answers a narrow question

00:27:14 - 00:27:53

### A technical score answers a narrow question

Matching a prompt does not establish artistic success.

Wang, Chen & Wang, Visual Art Creation, 2025

44

A technical score answers a particular question. It may estimate prompt alignment, visual similarity, or some aspect of perceptual quality. It cannot automatically establish whether an artwork succeeds. A deliberately ambiguous image might receive a weaker score for literal prompt matching while producing the intended experience. That does not make technical evaluation useless. It means we need to state what each measure is supposed to assess. For the studio, you might separately evaluate whether the room remains stable and whether the change in the painting feels unsettling. One concerns a visible constraint. The other concerns interpretation, which may require viewers and discussion rather than a single automated number.

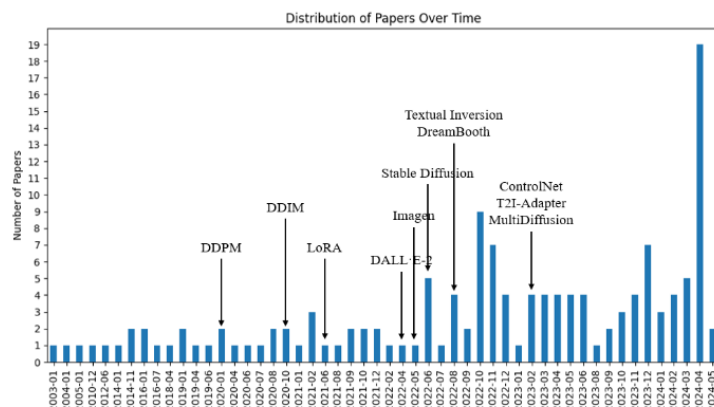
#### 中文讲解

技术分数回答特定问题，例如提示词匹配、视觉相似度或感知质量，不能自动决定艺术作品是否成功。有意模糊的图像可能字面匹配较低，却产生预期体验。这不意味着技术评价无用，而是要说明每项指标测什么。画室中可以分别评价房间是否稳定，以及画作变化是否令人不安。前者是可见约束，后者涉及理解，可能需要观众反馈与讨论，而不是单一自动分数。

# 45 / The literature changed after diffusion

00:27:53 - 00:28:30

## The literature changed after diffusion



The survey records growth in its selected research corpus.

Wang, Chen & Wang, Visual Art Creation, 2025, Fig. 5

45

This chart records changes in the survey's selected literature over time. The labels connect the distribution to influential developments such as diffusion models and methods for adaptation or control. Read it as a historical view of the corpus the authors assembled. It is not a complete census of AI-art research, and its final dates do not extend to today. The useful question is how new methods changed the kinds of tasks researchers could attempt. Greater access to generation can shift attention toward control, editing, and applications. To understand the strength of that claim, we would still need to examine the actual papers and selection process.

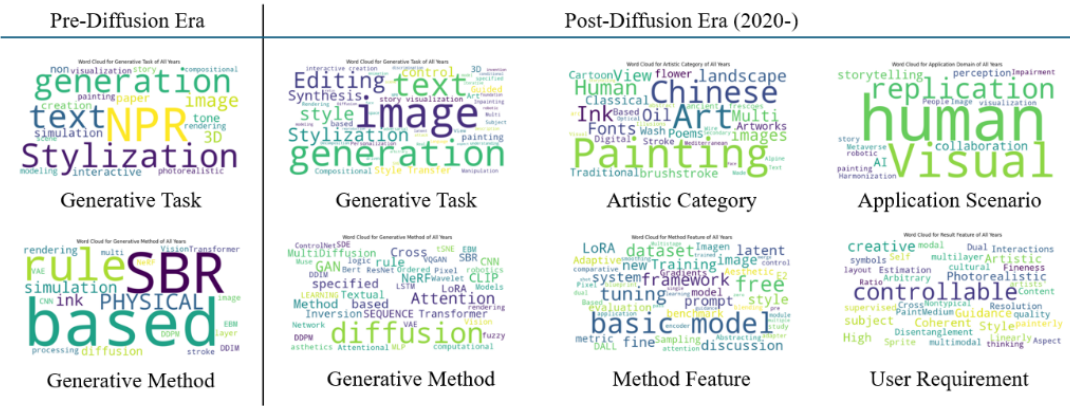
### 中文讲解

图表记录作者所选文献随时间的变化，并标出扩散、适配和控制方法等发展。它是历史文献集合的视图，不是AI艺术研究的完整普查，末端日期也不延伸到今天。值得问的是新方法如何改变研究者能够尝试的任务。生成变得容易后，注意力可能转向控制、编辑和应用。但要判断这种解释有多强，仍需阅读具体论文并了解文献筛选方法。

# 46 / The vocabulary of art research also changed

00:28:30 - 00:29:08

## The vocabulary of art research also changed



The survey compares tasks, methods, and artistic concerns.

Wang, Chen & Wang, Visual Art Creation, 2025, Fig. 7

46

The word clouds offer another view of the authors' research collection. Larger words indicate prominence within the coded material, rather than artistic importance or current popularity. Compare task terms such as generation and editing with method terms such as diffusion. Then look at the artistic categories and user requirements. These are different kinds of vocabulary, and mixing them can make a project description vague. If you say your project uses diffusion, you have named a method. If you say it explores fragmented memory through changing portraits, you have begun to describe an artistic purpose. A clear proposal explains the connection between those two descriptions.

### 中文讲解

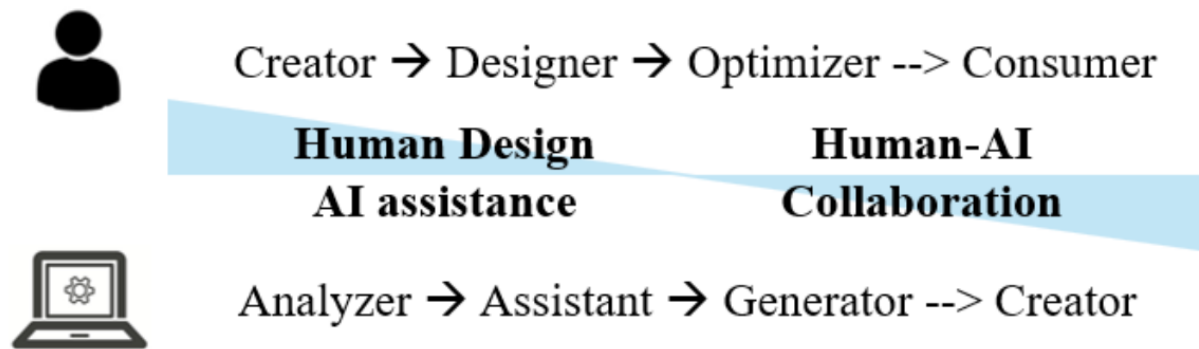
词云提供文献集合的另一种视图。词越大，表示在编码材料中越突出，不等于艺术重要性或当前流行程度。比较“生成”“编辑”等任务词与“扩散”等方法词，再看艺术类别和用户要求。混用不同层次的词，会让项目描述含糊。“使用扩散”说的是方法；“通过变化的肖像探索碎片记忆”才开始说明艺术目的。清楚的提案需要解释两者如何相连。



## 47 / Human involvement changes across a workflow

00:29:08 - 00:29:44

### Human involvement changes across a workflow



Choosing, revising, and judging remain creative acts.

Wang, Chen & Wang, Visual Art Creation, 2025, Fig. 9

47

Human involvement changes across a workflow. The figure presents possible roles for people and AI, from assistance and analysis toward more extensive generation. Treat this as a conceptual perspective, not an inevitable sequence in which one role replaces another. Choosing references, rejecting outputs, setting constraints, and deciding when to stop are all meaningful parts of creative work. Their importance may increase when generating more candidates becomes easy. For your own project, identify a decision that you would keep under direct human control. Then explain what evidence you would want from a tool before accepting its contribution to that decision.

#### 中文讲解

人在流程中的作用会变化。图展示从辅助、分析到更多生成的可能角色，它是一种概念视角，不是必然替代关系。选择参考、拒绝输出、设置约束和决定何时完成，都是有意义的创作工作。生成候选越容易，这些决定可能越重要。请指出项目中一个希望由人直接掌握的决定，并说明在接受工具贡献之前，希望看到什么证据。

## 48 / Our studio scene as an artwork

00:29:44 - 00:30:20

### Our studio scene as an artwork

Intention: an uneasy memory.

Decision: the painting changes before the room does.

Classroom exercise · inspired by the readings

48

Return to the studio as an artwork. Our intention is an uneasy memory, and our chosen intervention is that the painting changes before the room does. The order directs attention. At first, the viewer may wonder whether they noticed anything. Later, a second change can confirm that something is wrong. A different version might use abrupt movement and loud sound to create surprise instead. Both could use similar generation tools while producing different experiences. Describe the intended effect in terms that can guide production. We need to know where the change begins, how quickly it develops, and which stable details help the audience notice it.

#### 中文讲解

回到作为艺术作品的画室：目标是不安的回忆，介入是画作先变化、房间后变化。顺序引导注意力，观众最初可能怀疑自己是否看错，后续变化再确认异常。另一版本可用突然运动和巨响制造惊吓。工具相近，体验仍可能完全不同。需要以能指导制作的方式描述效果：变化从哪里开始，发展多快，哪些稳定细节帮助观众注意到它。

## 49 / One intention, one intervention

00:30:20 - 00:30:57

### One intention, one intervention

Choose a visual decision.

Explain how an audience might read it.

Classroom exercise · inspired by the readings

49

Choose one artistic intention and one visible intervention. For example, you might want the studio to feel welcoming, and decide that light gradually enters through the window. Or you might want it to feel unfamiliar, and change the scale of an object between shots. Explain how an audience could read the intervention. Then propose a way to judge whether it works, perhaps by showing two versions to classmates and asking what changed their interpretation. Pause the recording if you want time to work through an example. The aim is to connect a controllable property to an intended experience, rather than simply to request a more impressive output.

#### 中文讲解

选择一个艺术意图和一个可见介入。若想让画室温暖，可以让光逐渐进入；若想让它陌生，可以在镜头间改变物体尺度。解释观众可能如何理解，再提出检验办法，例如给同学看两个版本，询问什么改变了他们的感受。需要讨论时可暂停录音。练习目标是把可控制属性与预期体验联系起来，而不是笼统要求更惊艳的输出。

## 50 / Art becomes a sequence

00:30:57 - 00:31:34

### Art becomes a sequence

The next question is how the choices survive across shots.

Wang, Chen & Wang, Visual Art Creation, 2025

50

Once art becomes a sequence, decisions have to survive across shots and moments. A visual style may need to remain recognizable. A character may need to carry identity from a close-up into a wider view. The sequence also creates relationships through ordering, duration, and sound. Some art videos deliberately disrupt continuity, so consistency is not an absolute artistic rule. What matters is whether a change supports the work. We now move from the visual-art framework to filmmaking cases. As you look at each case, identify the intended effect, the tools that support it, and the human choices that connect the outputs into a finished experience.

#### 中文讲解

艺术成为序列后，选择需要跨镜头和时刻维持。风格要可辨认，角色要从特写延续到远景，顺序、时长和声音也产生关系。艺术视频可能有意破坏连续性，所以一致性不是绝对规则，关键是变化是否服务作品。接下来进入电影案例。每个案例都可以问：预期效果是什么，哪些工具支持它，人又如何把输出组织为完整体验？

# 03

## Film creation

Zhang et al.  
Generative AI for Film Creation  
CVPR CVEU Workshop, 2025

Zhang et al., Film Creation, CVPR Workshop 2025

51

The filmmaking survey by Zhang and colleagues appeared in the 2025 CVPR workshop on Computer Vision for the Creative Industries. It combines observations from an AI film hackathon with case studies and feedback from artists. This gives us a useful view of how people combine methods in practice. It also sets limits on the evidence. Hackathon participants are a selected group, and the tools reflect the period studied. We should not treat their behavior as a complete picture of the film industry. Focus on what the cases reveal about workflow: where generation helps, where creators intervene, and why a film still needs decisions beyond individual shots.

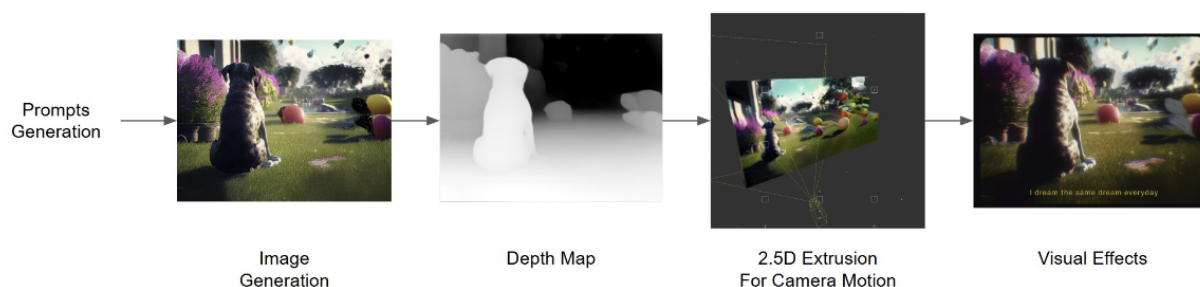
### 中文讲解

Zhang等人的电影制作综述发表于2025年CVPR创意产业计算机视觉研讨会，结合AI电影黑客松、案例和艺术家反馈。它展示创作者如何组合技术，也有证据边界：参赛者经过自我选择，使用工具反映研究时期，不能代表整个产业。阅读时关注生成在哪里提供帮助、人在哪里介入，以及为什么单个镜头之外仍需要制作决策。

## 52 / A film combines several kinds of work

00:32:12 - 00:32:49

### A film combines several kinds of work



Images become part of a production workflow.

Zhang et al., Film Creation, CVPR Workshop 2025, Fig. 7

52

This workflow comes from DOG: Dream of Galaxy. Follow the sequence from an image toward a depth map, a shallow spatial construction, and visual effects. The generated image is one contribution to the production process. Later stages create movement and shape the final presentation. This is a useful correction to the idea that making an AI film means asking one model for a finished movie. A creator may combine generation with familiar tools for composition, camera work, and editing. When evaluating a workflow, identify the contribution of each stage. Otherwise, you may attribute an effect to the generator that actually comes from later production work.

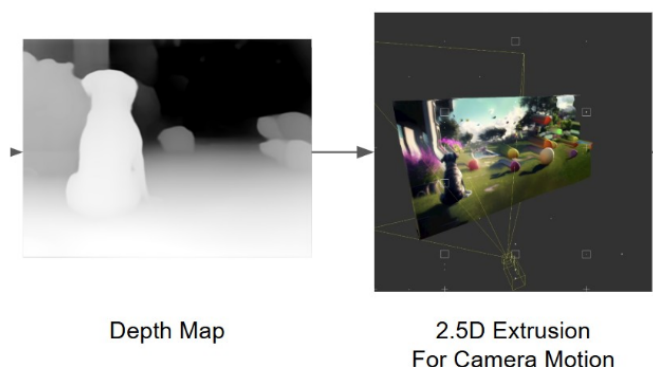
### 中文讲解

《DOG: Dream of Galaxy》的流程从图像走向深度图、浅层空间结构和视觉特效。生成图像只是其中一步，后续制作创造运动并塑造最终呈现。因此，AI电影不等于向一个模型索要完整影片。分析流程时，要辨认每个阶段的贡献，避免把后期合成或运镜产生的效果误归给生成器。

## 53 / Depth can turn an image into a camera move

00:32:49 - 00:33:26

### Depth can turn an image into a camera move



Near and far regions move differently as the camera shifts.

Zhang et al., Film Creation, CVPR Workshop 2025, Fig. 7 (detail)

53

Depth helps distinguish near and far regions in an image. If those regions are lifted into a shallow spatial arrangement, moving a virtual camera can create different amounts of apparent motion across the picture. This is one reason depth can make a still image feel more spatial. However, the original image does not show what is behind each object. A large camera movement may expose gaps or stretched regions. The figure's two-and-a-half-dimensional construction is therefore useful within a range of views. For the studio scene, a small camera drift might work well, while a complete orbit around furniture would demand much more information.

#### 中文讲解

深度区分图像中的远近区域。把这些区域放入浅层空间后，虚拟相机移动可产生不同程度的视差，让静态图像更有空间感。但原图没有物体背面的信息，大幅运镜可能暴露空洞或拉伸。图中的2.5D结构适合有限视角。画室中的轻微相机漂移可能有效，绕家具一整圈则需要更多信息。

## 54 / Film Hack evidence has a context

00:33:26 - 00:34:09

### Film Hack evidence has a context

2023: 8 films

2024: 67 films

2025: 118 films

Zhang et al., Film Creation, CVPR Workshop 2025

54

The survey's Film Hack samples contain eight films from 2023, sixty-seven from 2024, and one hundred eighteen from 2025. These numbers provide context for the observations that follow. The sample size changed, and the participants chose an event devoted to AI filmmaking. Adoption rates in this group therefore cannot directly estimate adoption across the whole industry. The earliest group is especially small. A useful reading habit is to ask who supplied the evidence and how they were selected. This does not dismiss the survey. It helps us use it appropriately, as a detailed account of a particular creative setting rather than a universal industry census.

#### 中文讲解

Film Hack样本包括2023年的8部、2024年的67部和2025年的118部影片。样本量发生变化，参与者又主动选择AI电影活动，所以采用比例不能直接推算整个电影产业，尤其最早一组很小。询问证据来自谁、如何选取，并不是否定研究，而是把它用作特定创作环境的详细记录，而非普遍产业普查。



## 55 / Creators combined tools

00:34:09 - 00:34:46

### Creators combined tools

The 2025 sample averaged 3.14 video-generation tools per film.

Zhang et al., Film Creation, CVPR Workshop 2025

55

The 2025 sample used an average of 3.14 video-generation tools per film. This describes the sample, rather than recommending that every film use three tools. Creators may combine tools because different stages have different requirements. One method may produce a useful starting look, another may preserve a performance, and an editing application may assemble the results. Additional tools also introduce coordination work and opportunities for mismatch. For your project, add a tool when you can identify the problem it solves. A workflow with fewer tools can be effective if it provides the controls you need and lets you preserve decisions that already work.

#### 中文讲解

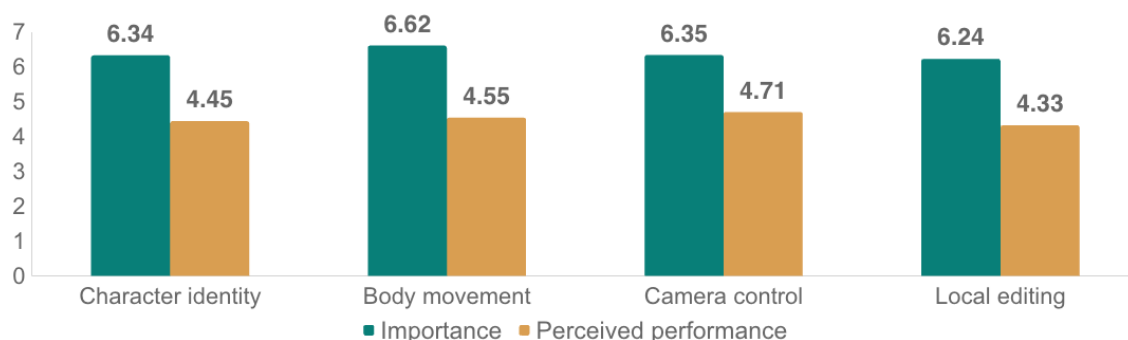
2025年样本平均每部影片使用3.14种视频生成工具。这是描述，不是建议每部电影都用三种。不同工具可能分别负责外观、表演保留和剪辑，但也带来协调与匹配成本。为项目增加工具时，应说清它解决什么问题。少量工具若能提供必要控制、保存已有效的决定，同样可以形成好流程。

## 56 / What filmmakers wanted

00:34:46 - 00:35:30

### What filmmakers wanted

Importance exceeded reported performance in each selected task.



Artist ratings on a 0–7 scale. n = 100. Selected tasks from Table 5.

Zhang et al., Film Creation, CVPR Workshop 2025

56

The chart compares importance with perceived performance for four selected tasks. The green bars represent how important artists considered the task. The gold bars represent their assessment of current performance during the survey period. Character identity, body movement, camera control, and local editing all show a gap. These are ratings from one hundred respondents on a zero-to-seven scale, not objective model benchmark scores. The result suggests that attractive generation alone did not satisfy all production needs. Notice how the tasks concern control over specific properties. For an artist, being able to repeat or revise a decision can matter as much as obtaining an impressive first output.

#### 中文讲解

图表比较四项任务的重要性与当时的感知表现：角色身份、身体运动、相机控制、局部编辑。绿色表示重要性，金色表示表现评价，均来自100位受访者的0到7分评分，不是客观模型基准。差距说明，漂亮的首次输出不足以满足制作需求。能够重复和修改具体决定，对艺术家同样重要。

## 57 / A visual style can begin with a drawing

00:35:30 - 00:36:06

### A visual style can begin with a drawing



#### A Dream About to Awaken

Zhang et al., Film Creation, CVPR Workshop 2025, Fig. 1

57

A Dream About to Awaken illustrates a workflow in which drawings help establish the visual language. The survey describes interpreting hand-drawn storyboards and remixing color and style. Look at the strong shapes and the repeated use of bright color in the examples. The drawings give the process a more specific starting point than a broad style label alone. They also preserve a place for the creator's composition decisions. When using a generator, you can begin with material you made yourself and ask it to develop selected aspects. The important question is which qualities of that starting material should survive through interpretation and animation.

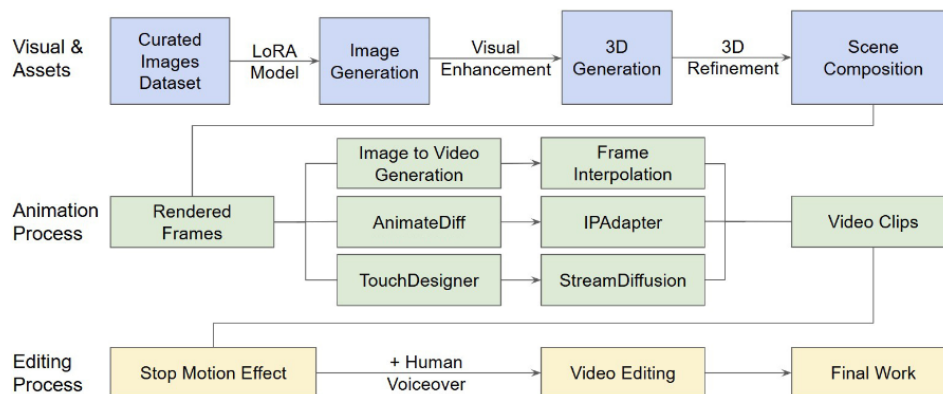
#### 中文讲解

《A Dream About to Awaken》用手绘故事板建立视觉语言，再解释和重混颜色与风格。观察强烈形状和重复的鲜艳色彩。绘画比宽泛的风格词更具体，也保留了创作者的构图选择。可以用自己制作的素材作为起点，让生成器发展某些属性，同时明确哪些特征必须在解释与动画过程中保留。

# 58 / Overthinking

00:36:06 - 00:36:45

## Overthinking



### A deliberately nostalgic visual language

Zhang et al., Film Creation, CVPR Workshop 2025, Fig. 2

58

Overthinking uses a deliberately nostalgic visual language. The workflow diagram includes image adaptation, asset creation, animation, and editing. Read the branches as parts of this particular production, rather than a mandatory recipe. Several techniques contribute to the final style, and their contributions occur at different stages. This matters because viewers experience the finished sequence as a whole. They do not separate an image model's contribution from a timing decision in the edit. For our own work, we can still separate those contributions analytically. Doing so helps us decide where to intervene when the final sequence feels wrong, even though the individual images look suitable.

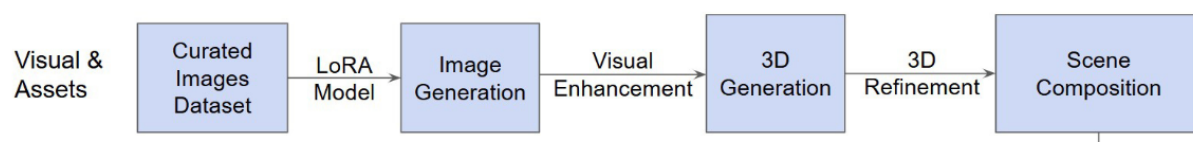
### 中文讲解

《Overthinking》采用怀旧的视觉语言，流程包括图像适配、资产制作、动画和剪辑。分支属于这部作品，并非通用配方。观众体验的是整个序列，不会把模型贡献与剪辑时机分开；但制作分析可以区分它们。这样，当单张图片合适而整段感觉不对时，才知道应该在哪个环节介入。

## 59 / Style begins before animation

00:36:45 - 00:37:22

### Style begins before animation



Overthinking builds assets before it builds movement.

Zhang et al., Film Creation, CVPR Workshop 2025, Fig. 2 (detail)

59

The upper branch of Overthinking begins with a curated image collection and adaptation of an image model. The paper discusses imagery associated with mid-century toys. A LoRA is a compact set of learned changes that can adapt a larger model to particular material. The later stages build and refine assets for the scene. The important creative decision happens before movement: the team chooses a visual world and prepares material that supports it. If you want a coherent art video, collecting references and deciding what belongs in that world can be more productive than generating unrelated shots and trying to make them match afterward.

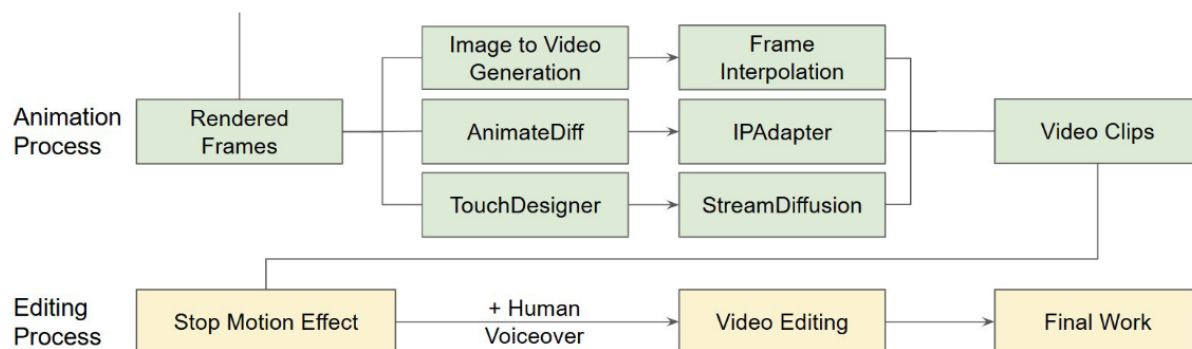
#### 中文讲解

《Overthinking》上方分支从精选参考图像和模型适配开始，论文讨论了与上世纪中叶玩具相关的图像。LoRA以一小组学习到的变化适配较大模型。关键创作决定在动画之前就发生：选择视觉世界并准备支持它的素材。要让艺术视频统一，整理参考和界定世界，可能比生成互不相关的镜头后再硬凑更有效。

## 60 / Timing can carry a style

00:37:22 - 00:38:00

### Timing can carry a style



Overthinking also shapes its motion during editing.

Zhang et al., Film Creation, CVPR Workshop 2025, Fig. 2 (detail)

60

The lower branches of the workflow show animation and editing. The survey describes adjusting playback cadence to evoke a stop-motion quality. This is a reminder that style has a temporal dimension. Two sequences can use similar images and still feel different because of how movement is sampled or paced. Smooth motion is appropriate for some intentions, while a stepped rhythm may suit another. For the studio painting, ask whether its movement should flow continuously or arrive in small unsettling changes. That choice may involve editing as well as generation. It should be judged against the intended experience, rather than against smoothness as a universal goal.

### 中文讲解

下方分支包含动画和剪辑，论文描述通过调整播放节奏营造定格动画感。风格也存在于时间之中：相近的图像可以因运动采样与节奏不同而产生不同体验。画室中的画应连续流动，还是以令人不安的小步变化？这个决定可能通过剪辑实现，应按艺术意图判断，而不是把流畅当作绝对目标。

## Clown



Variation becomes part of the character's expression.

Zhang et al., Film Creation, CVPR Workshop 2025, Fig. 3

61

Clown provides a different relationship to consistency. The survey discusses frame-by-frame style transformation and interprets visual variation as supporting a fragmented identity. Look at how the character's appearance changes while the performance remains related across the sequence. An inconsistency can become expressive material, but that does not make every accidental change meaningful. We need an account of how it affects the viewer. This case is useful because it challenges a simple rule that all AI artifacts must be removed. Instead, ask whether a particular variation supports the work, distracts from it, or communicates something the creator did not intend.

### 中文讲解

《Clown》使用逐帧风格转换，论文将视觉变化解释为对碎片化身份的表达。角色外观变化，而表演在序列中保持关联。不一致可以成为表达材料，但不是每个意外变化都有意义。要说明它如何影响观众：支持作品、分散注意，还是传达了非预期内容。因此，不能简单规定所有AI痕迹都必须消除。

## 62 / A source performance can survive a change of style

00:38:38 - 00:39:15

### A source performance can survive a change of style



Compare the hand gesture across both rows.

Zhang et al., Film Creation, CVPR Workshop 2025, Fig. 3 (detail)

62

Here we enlarge part of the comparison so that the hand gesture is easier to follow. Compare corresponding moments across the two rows. Which features of the performance remain legible even when the appearance changes? The position of a hand or the direction of a gaze can carry an action through stylization. At the same time, small changes around the face may alter the character's expression. These stills help us examine correspondence, but they cannot establish how smooth the full sequence feels. For that, we need playback. In your own project, compare both individual frames and the moving result before deciding that a transformation preserves the performance.

#### 中文讲解

放大的对比让手部动作更容易观察。比较两行对应时刻，辨认风格改变后仍可读的表演特征。手的位置或视线方向能延续动作，面部细节变化也可能改变表情。静帧能帮助分析对应关系，不能证明整段流畅。自己的项目应同时查看单帧与播放结果，判断转换是否保留了表演。



## 63 / Continuity can be a choice

00:39:15 - 00:39:52

### Continuity can be a choice

Would a stable face or a changing face serve this scene?

Zhang et al., Film Creation, CVPR Workshop 2025

63

Would a stable face or a changing face better serve this scene? There is no answer without an artistic intention. A conventional dialogue scene may depend on stable identity so that the audience follows the conversation. A work about unstable memory may use changing appearance as part of its meaning. Propose one intention for each version, and explain what the viewer should notice. Then identify a limit. Even a film about instability might need a consistent gesture, setting, or sound to hold the sequence together. Pause here if you want to discuss alternatives. The goal is to make continuity a considered choice rather than an automatic slogan.

#### 中文讲解

稳定的脸还是变化的脸更适合场景，取决于艺术意图。常规对话需要稳定身份帮助观众理解，关于不稳定记忆的作品则可能利用外观变化。为两个版本各提出意图，并说清观众应注意什么。即使表现不稳定，也可能需要持续的动作、场景或声音维序列。让连续性成为有意识的选择。

### For Pixi



A recurring character needs recognizable features.

Zhang et al., Film Creation, CVPR Workshop 2025, Fig. 4

64

For Pixi illustrates a case where a recurring character needs recognizable features. Look across the selected images. The blue material, large eyes, proportions, and general silhouette help establish continuity, although the appearances are not identical. The survey describes iterative generation and region editing as parts of the workflow. This is a production process that includes selecting and correcting, rather than expecting one instruction to guarantee every shot. The stills show the kinds of identity cues the creators work with. They do not prove perfect consistency across the entire film. As a viewer or researcher, keep the difference between a selected example and a complete evaluation clear.

### 中文讲解

《For Pixi》的角色通过蓝色材质、大眼睛、比例和轮廓保持可识别性，但并非每张都完全相同。论文描述了反复生成与区域编辑，说明制作包含选择和修正，不是一次指令就保证所有镜头。所选静帧展示身份线索，却不能证明整部电影完美一致。应区分精选案例与完整评估。

## 65 / A character reference is specific

00:40:31 - 00:41:09

### A character reference is specific



Which features make this character recognizable?

Zhang et al., Film Creation, CVPR Workshop 2025, Fig. 4 (detail)

65

The enlarged examples let us be more specific about a character reference. A phrase such as blue puppet supplies a category, but it leaves the shape of the eyes, the head proportions, and the material open. A reference can communicate these details more directly. Identify two features that must survive changes in view and expression. Then decide which variations are acceptable. Lighting should be allowed to change, for example, without changing what the character is made of. This exercise turns a vague request for consistency into a set of visible criteria. Those criteria can guide reference selection, local editing, and the rejection of unsuitable shots.

#### 中文讲解

“蓝色木偶”只是类别，没有确定眼睛形状、头部比例和材质，具体参考图能更直接地表达这些细节。指出两个在视角与表情变化时必须保留的特征，再界定允许的变化。例如，照明可以改变，但不应无意改变材质。把模糊的一致性要求变成可见标准，才能指导参考选择、局部编辑和镜头筛选。

## 66 / Fish Tank

00:41:09 - 00:41:48

### Fish Tank



Generated assets still enter a 3D production process.

Zhang et al., Film Creation, CVPR Workshop 2025, Fig. 5

66

Fish Tank connects generated material to a three-dimensional production workflow. The figure places finished scenes beside assets and a scene model. Notice that the final image depends on how objects are arranged and rendered, rather than only on their individual appearance. The case study reports manual work on UV mapping and topology. Those are ordinary production requirements that do not disappear when generation supplies an initial asset. For a filmmaker, the useful output is something that can participate in the intended scene. A beautiful preview may still require substantial work before the object can be lit, placed, or animated as the production demands.

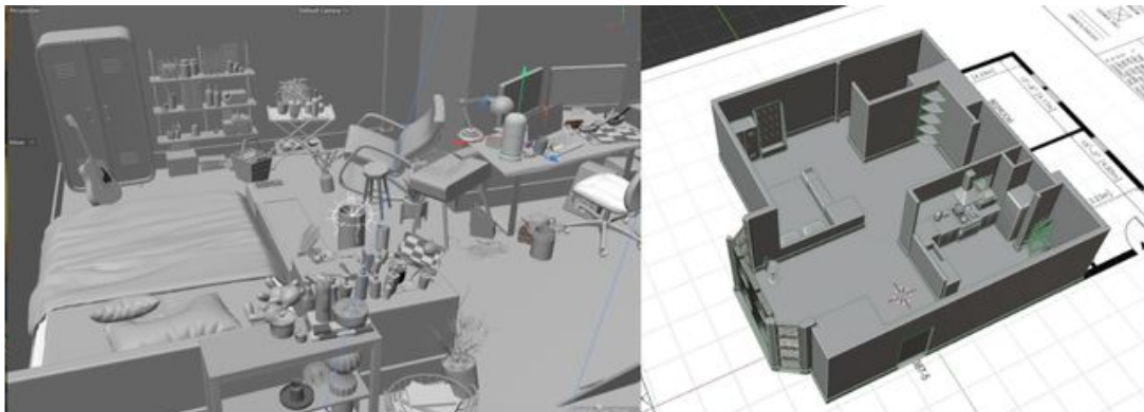
#### 中文讲解

《Fish Tank》把生成素材接入三维制作流程，图中并列最终场景、资产和场景模型。最终画面依赖物体如何排列和渲染，不只是每件资产本身好看。案例提到UV映射和拓扑的手工处理。生成起始资产不会消除这些要求；漂亮预览仍可能需要大量工作，才能用于布光、摆放和动画。

## 67 / A generated asset still needs production work

00:41:48 - 00:42:25

### A generated asset still needs production work



A scene needs usable geometry and consistent placement.

Zhang et al., Film Creation, CVPR Workshop 2025, Fig. 5 (detail)

67

Look more closely at the scene construction. Geometry determines the shape and placement of surfaces. Topology describes how a mesh connects, which matters when it needs to deform. UV mapping associates locations on that surface with texture information. A generated asset may need correction in any of these areas. This is why it helps to decide how an asset will be used before judging it. A background object seen briefly from one angle has different requirements from a character that bends in a close-up. In the studio, a chair that the returning person sits on needs more dependable structure than a distant decorative object.

#### 中文讲解

几何决定表面的形状和位置；拓扑描述网格如何连接，影响变形；UV映射把表面位置与纹理关联起来。生成资产可能在这些方面需要修正。先决定用途再评估：只从一个角度短暂出现的背景物，与需要在特写中弯曲的角色，要求不同。画室里供人坐下的椅子，比远处装饰物更需要可靠结构。

### Volumography



A captured scene can support later camera decisions.

Zhang et al., Film Creation, CVPR Workshop 2025, Fig. 6

68

Volumography refers here to creating moving images through spatial capture and rendering. The survey presents examples including Former Garden, Metanoia, and Dressage: Marching Through Memories. Instead of recording only a fixed sequence of frames, a spatial representation can permit later decisions about the viewpoint. This opens a different kind of creative process. Camera movement can become something explored after capture. The freedom is still limited by what the representation contains and how accurately it reconstructs the scene. Missing surfaces, reflective materials, and insufficient observations may produce gaps or distortions. Such effects can be production problems or, in some cases, material for artistic interpretation.

#### 中文讲解

这里的Volumography指通过空间捕捉和渲染创作动态图像。论文展示《Former Garden》《Metanoia》和《Dressage: Marching Through Memories》等案例。空间表示让视点选择能够发生在捕捉之后，运镜成为可继续探索的创作决定。但自由受到表示所含信息及重建质量限制，不能假定任何新视角都可靠。



## 69 / Fragmented space as an aesthetic

00:43:05 - 00:43:44

### Fragmented space as an aesthetic



Gaps and distortions can suggest a memory.

Zhang et al., Film Creation, CVPR Workshop 2025, Fig. 6 (detail)

69

This enlarged image shows a fragmentary spatial appearance. Parts of the scene seem incomplete, stretched, or loosely connected. In the cases discussed by the survey, such qualities can contribute to an experience of memory or uncertain space. We should avoid assuming that every visible gap was individually designed. The artistic decision may instead involve selecting and framing the behavior of a reconstruction process. Ask what the fragments direct your attention toward. Do they suggest loss, movement, or an unstable viewpoint? Then consider whether a more complete reconstruction would strengthen or weaken that reading. Technical fidelity and expressive effect can point toward different choices.

#### 中文讲解

放大的图像呈现碎片化空间：部分区域不完整、拉伸或松散连接。在论文案例中，这些特征可以营造记忆或不确定空间的体验，但不能假定每个空洞都经过单独设计。艺术决定也可能是选择并呈现重建过程的行为。碎片让你注意到失落、运动，还是不稳定视角？更完整的重建会强化还是削弱这种解读？技术忠实度与表达效果可能指向不同选择。

## 70 / A short film still needs decisions

00:43:44 - 00:44:22

### A short film still needs decisions

What changes?

When does it change?

What should the audience feel?

Classroom exercise · inspired by the readings

70

A short film still needs decisions about what changes, when it changes, and what the audience should feel. Return to the empty studio. Perhaps the painting moves only after the viewer has had time to notice its stillness. Perhaps the returning person never sees what happened. These choices shape the audience's knowledge and expectation. A generator can help produce the required material, but the relationship between moments still needs planning. This does not mean every art video must use conventional narrative. A work organized through repetition or rhythm also makes decisions about change and duration. Describe those decisions clearly enough that you can evaluate alternatives.

#### 中文讲解

短片仍需决定什么发生变化、何时变化、观众应有什么感受。画室中的画也许在观众充分注意到静止后才动，返回的人也许从未看到变化。这些选择塑造观众的知识与期待。生成器提供素材，但时刻之间的关系仍需规划。艺术视频未必要采用常规叙事，以重复或节奏组织的作品同样要决定变化与时长。



# 71 / A shot plan makes the next step concrete

00:44:22 - 00:44:59

## A shot plan makes the next step concrete

Opening: establish the empty studio.

Change: reveal movement inside the painting.

Ending: show who returns.

Classroom exercise · inspired by the readings

71

Here is one possible shot plan. The opening establishes the empty room. The next shot reveals movement inside the painting. The ending shows someone returning. Now make the plan more concrete. Is the first view wide enough to establish the door and painting together? Does the reveal use a camera move, an edit, or movement inside a fixed frame? Does the returning person's expression need to be readable? Each answer creates a technical requirement. You can then decide where to use an initial image, a motion reference, or local editing. The shot plan helps connect an artistic idea to manageable production tasks without determining every detail in advance.

### 中文讲解

一种镜头计划是：开场建立空房间，下一镜揭示画中运动，结尾有人返回。把它具体化：开场是否同时交代门与画？揭示通过运镜、剪辑还是固定画面内运动？返回者的表情是否必须可读？每个回答都产生技术要求，再据此选择初始图像、动作参考或局部编辑。镜头计划把艺术想法转成可执行任务。

## 72 / Film-making checkpoint

00:44:59 - 00:45:38

### Film-making checkpoint

A generated shot becomes useful when it serves the sequence.

Zhang et al., Film Creation, CVPR Workshop 2025

72

The filmmaking cases show several ways to turn generated material into useful shots and sequences. Some creators establish style through adapted image models. Others preserve a source performance, construct three-dimensional assets, or use spatial capture. Editing connects these decisions through timing and order. A generated shot becomes useful when it serves that larger sequence. Before moving on, choose one case that relates to a project you might make. Name the production problem it addresses and one limitation you would test. We will now examine shared technical ideas behind these workflows, including conditioning, scene representations, and the difference between generating pixels and generating movement data.

#### 中文讲解

电影部分把注意力从单个输出转向整个制作过程。生成提供素材，创作者还要组织表演、风格、空间和节奏。接下来回到视觉计算，理解不同表示为何支持不同编辑。选择技术时，不仅要问它能生成什么，也要问输出之后还能做什么。

# 04

## Visual computing

Po et al.  
State of the Art on Diffusion Models  
Computer Graphics Forum, 2024

Po et al., Visual Computing, 2024

73

Our fourth reading is State of the Art on Diffusion Models for Visual Computing by Po and colleagues, published in Computer Graphics Forum in 2024. It connects diffusion methods across images, video, geometry, and motion. We will use it to explain why related model ideas can support different kinds of output. The date matters here: these are foundational research examples, rather than a claim about the newest available products. Focus on the representation and the control signal. An image, a scene, and a motion sequence may all use learned generative knowledge, but they create different opportunities for later editing and rendering.

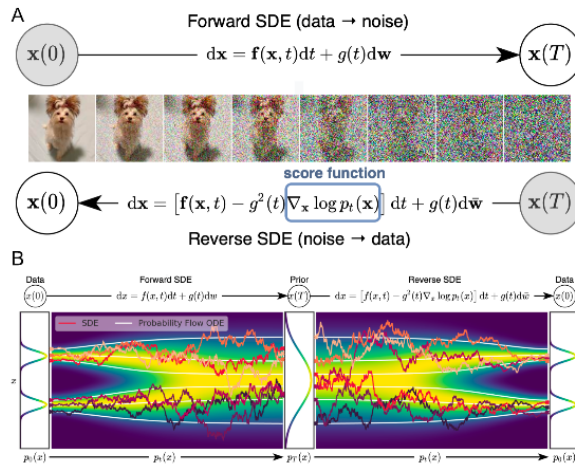
### 中文讲解

视觉计算综述把扩散模型放入更广泛的任务中，包括图像、视频、三维与运动。相似的生成思想可以作用于不同表示，但输入、输出和可控制属性并不相同。阅读时把模型名称与具体任务联系起来，辨认它生成或编辑的对象，以及适合怎样的创作需求。

## 74 / The same process across visual media

00:46:15 - 00:46:54

### The same process across visual media



A learned prior guides plausible visual outputs.

Po et al., Visual Computing, 2024, Fig. 1

74

A learned prior captures regularities in the material a model has seen during training. In practical terms, it influences what kinds of outputs the model considers plausible. The figure shows a process that moves from data toward noise and a learned process that moves back toward data. It also includes different mathematical formulations of those paths. You do not need the equations to understand the creative implication. The prior helps supply details that your inputs leave unspecified. Those details can be useful, but they are not guaranteed to be true. Plausibility should therefore be distinguished from faithful reconstruction, factual accuracy, and artistic appropriateness.

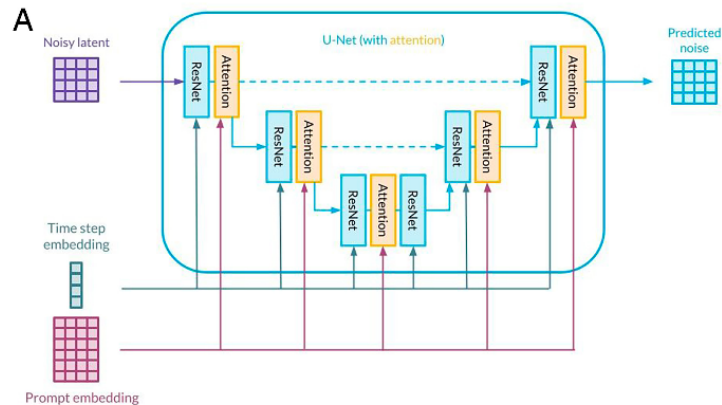
#### 中文讲解

生成先验来自模型从训练数据中学到的规律，它帮助补充输入没有确定的细节。无需掌握全部数学，也能理解创作含义：这些补充可能合理、有用，却不保证真实。应区分视觉可信、忠实重建、事实准确与艺术适切性；它们不是同一个评价标准。

## 75 / The denoiser uses the prompt at several scales

00:46:54 - 00:47:33

### The denoiser uses the prompt at several scales



The network combines noisy content with conditioning.

Po et al., Visual Computing, 2024, Fig. 4 (detail)

75

The upper diagram shows a denoising U-Net using information at several spatial scales. The U shape reflects a process that reduces and then rebuilds spatial detail, while connections carry information between stages. Prompt embeddings provide text conditioning. The time-step embedding tells the network about the current noise-removal stage. It does not indicate the moment of an action inside a video. As you trace the diagram, identify the noisy input, the conditions, and the predicted output. This reading strategy is more important than memorizing every block. It lets you connect a model's architecture to the information available when it makes each refinement.

#### 中文讲解

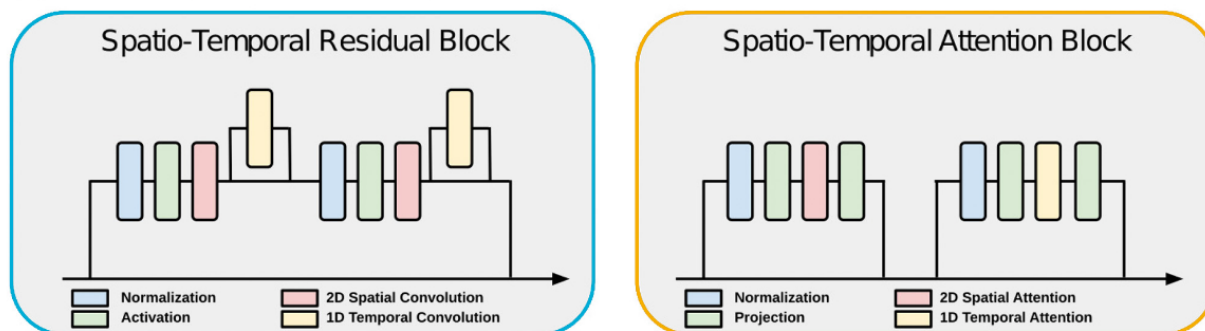
上图的去噪U-Net在多个空间尺度处理信息，通过缩小、重建和跨层连接协调细节。提示词嵌入提供文本条件，时间步嵌入表示当前去噪阶段，并非视频动作发生的时刻。沿图找到含噪输入、条件与预测输出，比记住每个模块名称更重要。

## 76 / Temporal layers connect the frame features

00:47:33 - 00:48:13

### Temporal layers connect the frame features

B



An image model gains operations that work across time.

Po et al., Visual Computing, 2024, Fig. 4 (detail)

76

The lower diagram shows a common way to extend image-processing blocks across time. Spatial operations connect information within each frame. Temporal operations connect information across frames. The legend distinguishes these roles, and the inserted temporal components show where the sequence gains additional connections. This is one architectural family, rather than a description of every modern video model. The principle is that treating every frame independently leaves consistency to chance. Connecting frames gives the system a way to coordinate their features. For our moving painting, such coordination can help preserve structure while appearance changes. It still needs suitable training and conditioning to produce the particular change we want.

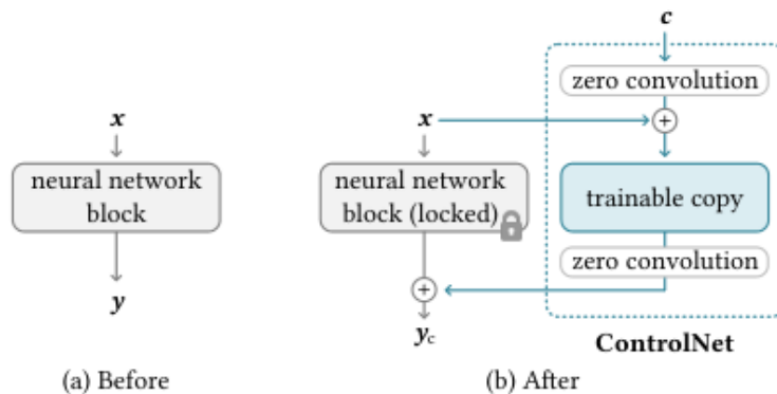
#### 中文讲解

下图展示将图像处理模块扩展到时间维度的一类方法：空间操作连接帧内信息，时间操作连接跨帧信息。它不是所有现代视频模型的统一结构，但说明一个原则：独立处理每帧会让一致性缺乏保障，跨帧连接提供协调特征的途径。具体变化仍依赖训练和条件。

## 77 / Conditioning adds a control signal

00:48:13 - 00:48:51

### Conditioning adds a control signal



A sketch or depth map can guide structure.

Po et al., Visual Computing, 2024, Fig. 3

77

Conditioning adds information that guides generation. ControlNet is a well-known example of a method that incorporates structural signals such as edges or depth. The diagram shows a trainable branch alongside a fixed base network. At a high level, the added branch learns how to use the extra condition. Think of a sketch that places the painting on the left wall of the studio. The sketch provides a spatial constraint that a broad text prompt might not preserve reliably. Different conditions constrain different properties. An edge map says little about the intended material or mood, so you may still need text, references, and later selection to complete the result.

#### 中文讲解

条件信息引导生成。ControlNet通过可训练分支接入边缘或深度等结构信号，与固定基础网络配合。草图可以比宽泛文字更可靠地指定画在左墙的位置。不同条件约束不同属性，边缘并不确定材质和情绪，所以仍可能需要文字、参考、筛选和编辑。

## 78 / Inversion and personalization

00:48:51 - 00:49:30

### Inversion and personalization

Inversion starts with an existing result.

Personalization adapts to a subject or style.

Po et al., Visual Computing, 2024

78

Inversion and personalization have different purposes. Inversion starts with an existing image and seeks a representation or generation path that can reproduce it, often approximately. That representation can then support edits. Personalization adapts a model or concept representation to a particular subject or style. A workflow may use both, but they answer different questions. If you want to edit one existing studio image, inversion may be relevant. If you want to generate many new views of a recurring character, personalization may be relevant. Neither label guarantees exact preservation. Check how well the method reconstructs the input or maintains the subject under the changes your project needs.

#### 中文讲解

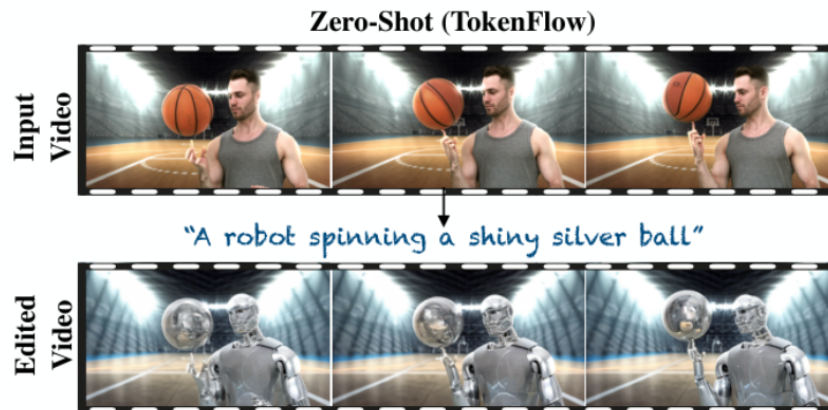
反演从已有图像寻找能近似重现它的表示或生成路径，以支持编辑；个性化则适配模型或概念表示，使其对应某个主体或风格。编辑一张已有画室图像可能需要反演，生成同一角色的多种新视角可能需要个性化。两者都不自动保证精确保留，应按实际变化需求测试。



## 79 / Video editing can preserve movement

00:49:30 - 00:50:09

### Video editing can preserve movement



TokenFlow changes appearance across a source sequence.

Po et al., Visual Computing, 2024, Fig. 5 (detail)

79

TokenFlow illustrates video editing that aims to preserve coherent movement while changing appearance. The top row shows the input performance. The lower row changes the subject and object while retaining a related sequence of poses and positions. The method uses correspondences between features to help keep the edit consistent. This differs from independently generating a new image for every frame. The example is useful for understanding the task, but selected frames cannot show every temporal issue. For a performance-based art video, inspect whether the intended gesture remains clear and whether changing the surface appearance has unintentionally changed the emotion or physical meaning of the action.

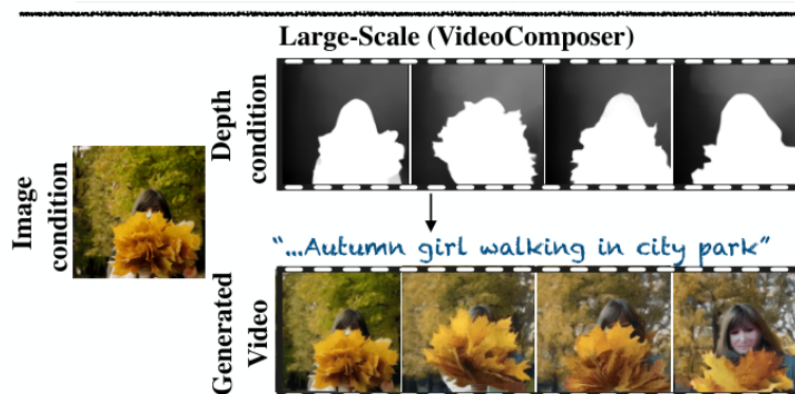
#### 中文讲解

TokenFlow在改变外观时利用特征对应关系保持视频编辑的一致性。比较输入表演与编辑结果，观察姿态和位置是否仍相关。这不同于独立重画每一帧。精选静帧不能展示所有时间问题。表演型艺术视频还要检查动作是否清晰，以及表面变化有没有意外改变情绪或身体动作的含义。

## 80 / Video generation can follow structure

00:50:09 - 00:50:50

### Video generation can follow structure



VideoComposer uses an image and depth information.

Po et al., Visual Computing, 2024, Fig. 5 (detail)

80

VideoComposer illustrates a different use of conditions. Here, an image supplies appearance information, while a depth sequence supplies structural guidance through time. The generated video combines those inputs with a text description. Look at how each condition answers a different question. The image helps specify what the subject looks like. Depth helps describe spatial structure and its progression. The text describes the desired content or action more broadly. This is a useful way to plan a controlled workflow: choose inputs according to the property each one constrains. More inputs do not automatically produce a better result, especially when their instructions or structures conflict.

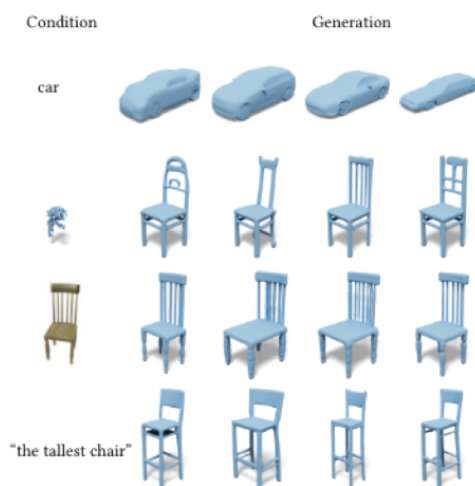
#### 中文讲解

VideoComposer的例子结合图像、深度序列和文字：图像提供外观，深度提供随时间变化的结构，文字描述内容或动作。规划可控流程时，应按每个输入约束的属性来选择。条件越多不一定越好，互相冲突的指令或结构也可能降低效果。

# 81 / A 3D object has many views

00:50:50 - 00:51:29

## A 3D object has many views



Geometry must agree when the viewpoint changes.

Po et al., Visual Computing, 2024

81

A three-dimensional object has many views, including surfaces that a single picture does not show. The examples display objects under different conditions and viewpoints. A front view can be attractive while the side or back contains implausible geometry. This is a distinctive requirement of reusable three-dimensional material. The representation needs enough consistency to support the views you intend to render. For our studio chair, consider whether it will appear only in the background or whether the camera will circle it. The second use demands more complete inspection. Evaluate generated geometry as an object with a purpose, rather than as a single favorable preview image.

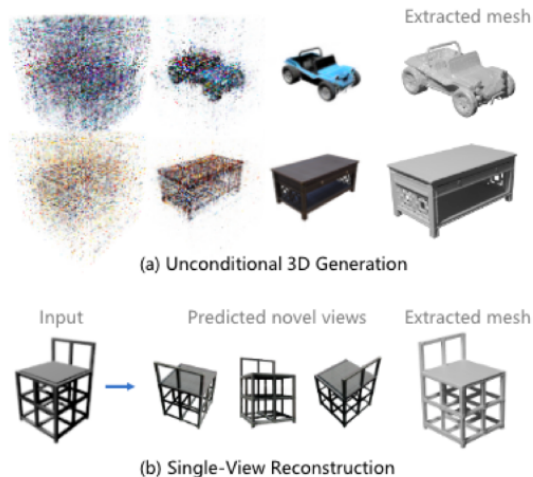
### 中文讲解

三维物体包含单张图片看不到的表面。正面好看，不表示侧面与背面合理。可重用三维素材需要支持计划中的视角。画室椅子若只在背景出现，与相机要围绕它移动，检查要求不同。把生成几何当作有用途的物体，而不只是一张有利角度的预览。

## 82 / Shape and appearance are different outputs

00:51:29 - 00:52:09

### Shape and appearance are different outputs



A mesh describes surfaces.  
Appearance describes how they look.

Po et al., Visual Computing, 2024

82

Shape and appearance are related but distinct outputs. A mesh describes surfaces through connected geometric elements. Materials and textures describe how those surfaces look when rendered. The figure separates visual examples from extracted meshes, making the distinction visible. A plausible texture can hide a structural problem from one angle. A clean mesh can also look unfinished until materials and lighting are added. If the object must animate, connectivity and deformation introduce further requirements. For production, inspect the representation that you will actually use. A rendered preview is useful evidence, but it does not replace checking whether the underlying asset supports the intended movement and camera views.

#### 中文讲解

形状与外观相关但不同：网格描述连接的表面，材质和纹理决定渲染时的外观。漂亮纹理可能从某个角度掩盖结构问题，干净网格也需要材质和照明。若要动画，还需检查连接与变形。评估真正要使用的表示，而不只看渲染预览。

## 83 / A scene adds spatial relationships

00:52:09 - 00:52:47

### A scene adds spatial relationships



A camera moves through a space with persistent structure.

Po et al., Visual Computing, 2024

83

A scene adds relationships between objects and spaces. As a camera moves through a room, doorways, furniture, and walls should form a coherent arrangement. The figure illustrates scene-level generation and related views. This is more demanding than producing an isolated object against a simple background. It is also important to distinguish a video that looks like a camera move from an editable scene representation. The video may look convincing without providing geometry that you can manipulate. If your project needs to change the path later, determine whether the workflow actually supports that operation. The appearance of spatial movement alone does not establish control over three-dimensional space.

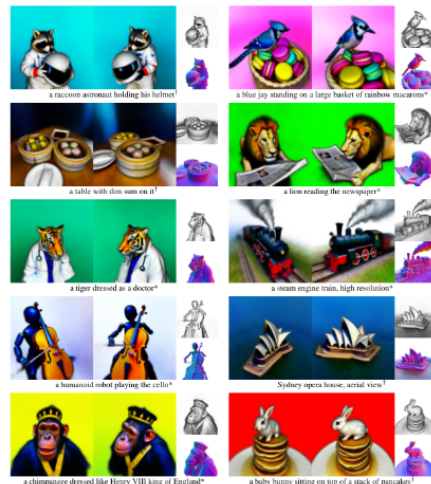
#### 中文讲解

场景增加物体与空间之间的关系。相机穿过房间时，门、家具和墙应形成连贯布局。看似运镜的视频不等于可编辑的三维场景；它可能没有可操纵几何。如果后续需要修改路径，必须确认流程支持这项操作。空间运动的外观本身不能证明三维控制能力。

## 84 / Image models can guide 3D creation

00:52:47 - 00:53:27

### Image models can guide 3D creation



DreamFusion uses image-based guidance to optimize a 3D result.

Po et al., Visual Computing, 2024, Fig. 7

84

DreamFusion uses image-based generative guidance to optimize a three-dimensional result. The basic idea is to render views of a candidate scene and use a learned image prior to guide improvements. The figure shows outputs from multiple viewpoints, along with other visualizations of their structure. This connects the knowledge of an image model to a different representation. It is an influential idea because large image collections can provide information that is difficult to obtain directly for every three-dimensional object. However, an image prior does not automatically ensure agreement across all views. Look for repeated features, distorted geometry, or inconsistent appearances when evaluating a generated object.

#### 中文讲解

DreamFusion利用图像生成先验指导三维结果优化：渲染候选场景的视图，再用学习到的图像规律提供改进方向。它把图像模型知识连接到另一种表示，但不会自动保证所有视图一致。检查生成物体时，要留意重复特征、几何扭曲和外观不一致。



## 85 / Editing a scene changes multiple views

00:53:27 - 00:54:07

### Editing a scene changes multiple views



InstructNeRF2NeRF edits a reconstructed scene.

Po et al., Visual Computing, 2024, Fig. 8

85

InstructNeRF2NeRF illustrates editing a reconstructed scene using language instructions. The examples alter features while showing the results from related viewpoints. A neural radiance field, or NeRF, represents a scene through a learned function used for rendering. Editing such a representation differs from repainting one isolated photograph. The goal is for the change to appear coherently across views. This can be useful when a creator wants to revise the appearance of an existing captured space. It does not imply that every edit creates physically complete or animation-ready geometry. Always connect the demonstration to the operations you require, such as new camera views or consistent appearance changes.

#### 中文讲解

InstructNeRF2NeRF展示以语言编辑重建场景，并从相关视角观察结果。NeRF通过学习函数表示可渲染场景，编辑它不同于重画一张照片，目标是让改变跨视角一致。这不等于每次编辑都生成物理完整、可直接动画的几何。应连接演示与实际需要的操作。

## 86 / The output format changes what we can edit

00:54:07 - 00:54:45

### The output format changes what we can edit

Pixels support direct image edits.

A scene representation supports new rendered views.

Po et al., Visual Computing, 2024

86

The output format changes what we can edit later. Pixels give us a finished image surface. A scene representation can support rendering from additional viewpoints. Movement data can separate an action from the character that eventually performs it. Consider two projects: a fixed-view poster and a camera orbit around a sculpture. The poster may need no three-dimensional representation at all. The orbit does. Choosing a more complex representation can add flexibility, but also adds preparation and evaluation work. Pause and name the most important later edit in your project. Then ask which output format makes that edit possible, and what limitations you would need to test before committing.

#### 中文讲解

输出格式决定后续能编辑什么。像素给出图像表面，场景表示支持新视角，运动数据把动作与最终角色外观分开。固定海报与绕雕塑一周的镜头需要不同表示。更复杂的表示带来灵活性，也增加准备和检查工作。先指出最重要的后续修改，再选能支持它的格式。



## 87 / Motion can be generated as movement data

00:54:45 - 00:55:24

### Motion can be generated as movement data



Audio can guide a dance before final rendering.

Po et al., Visual Computing, 2024, Fig. 9

87

Motion generation can produce articulated movement rather than finished video pixels. The example shows audio-conditioned movement sequences that can later be rendered with a character. This separates what the body does from how the final image looks. The same movement might be applied to different character designs, subject to the capabilities of the rig and retargeting process. This is valuable when timing and performance matter more than a particular rendered appearance. It also introduces its own evaluation questions. Are contacts plausible? Does the body maintain balance? Does the movement relate to the music in a meaningful way? A visually attractive render can conceal problems in the underlying motion.

#### 中文讲解

运动生成可以输出关节动作，而非最终视频像素。音频条件动作之后可应用到角色，分开“身体做什么”与“画面看起来怎样”。这取决于绑定与动作重定向能力。评估时检查接触、平衡和与音乐的关系，漂亮渲染可能掩盖底层运动问题。

# 05

## Creative industries

Anantrasirichai, Zhang, and Bull  
Advances in Artificial Intelligence  
Artificial Intelligence Review, 2026

Anantrasirichai et al., Creative Industries, 2026

88

Our final reading is *Advances in Artificial Intelligence: A Review for the Creative Industries* by Anantrasirichai, Zhang, and Bull, published in *Artificial Intelligence Review* in 2026. Its scope extends beyond generation to analysis, enhancement, spatial reconstruction, and other parts of production. Some illustrated examples are older than the publication date, so we will keep those dates clear. This wider view helps explain why improving a generator may not solve the main problem in a project. The difficulty may involve selecting a region, preserving a performance, refining an asset, or judging a result. We now connect those supporting capabilities to the creative workflow as a whole.

### 中文讲解

2026年创意产业综述把范围扩展到分析、增强、空间重建等制作环节。部分插图早于发表日期，需保留时间背景。改进生成器未必解决项目主要困难，瓶颈也可能是区域选择、表演保留、资产修整或结果评估。接下来把辅助能力连接到完整创作流程。

## 89 / Creation sits inside a larger workflow

00:56:02 - 00:56:40

### Creation sits inside a larger workflow

Before generation: intention and references.

After generation: selection, editing, sound, and review.

Anantrasirichai et al., Creative Industries, 2026

89

Creation sits inside a larger process. Before generation, we establish an intention and gather material that helps communicate it. After generation, we select, revise, combine, and review the results. Sound and editing can change the experience substantially. Think about where time actually goes in your own work. If the problem is unclear pacing, generating a sharper image may not help. If the problem is identity drift, a new music track may not help. Identify the stage responsible for the mismatch between intention and result. This lets you choose a targeted intervention and evaluate whether it resolves the problem, instead of treating every issue as a reason to generate again.

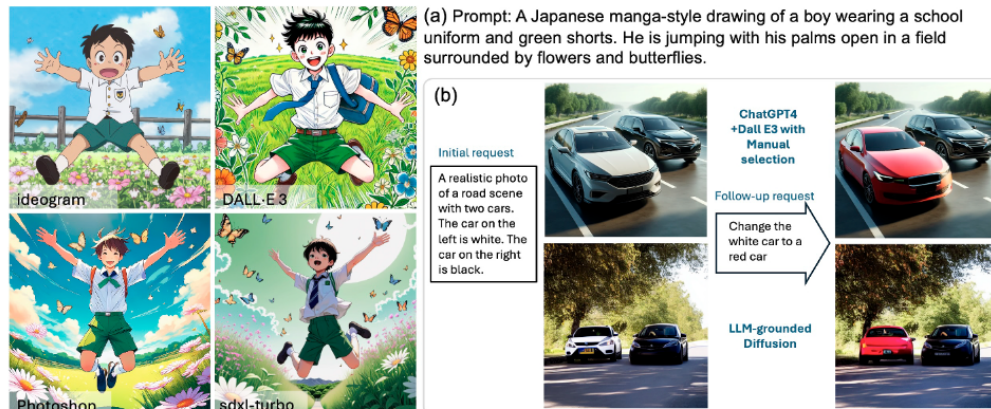
#### 中文讲解

生成处在更大的过程中：之前明确意图、收集素材，之后选择、修改、组合和审查。声音与剪辑会显著改变体验。如果问题是节奏，生成更清晰图像未必有用；身份漂移也不是加音乐就能解决。找出意图与结果偏差来自哪个环节，再做有针对性的介入。

# 90 / One prompt can produce different images

00:56:40 - 00:57:18

## One prompt can produce different images



Models interpret composition and style differently.

Anantrasirichai et al., Creative Industries, 2026, Fig. 3; examples dated 27 Nov 2024

90

This figure compares image examples produced from a prompt and includes an editing comparison. The examples are dated November 27, 2024. They are historical illustrations, not a current product ranking. Look at composition, object relationships, and the interpretation of style. A shared prompt can leave room for noticeably different results because models and generation settings differ. Even one model can vary across samples. For a fair creative comparison, decide which visible properties matter and inspect more than one output. Declaring a winner from a single attractive picture tells us little about reliability or whether the system can produce the specific sequence our project requires.

### 中文讲解

这组提示词生成与编辑示例日期为2024年11月27日，是历史图示，不是当前产品排名。比较构图、物体关系和风格解释。同样提示留下的空间、模型与设置差异，都可能导致不同结果。公平比较应先明确可见标准，并检查多个输出，而非用一张漂亮图片宣布胜者。

# 91 / A local instruction can affect the whole image

00:57:18 - 00:57:55

## A local instruction can affect the whole image



Compare the white car with its edited red version.

Anantrasirichai et al., Creative Industries, 2026, Fig. 3 (detail); examples dated 27 Nov 2024

91

The enlarged editing example asks for a white car to become red. This sounds local, but we should inspect the whole image. Did the second car remain stable? Did the road, lighting, and viewpoint change? A method can satisfy the most obvious part of the request while introducing unwanted differences elsewhere. The figure presents selected historical examples, so it should not be used as a current benchmark. Its teaching value is the evaluation procedure. Define the intended change and the features to preserve, then inspect both. This is the same procedure we would use when editing the painting inside our otherwise stable studio scene.

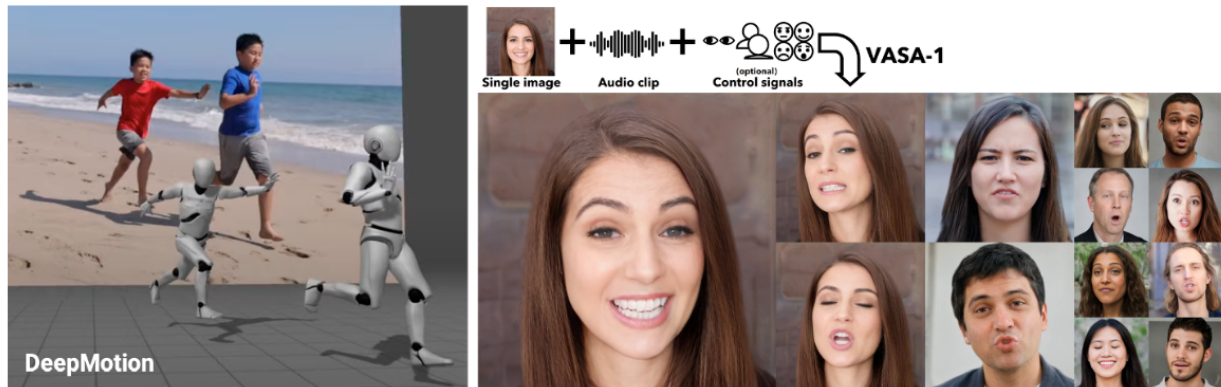
### 中文讲解

把白车改红看似是局部任务，但必须检查整幅图：第二辆车、道路、照明、视角是否稳定？模型可能完成明显要求，同时引入不希望的变化。图中精选历史案例的教学价值是评估程序：定义该改什么、该保留什么，然后同时检查。画室里的局部编辑也一样。

## 92 / Performance can come from captured motion

00:57:55 - 00:58:33

### Performance can come from captured motion



Body movement and facial animation use different inputs.

Anantrasirichai et al., Creative Industries, 2026, Fig. 4

92

Performance can come from different kinds of source information. The left example involves captured body movement and its use with an animated figure. The right example combines an image with audio for facial animation. These are separate systems, and their inputs constrain different properties. A body-motion reference tells us something about pose and timing. Audio can guide speech-related facial movement. A portrait supplies appearance but does not itself specify a whole performance. When planning a character shot, identify what you already have: a voice recording, an acted movement, or only a design. The available material helps determine which workflow is appropriate and what the model must invent.

#### 中文讲解

表演可以来自不同信息。左侧以捕捉的身体运动驱动动画，右侧结合图像与音频进行面部动画，它们是不同的系统。身体参考约束姿态和时机，音频支持说话相关运动，肖像提供外观却不指定完整表演。先辨认已有材料，才能决定合适流程以及模型还需创造什么。



## 93 / Post-production can change the result

00:58:33 - 00:59:13

### Post-production can change the result



Enhancement and editing continue after generation.

Anantrasirichai et al., Creative Industries, 2026, Fig. 5

93

Post-production continues after generation. The figure places image enhancement beside portrait editing to show different kinds of intervention. Enhancement may improve apparent detail or resolution. Editing changes selected content or expression. Both can be useful, but they require evaluation according to the intended task. A sharper face is not necessarily a more faithful face, and an expressive alteration may affect identity. For an art video, you might welcome some changes while rejecting others. Keep an original version so that you can compare what happened. It is easier to judge a revision when the intended improvement is specific and the earlier state remains available for reference.

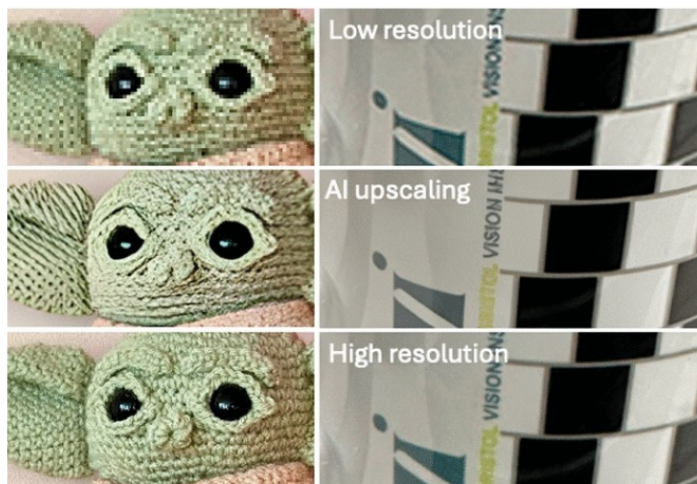
#### 中文讲解

生成后仍有后期工作。增强改变清晰度或细节，编辑改变内容或表情，两者应按目的评估。更清晰的脸未必更忠实，表情修改也可能影响身份。保留原版本便于比较，明确预期改善后，才能判断修改是否有效。

## 94 / Upscaling adds detail that needs inspection

00:59:13 - 00:59:51

### Upscaling adds detail that needs inspection



Compare the generated texture with the high-resolution reference.

Anantrasirichai et al., Creative Industries, 2026, Fig. 5 (detail)

94

Compare the low-resolution input, the AI-upscaled image, and the high-resolution reference. Pay attention to texture rather than just overall sharpness. The upscaled version can create convincing detail that differs from the reference. This happens because a model uses learned regularities to fill in information that the input does not fully contain. In a new artwork, that contribution may be acceptable or desirable. In faithful restoration, it may be a problem. For video, repeat the comparison across time to check whether the added texture flickers. The evaluation should follow the purpose of the work, rather than assuming that more visible detail always means better quality.

#### 中文讲解

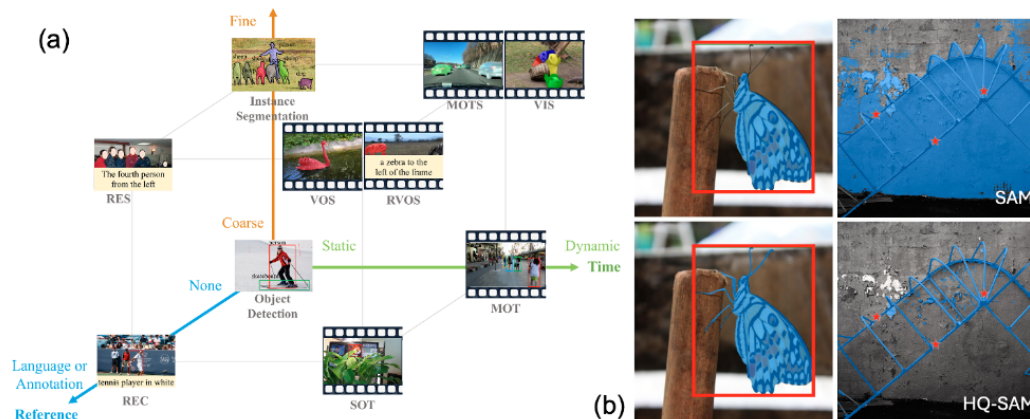
比较低分辨率输入、AI放大结果与高分辨率参考，要看纹理而非只看锐利程度。模型填补输入未确定的信息，可能产生可信却不同的细节。新艺术作品可能欢迎这些变化，忠实修复则未必。视频还需跨时间检查纹理闪烁。细节更多并不自动意味着质量更好。



# 95 / Selection makes local editing possible

00:59:51 - 01:00:30

## Selection makes local editing possible



Segmentation identifies a region.  
Tracking follows it through time.

Anantrasirchai et al., Creative Industries, 2026, Fig. 6

95

Segmentation identifies regions or objects in an image. Tracking follows an object through a sequence. Together, they can make local video editing more manageable. Suppose we want to alter the returning person's coat while leaving the room unchanged. We need to identify the coat and keep the selected region attached to it as the person moves. The figure organizes several object-understanding tasks and includes segmentation examples. These capabilities support generation and editing without necessarily generating the final content themselves. They are part of the larger workflow. Failures around occlusion or fast movement can cause an edit to jump, leak into the background, or disappear temporarily.

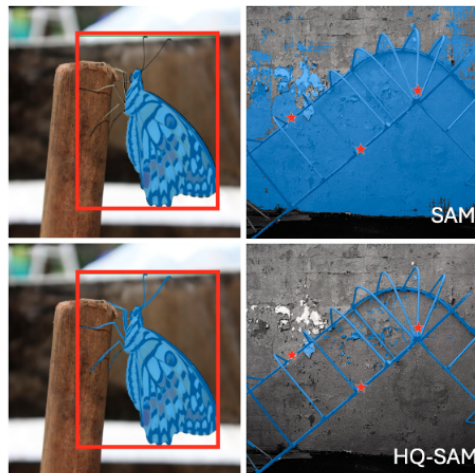
### 中文讲解

分割确定图像区域或物体，跟踪在序列中追随它。二者帮助局部视频编辑：改变外套时，要让选区随人物运动并保持背景不变。它们支持生成与编辑，却未必生成最终内容。遮挡或快速运动中的失败，会让编辑跳动、泄漏到背景或暂时消失。

## 96 / Fine boundaries make selection difficult

01:00:30 - 01:01:08

### Fine boundaries make selection difficult



Thin structures need careful masks.

Anantrasirichai et al., Creative Industries, 2026, Fig. 6 (detail)

96

Fine boundaries are difficult because small selection errors can be highly visible. Look at the butterfly outline and the thin structure in the comparison. Hair, transparent material, and narrow gaps create similar challenges in creative work. The figure compares selected outputs from SAM and HQ-SAM. It illustrates the importance of boundary detail, rather than proving that one method always performs better. A mask that looks acceptable in a still frame may still change inconsistently during movement. For our studio, the edge of the painting is a useful inspection point. A clean interior transformation can be undermined by a shifting outline or an edit that spills onto the frame.

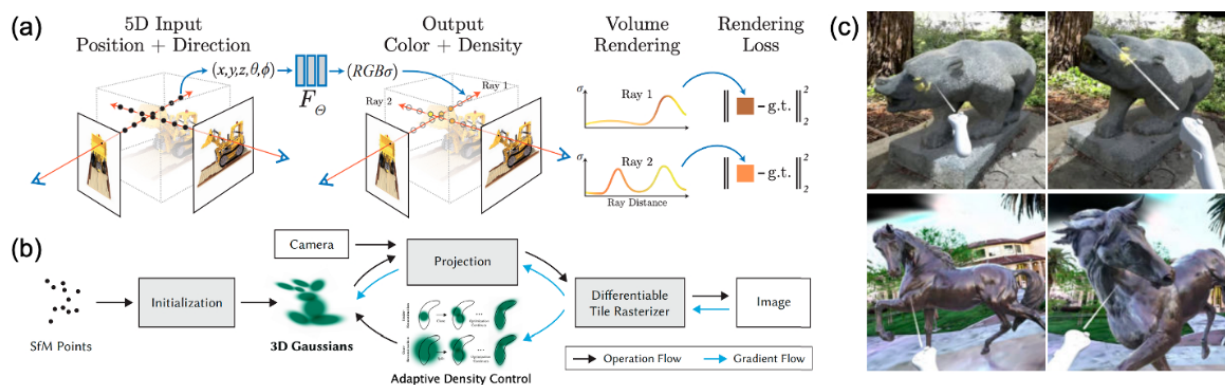
#### 中文讲解

细边界的小误差很容易显眼。蝴蝶边缘、细结构，以及头发、透明材料和狭窄空隙，都给选区带来挑战。SAM与HQ-SAM的精选对比说明边界细节的重要性，不证明某方法永远更好。静帧可接受的掩码也可能随时间不稳定，画框边缘尤其值得检查。

# 97 / Spatial capture supports new viewpoints

01:01:08 - 01:01:48

## Spatial capture supports new viewpoints



NeRFs and Gaussian splats represent scenes differently.

Anantasisrichai et al., Creative Industries, 2026, Fig. 7

97

NeRFs and Gaussian splats are different ways to represent scenes for rendering. A NeRF uses a learned field that describes properties queried along viewing rays. Gaussian splatting uses many spatial primitives with properties such as position and appearance. The figure places their processes alongside examples of spatial interaction. You do not need to master their mathematics to understand the production question: what views can this representation render well, and what can I edit? Capture coverage and reconstruction quality still matter. A scene representation does not automatically contain reliable hidden surfaces. Test the intended viewpoint range before building a film or interactive experience around that freedom.

### 中文讲解

NeRF与Gaussian splatting是不同的场景渲染表示。前者使用沿视线查询的学习场，后者使用带位置和外观等属性的空间基元。制作问题是它们能可靠渲染哪些视角、允许哪些编辑。捕捉覆盖与重建质量仍然重要，表示不会自动补齐可靠的隐藏表面。先测试计划中的视角范围。

## 98 / Quality has several meanings

01:01:48 - 01:02:26

### Quality has several meanings

A sharp image can still fail the scene.

A strange image can serve an artistic intention.

Anantrasirichai et al., Creative Industries, 2026

98

Quality has several meanings. An image can be sharp but fail to convey the intended action. A sequence can match a prompt yet feel emotionally wrong. A distorted image can also serve a deliberate artistic purpose. These observations call for separate judgments, rather than one vague claim that the output is good. For the studio, you could assess technical continuity, the clarity of the reveal, and the audience's interpretation. Different viewers may disagree about the last part, and that disagreement can be useful evidence. Explain the criteria you chose and why they suit the work. Automated measures can contribute, but they cannot settle every question about artistic experience.

### 中文讲解

质量有多个含义：图像可以清晰却动作不明，序列可以符合提示却情绪不对，扭曲也可能服务艺术目的。应分别判断技术连续性、揭示是否清楚、观众如何理解，而非笼统说好。不同观众的分歧也是证据。解释标准为何适合作品，自动指标不能解决所有艺术体验问题。

## 99 / A process others can understand

01:02:26 - 01:03:07

### A process others can understand

Record references and major edits.

Explain the decisions behind the result.

Anantrasirichai et al., Creative Industries, 2026

99

A process record helps other people understand what you made and how you made it. Save important references and meaningful before-and-after versions. Record model names and versions when available, along with settings needed to repeat a comparison. Explain the decisions that changed the work, including choices made in editing or sound. This is more informative than presenting a final output with only a prompt. It also helps you diagnose your own workflow when a later result differs. For coursework, follow the assignment's actual requirements. The broader habit is to make your contribution and source material understandable without turning the record into an unstructured collection of every experiment.

#### 中文讲解

过程记录帮助别人理解作品如何形成。保存重要参考与前后版本，记录可获得的模型名称、版本和比较所需设置，并说明剪辑、声音等关键决定。这比只附提示词更有信息，也便于诊断后来结果的差异。课程作业以实际要求为准；记录应解释贡献与来源，而不是堆积所有试验。

# 100 / Five readings to keep

01:03:07 - 01:03:47

## Five readings to keep

- 01 Video models: how a sequence forms
- 02 Visual art: what the artist controls
- 03 Film: how shots work together
- 04 Visual computing: what transfers across media
- 05 Creative industries: how the work reaches an audience

Anantrasirichai et al., Creative Industries, 2026

100

We have used five readings to follow a creative process from model behavior to finished work. The video survey explains how sequences form and how inputs constrain them. The visual-art survey connects methods to intention. The film survey shows creators combining tools and making production decisions. Visual computing explains how related ideas transfer across different representations. The creative-industries review places generation alongside supporting tools and evaluation. Choose one decision you would now change in your own project. Explain which reading informed it, what visible result you expect, and how you would test that expectation. That connection between a method, an intention, and evidence is the central skill to take from today's lecture.

### 中文讲解

五篇阅读沿创作过程展开：视频综述解释序列和条件，艺术综述连接方法与意图，电影综述展示工具组合与制作决定，视觉计算讨论不同表示，创意产业综述连接辅助工具与评估。选一个你会改变的项目决定，说明哪篇阅读启发它、期待什么可见结果，以及如何检验。方法、意图和证据之间的联系，是本讲核心能力。