第 3 章

AI 生图

3.1 了解AI生图的底层逻辑

AI 生图的底层逻辑相当复杂，涵盖了深度学习框架与神经网络模型、生成对抗网络、变分自编码器、扩散模型、潜在空间与多模态模型以及图像生成与优化等诸多领域。正是这些原理共同构筑了 AI 绘画的核心技术基石，使 AI 得以创作出富有艺术韵味的视觉作品。为了帮助大家更深入地理解 AI 生图的底层逻辑，作者将通过 ComfyUI 的文生图工作流，引导大家逐步探悉 AI 生图的过程。

首先，打开 ComfyUI，并加载文生图工作流，具体界面如图3-1 所示。这一工作流构成了 AI 文生图最基础的操作流程，任何更为复杂的 AI 生图任务都是基于这一流程进行拓展的。

图3-1

在文生图流程的起始阶段，首先需要加载 Stable Diffusion 大模型。在“Checkpoint 加载器 (简易)”节点，可以选择生成图片时所使用的底模。简而言之，想让 AI 生成何种风格的图片，就选择对应风格的底模。

为了生成满足特定要求的图像，还需要输入正向和反向提示词。两个“CLIP 文本编码器”节点分别代表正向提示词和反向提示词的功能，它们能将提示词经过编码，转化为机器可理解的格式并传入潜空间，后续会对此进行详细讲解。在“正向提示词”文本框中输入的内容，AI 会视为希望在生成图片中呈现的元素，并据此进行绘制；而在反向提示词文本框中输入的内容，AI 则会视为不希望在生成图片中出现的元素，从而在生成时减少或去除这些负面提示词所描述的图像元素。

为了控制生成图像的大小，还需要设置相应的图像尺寸。在“空 Latent”节点，可以设置生成图片的尺寸。该节点会创建一个初始的潜在空间图像，为后续操作奠定基础。这可以简单理解为生成了一张特定大小的画布，AI 将在这张画布上生成图片，而不会随意产生不符合要求的图像尺寸。

为了调整生成图像的质量和风格，还需要对采样器参数进行设置。“K 采样器”节点结合所设置的参数以及传输来的模型、提示词、图片等进行图像的绘制。具体而言，系统在潜空间内生成一张随机种子噪声图，并在采样器的引导下逐步降噪，最终生成图像。

“K 采样器”节点所绘制的图像信息仍处于潜空间状态，需要经过解码才能转换到像素空间，后文会详细阐述这一过程。此时，需要利用“VAE 解码”节点对图像进行解码，再将其输出给“保存图像”节点，从而生

成我们所需的图像。

综上所述，整个文生图的流程可以概括为：CLIP 模型将输入的提示词进行转换，接着由 K 采样器在潜空间中生成图像，再经过 VAE 将图像解码为像素空间的图像，并最终输出图像。整个过程的流程如图3-2 所示。

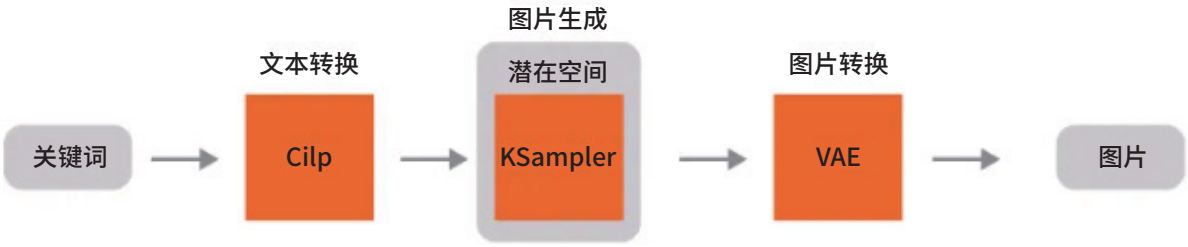

图3-2

通过工作流，我们可以更直观地感受 AI 生成图像的过程。实际上，AI 生成图像的过程相对抽象。为了让大家能更直观地了解图像的降噪过程，作者特此截取了“K 采样器”节点在降噪过程中的图片，如图3-3 所示。

图3-3

在讲解完 ComfyUI 的文生图工作流程后，相信单击对 AI 生成图像的过程已经有了初步的了解。这将有助于大家更轻松地理解 AI 生成图像的底层逻辑。接下来，将对相关概念进行详细的阐述。

3.2 CLIP模型

在 Stable Diffusion 模型中，CLIP 扮演着一个至关重要的角色。它负责将文本与图像相互关联，从而确保模型能够根据给定的文本描述生成相应的图像。

3.2.1　CLIP的基本概念

CLIP 模型是由 OpenAI 公司开发的一种多模态模型，它能够理解文字和图片之间的关联，进而实现更为精确和一致的图像生成。CLIP 模型包含两个主要组件：一个是文本编码器，负责理解文字；另一个是图像编码器，负责解析图像。

1. 文本编码器

文本编码器宛如一位翻译官，其职责在于将我们输入的文字描述转换为模型可理解的一系列数字代码，这些代码被称为“文本向量”。由此，模型便能够“领悟”我们意欲生成何种图像。

2. 图像编码器

图像编码器宛如一位摄影师，其功能在于将图片转换成一系列数字代码，这些代码被称作“图像向量”。由此，模型便能将图片的内容转换为可处理的信息。

3.2.2　CLIP在Stable Diffusio中的应用

1. 文本指导图像生成

CLIP 技术使 Stable Diffusio 模型能够依据创作者提供的文本提示来生成相应的图像。例如，当输入“一只可爱的猫咪”时，CLIP 会将这个文字描述转换成模型可理解的数字代码，随后 Stable Diffusio 根据这些代码生成相应的图片。

2. 多模态对齐

CLIP 技术能够让文字和图片在同一个“空间”内实现对话。举例来说，当我们输入“夕阳下的海滩”时，模型不仅能够领悟“夕阳”与“海滩”这两个词汇的含义，还能够在所生成的图片中将这两个元素巧妙地融合在一起。这种对齐机制对于生成高质量且紧密贴合文本描述的图像至关重要。

3. 条件生成

Stable Diffusio 模型以 CLIP 生成的文本向量为条件，将其与初始噪声相融合，并在扩散模型的逐步去噪过程中发挥指导作用。这一过程好比依照一份食谱进行操作，其中输入的文本描述就如同食谱中的原料与制作步骤，Stable Diffusio 模型则依据这份“食谱”逐步生成图片。而 CLIP 在此过程中助力模型理解这份“食谱”，从而确保最终生成的图片与我们的描述相吻合。

3.2.3　CLIP参数的影响

在 Stable Diffusio 模型中，CLIP 终止层数的设定确实会对生成的图像产生显著影响。具体而言，CLIP 终止层数决定了在将文本提示转化为文本向量时，CLIP 模型所使用的层数深度。若 CLIP 终止层数设定得较小，模型将会利用更多的层次来深入提取文本特征，从而生成与文本描述更为贴切的图像。反之，若 CLIP 终止层数较大，模型则会提前终止文本特征的提取过程，这可能导致所生成的图像与原始文本描述之间存在一定的偏差。因此，在实际操作中，我们需要根据具体需求和期望效果来灵活调整 CLIP 的相关参数。

在 ComfyUI 中，几个常用的 CLIP 节点，包括“CLIP 文本编码器”节点、“双 CLIP 加载器”节点以及“CLIP 视觉加载器”节点。这些节点的面板图示及彼此间的连接情况，如图3-4 所示。

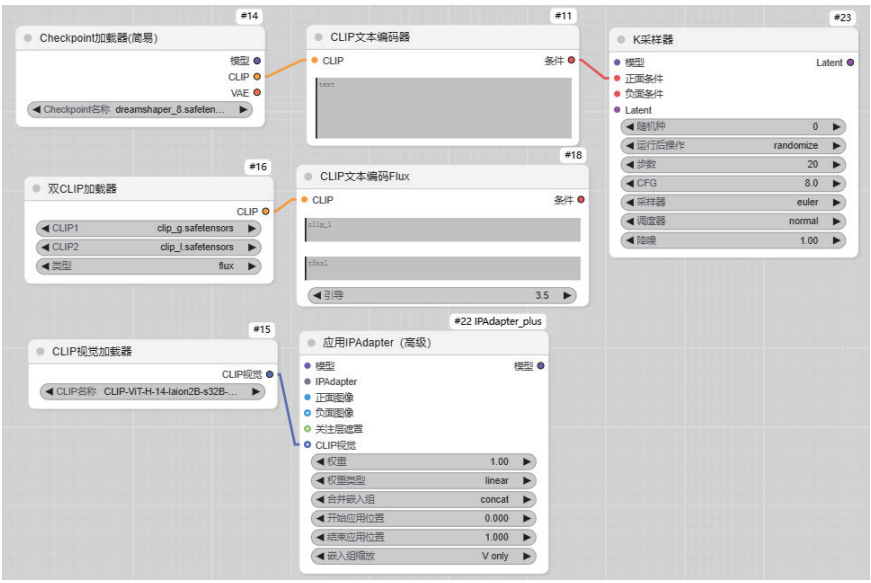

图3-4

3.3 Latent空间

在 Stable Diffusion 模型中，Latent 是一个核心概念，它关乎模型如何处理并生成图像。

3.3.1 Latent空间的定义

Latent 空间，亦被称作“潜在空间”，可被视为一个“图像的秘密仓库”。在此仓库内，图像不再是我们日常所见的由无数像素点构成的复杂形态，而是被转化为一种更为简洁与抽象的形态。这种形态便是图像的“潜在表示”，即 Latent 空间中的存储内容。

Latent 空间在 Stable Diffusion 模型中扮演着一个“中转站”的角色。编码器（encoder）如同打包员，将复杂的图像“打包”为一个小巧的包裹，即低维的潜在表示，随后将其存入 Latent 空间中。接着，扩散模型的主体部分，例如 UNet，作为包裹处理员，会取出仓库中的包裹进行各种变换与处理。这些操作旨在使图像更为丰富、更贴合我们的期望。最后，解码器（decoder）则扮演拆包员的角色，将处理过的包裹重新打开，呈现经过处理的图像，供我们查看最终成果。

简而言之，Latent 空间助力模型将复杂图像简化为更易于处理的形式，再通过一系列的处理与变换，最终生成创作者所期望的图像。这一过程宛如我们日常生活中的快递寄送：先将物品打包妥当，再交由快递公司进行后续的处理与运输，最终由收件人拆包验收。

3.3.2 Latent空间在Stable Diffusio中的应用

Latent 空间在 Stable Diffusio 模型中发挥着核心作用，从图像生成到编辑，再到超分辨率处理，其潜力

被充分挖掘。接下来，将深入探讨 Latent 空间在这些应用中的具体作用。

- 图像生成：Stable Diffusion模型巧妙利用Latent空间中的潜在表示来塑造图像。通过精细调整这些Latent表示，模型能够生成各具特色、风格迥异或细节丰富的图像作品。
- 图像编辑：Stable Diffusion的局部重绘功能赋予了用户直接在Latent空间内对图像特定区域进行编辑的能力。例如，用户可以轻松地删除图中的多余元素或增添新元素，而这一切都无须直接触及原始图像的像素层面。
- 图像超分辨率：Stable Diffusion 2.0更进一步，引入了强大的Upscaler Diffusion模型。这一高阶模型在Latent空间上施展“魔法”，能够将图像的分辨率提升至原先的4倍，带来更细腻的视觉体验。

此外，在 ComfyUI 的操作界面中，几个常用的 Latent 节点如“空 Latent”节点、“Latent 批次大小”节点以及“Latent 缩放”节点，为用户提供了灵活的操作选项。这些节点的面板图示及连接情况，如图3-5 所示。

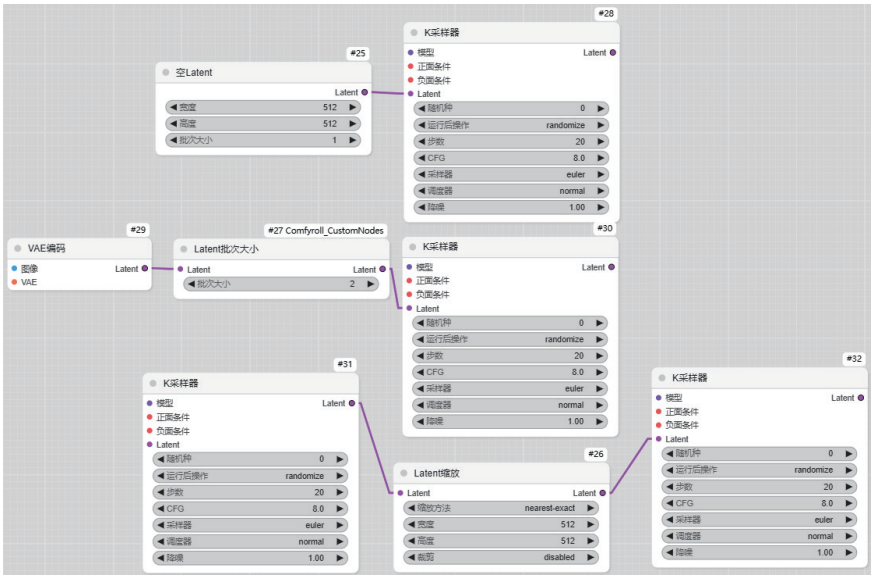

图3-5

3.4 UNet神经网络

在 Stable Diffusion 模型中，UNet 是一种至关重要的神经网络架构，它并非专门用于图像分割任务，而是在该模型的图像生成过程中发挥着关键作用。

3.4.1 UNet的工作流程

在 Stable Diffusion 模型的推理阶段，UNet 的工作可以想象成一位画家在创作一幅画。这个过程具体如下。

- 接收指令：首先，UNet会接收一个指令，其中包含了创作者想要生成的图像描述，例如“一只坐在

草地上的小猫”。

- 开始作画：接收到指令后，UNet便在一张“白纸”上开始作画。然而，这张“白纸”并非真正的白纸，而是一张充满随机噪声的图像，类似电视无信号时出现的雪花屏。
- 逐步清晰：运用其独特的技术，UNet逐步从噪声中提炼出清晰图像。这一过程宛如画家用橡皮擦轻轻擦去多余的线条，使图像逐渐变得清晰。
- 细节描绘：随着UNet的持续工作，图像的细节开始逐渐呈现。例如，小猫的皮毛、眼睛和胡须等细节都会慢慢浮现。
- 完成作品：当UNet将所有多余的线条擦除后，原本的噪声图像便转变为创作者最初所描述的那只坐在草地上的小猫。

综上所述，UNet 在此过程中扮演了一个根据文本描述和随机噪声创作图像的艺术家角色。

3.4.2 UNet在Stable Diffusio中的作用

在 Stable Diffusion 中，UNet 发挥着举足轻重的作用，其作用具体体现在以下几个方面。

- 语义分割：借助UNet，可以对生成的图像进行精准的语义分割，从而将图像中的不同物体或区域有效区分。例如，它能够将图像的前景与背景进行分离，或者识别并分割出图像中的各个独立物体，这对于提升生成图像的质量和真实感至关重要。
- 迭代降噪：在Stable Diffusion模型的工作流程中，UNet是负责从噪声中逐步还原出图像的核心组件。通过其反复的调用和运算，模型能够逐步剔除预测输出中的噪声成分，进而获得越来越清晰的图像表示。
- 图像生成：在预测阶段，UNet会经历多次迭代过程。在每一次迭代中，它都会根据前一次的输出以及输入的文本描述进行细致的调整和优化，从而确保最终生成的图像能够更加贴近目标效果。
- 精准分割：得益于其独特的编码器—解码器架构以及跳跃连接设计，UNet能够高效地整合和利用来自不同层级的特征信息。这种设计不仅增强了模型的表征能力，还显著提高了图像分割的精确度。

在 ComfyUI 中，涉及 UNet 模型的节点包括“Checkpoint 加载器（简易）”节点和“K 采样器”节点。这些节点的面板图示及其之间的连接关系，如图3-6 所示。

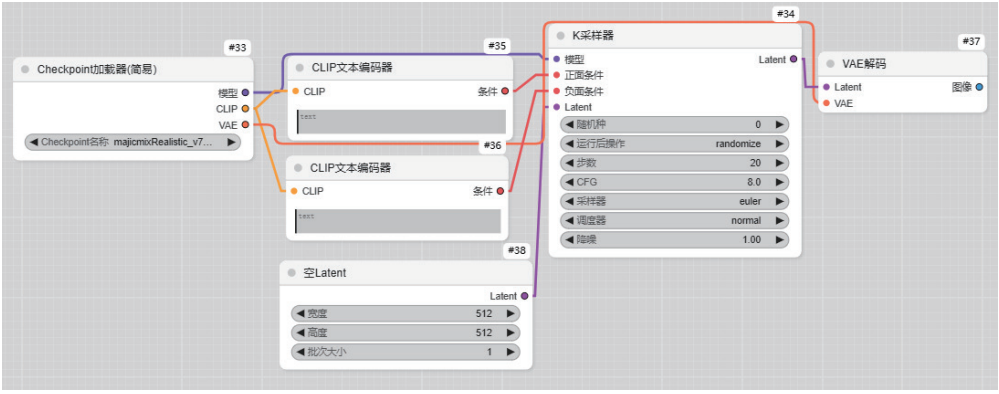

图3-6

3.5 VAE变分自编码器

在 Stable Diffusion 模型中，VAE（变分自编码器）是一个核心概念，它对提升模型的图像生成能力起着至关重要的作用。

3.5.1 VAE的基本概念

VAE（变分自编码器）是一种生成模型，在 Stable Diffusion 中扮演着将输入的文本表示转化为潜在变量，并基于这一潜在表示空间生成新图像的角色。我们可以将 VAE 的作用类比为一个压缩与还原的过程。

- 压缩功能：VAE编码器负责将图像数据压缩成潜在空间中的低维表示。这一过程类似将一本书的详尽内容精简为一个概括性的提纲或摘要。具体来说，编码器会将高维的图像数据（例如512×512像素的图像）“打包”成一个更为紧凑的形式，如4×64×64的矩阵。
- 还原功能：相应地，VAE解码器则承担着将这些潜在表示还原回图像的任务。当我们从潜在空间中提取这些低维表示，并期望查看它们所对应的原始图像时，解码器能够将这些编码还原为图像。尽管还原后的图像可能与原始图像不完全一致，但在整体上会呈现高度的相似性。

3.5.2 VAE在Stable Diffusion中的作用

VAE 在 Stable Diffusion 中发挥着举足轻重的作用，其贡献主要体现在以下几个方面。

- 图像生成：借助潜在空间中的随机采样以及解码器的精准转换，VAE能够生成与输入文本描述高度契合的图像。
- 图像质量提升：VAE对于提升图像质量同样功不可没。通过精心优化编码与解码过程，它可以生成色彩更为鲜艳、细节更为丰富的图像。
- 细节修复：VAE在细节修复方面同样表现出色，例如修复人脸、手部等区域的瑕疵。这得益于它在潜在空间中深刻捕捉到了图像的语义与结构信息，以及训练过程中积累的大量细节特征知识。

在 Stable Diffusion 框架下，用户可以根据需求选择多种 VAE 模型，如 kl-f8-anime、vae-ft-mse-840000-ema-pruned 等。这些模型在训练数据集、模型架构及优化策略上各有千秋，因此，所生成的图像在质量和风格上也各具特色。创作者可以根据自己的艺术追求和实际需求，挑选最合适的 VAE 模型，并通过调整采样步数、噪声水平等关键参数，进一步优化生成结果。

在 ComfyUI 中，常用的 VAE 节点包括“VAE 编码”节点、“VAE 解码”节点、“VAE 加载器”节点以及“VAE 内补编码器”节点。这些节点的具体面板图示及其连接关系如图3-7 所示。

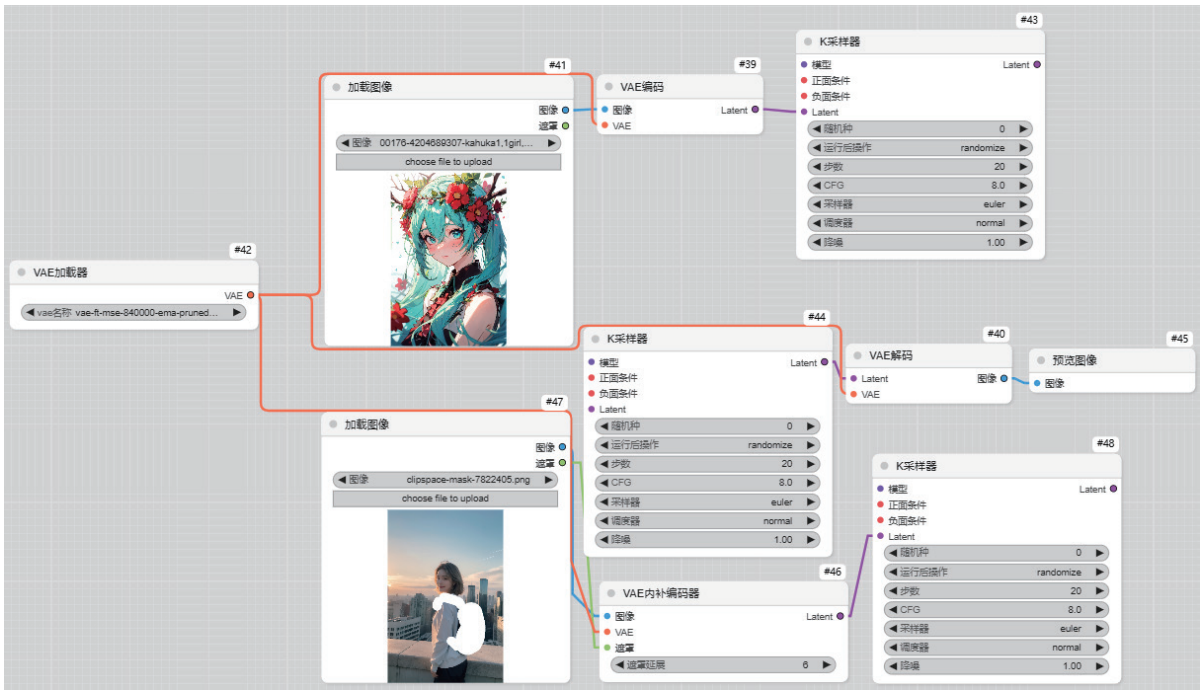

图3-7

3.6 ComfyUI图生图的底层逻辑

在 ComfyUI 中，图生图的流程如图3-8 所示，该流程的底层逻辑可简化为以下几个核心步骤。

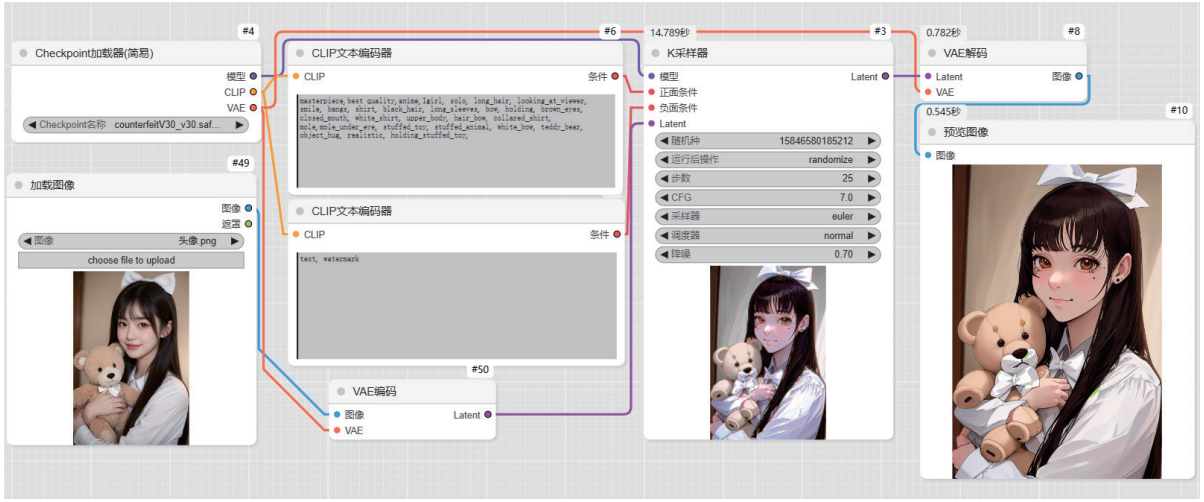

图3-8

- 图像编码：在图生图流程中，需要先上传参考图像，然后通过“VAE编码”节点，将参考图像从高维的像素空间转换至低维的潜在空间，从而得到一个紧凑且信息丰富的特征向量。这个过程既实现了图像的压缩，也有效提取了图像的主要特征和结构信息。该特征向量将作为后续新图像生成的基础，它承载了图像的核心内容，确保新生成的图像能够在一定程度上承袭参考图像的特征。
- 条件向量的生成与应用：创作者所输入的文本提示词，对于新图像的生成起着至关重要的指导作用。这些条件通过“CLIP文本编码器”节点转换为向量形式，即条件向量。在潜在空间中，条件向量与先前的特征向量相结合，形成一个融合了特征信息与创作者条件信息的新向量。这种结合确保了新生成的图像既契合创作者的意图，又具备足够的创意与多样性。
- 采样器的选择与参数配置：在图生图过程中，采样器扮演着举足轻重的角色。不同的采样器及其参数设置（例如降噪强度、采样步数等）会显著影响新图像的特征与风格。采样器根据潜在空间中的结合向量生成新的特征向量，这一过程可视为对潜在空间特征的精细调整或重新组合。通过调整采样器参数，创作者能够精准掌控新图像的生成过程，使其更加贴近个人期望与需求。
- 图像解码：在新图像生成的最后阶段，需要将新的特征向量从潜在空间解码回像素空间。这一步骤通过“VAE解码”节点完成，它能够将潜在空间中的特征向量转化为真实可见的图像。解码过程不仅涉及对潜在空间特征的解码与重构，更是将创作者的条件与参考图像特征完美融合，从而生成既具创意又富多样性的新图像。

关于图生图中噪声的应用，这也是一个值得关注的环节。噪声的引入能够增加生成图像的多样性与艺术性，使其呈现更为自然与逼真的效果。通过调整噪声强度，创作者可以控制生成图像与参考图像之间的相似度与差异度。在采样器节点中设置降噪强度参数是实现这一目的的有效手段。较小的降噪强度参数会使生成图像更贴近参考图像，保留更多原始特征；而较大的降噪强度参数则会使生成图像与参考图像产生更大差异，展现出更多创新元素与多样性。

综上所述，ComfyUI 图生图的底层逻辑是基于 Stable Diffusion 模型及其工作流节点的精心组合与配置。通过加载模型、上传图像、进行图像编码、设置采样器参数、生成新图像以及保存成果等一系列步骤，创作者能够将输入的图像转化为符合个人要求的新作品。在整个流程中，关键节点的选择与参数设置对于生成图像的质量与风格具有直接影响。