跨模态生成是一种生成任务，其目的是通过对不同模态对象的对齐映射，生成符合目标模态的数据。在生成技术中，生成的定义是通过特定算法和模型，从给定的输入中生成所需的输出数据。生成技术在现代人工智能应用中扮演着至关重要的角色，例如对话系统中的生成技术是其核心基础。

当前，对于单一模态生成技术，已经存在大量基于深度学习的架构和算法。

- 文本模态：主要使用基于RNN和Transformer的架构，其任务目标是预测下一个词元（next token prediction）。这些模型通常包含一个编码器，用于处理输入文本，并通过beam search或greedy search等策略控制生成的序列。
- 视觉模态：包括VAE（变分自编码器）、GAN（生成对抗网络）、扩散模型（Diffusion Model）等基础架构，不同于文本模态，其训练目标多种多样，包括去噪、与判别器对抗、重构原始图像等。

在跨模态生成任务中，尽管其输出最终会落脚到某一个模态上，但面临着一系列独特的挑战。这些挑战在单一模态生成任务中并不常见或不显著。

- 可控性（Controllability）：用户对模型输出的可控性在跨模态生成中尤为重要。例如，在使用跨模态生成模型生成图片时，用户可能希望生成一张在沙滩上玩耍的儿童图片。可控性意味着用户可以具体控制图片中的细节，如沙滩的颜色、儿童的数量和姿势、背景中的物体等。
- 组合性（Compositionality）：跨模态生成模型需要处理复杂的组合关系。例如，用户希望生成一张图片，内容是一个红色汽车停在蓝色房子旁边。组合性要求模型能够理解并正确生成这种复杂的组合关系，确保生成结果符合用户的期望。
- 同步性（Synchronization）：不同模态可能具有不同的特征空间和分布，需要处理好不同模态之间的同步关系。例如，在制作视频时，字幕需要与视频内容同步。当一个人在视频中说话时，字幕必须准确地与其讲话时间匹配。
- 长尾效应（Long Tail Phenomena）：跨模态生成可能面临模式崩溃的问题，即无法生成小概率事件。例如，生成一张非常特殊的图片，如独角兽在城市街道上漫步。尽管这种情景在现实中很少见，在训练数据中也可能没有出现过，理想情况下，生成模型应该能够成功生成这样的特殊内容。

视觉到文本生成

图像描述生成（Image Captioning）是计算机视觉和自然语言处理交叉领域的一项重要任务，旨在使计算机能够理解图像内容并生成自然语言描述。其主要目标是从输入的图像中提取出有意义的视觉特征，并将这些特征转化为连贯、详细的文字描述。这一任务不仅需要模型具备强大的图像理解能力，还需要具备良好的语言生成能力。

图像描述生成的核心困难包括多模态信息的融合、语义一致性和细节捕捉。多模态信息的融合涉及如何有效地将视觉信息和语言信息结合在一起，确保生成的描述既符合图像内容又流畅自然。语义一致性要求生成的描述在语义层面上连贯一致，避免生成与图像内容不符的描述。细节捕捉则关注如何在生成描述时充分反映图像中的细节和复杂关系，例如对象之间的相对位置、动作和场景背景等。

传统方法

1. **Every picture tells a story: Generating sentences from images (2010)**：这篇论文提出了通过检索和重排生成图像描述的方法。该方法首先检测图像中的视觉元素，然后在语料库中检索相关描述片段，最后通过重新排列组合生成描述。虽然这种方法简单直观，但依赖于预定义的模板和现有的语料库，缺乏灵活性和多样性。
2. **Composing simple image descriptions using web-scale n-grams (2011)**：该研究采用了n-gram检索与模板填充技术，通过在大规模n-gram语料库中查找相关高频短语和句子组合生成图像描述。这种方法利用了大量的语料库数据，但同样面临着灵活性不足和对模板依赖的问题。

3. **Baby talk: Understanding and generating image descriptions (2012)**：这篇论文使用条件随机场方法结合模板填充技术来生成图像描述。通过设置模板并检测相关图片内容，该方法能够根据检测到的对象和模板生成图像描述句子，确保生成的描述符合语法规则，但缺乏多样性和复杂性。
4. **Im2text: Describing images using 1 million captioned photographs (2012)**：提出了基于大规模图像描述数据库的检索方法。该方法构建了一个包含100万张带有描述的图像数据库，利用全局图片特征匹配和转移方法，从数据库中检索出与查询图像最相似的图片，并将其描述转移到查询图片上生成描述。
5. **Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics (2024)**：这篇论文将图像描述生成任务视为排序任务，使用RankNet模型学习图像和描述之间的排序关系。通过构建包含200万张图像和相应标题的数据集，该方法利用排序模型提高了图像描述生成的准确性和相关性。

深度学习方法

驱动深度学习方法在图像描述生成领域应用的主导因素是视觉编码器架构的发展，特别是注意力机制的推广与应用。早期的研究主要依赖于像MS-COCO这样的image-text pair数据集，随着视觉特征提取能力的提升和深度学习技术的进步，图像描述生成领域也逐渐细化，形成了更加专业的相关数据集。深度学习方法的发展经历了以下几个关键阶段：

1. **Encoder-Decoder架构的引入**：这一阶段的工作奠定了图像描述生成的基础框架，通过将图像编码为特征向量，再通过循环神经网络（RNN）解码生成文本描述。
2. **注意力机制的应用**：注意力机制的引入，使得模型可以关注图像中的不同区域，提升了描述生成的细粒度和准确性。
3. **多模态融合与关系建模**：通过构建视觉关系图和多模态嵌入模型，模型能够更好地理解图像中的复杂关系和语义信息。
4. **Transformer和生成对抗网络（GAN）的应用**：这些新架构的引入，进一步提升了图像描述生成的灵活性和多样性。

主要工作有：

1. **Show and Tell: A Neural Image Caption Generator (2015)**：这篇论文提出了基础的Encoder-Decoder架构。模型首先将图像编码为特征向量，然后通过LSTM解码生成描述文本。该方法解决了图像到文本生成的初步问题，提供了一个基准框架，为后续研究奠定了基础。
2. **Show, Attend and Tell: Neural Image Caption Generation with Visual Attention (2015)**：引入了视觉注意力机制，通过关注图像的不同区域，生成更细粒度和准确的描述文本。这一创新使模型能够更好地理解和生成复杂场景的描述，提高了生成文本的质量和相关性。
3. **Sca-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning (2017)**：进一步改进了注意力机制，不仅关注空间特征，还引入了通道注意力机制。通过同时考虑空间和通道信息，提升了图像特征的表示能力，使得特征提取更加全面和精细。
4. **Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning (2017)**：提出了自适应注意力机制，能够动态决定何时使用视觉信息，何时依赖语言模型。这种机制提高了描述生成的灵活性和准确性，使模型在生成过程中能够根据需要灵活调整注意力。
5. **Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering (2018)**：采用了双重注意力机制，先检测候选区域，再对这些区域进行注意力分配。这种方法显著提升了生成文本的准确性和细节描述能力，使得模型能够更精准地描述图像内容。
6. **Exploring Visual Relationship for Image Captioning (2018)**：在Faster R-CNN基础上，通过构建视觉关系图，将语义和空间关系引入图像特征的表示中。该方法提升了描述的上下文理解能力，使生成的文本能够更好地反映图像中的复杂关系。

7. **Learning to Collocate Neural Modules for Image Captioning (2019)**：提出了模块化的图像描述生成方法，通过多个网络模块并行处理图像特征，提高了模型的灵活性和生成能力。这种模块化设计使得模型能够更好地适应不同类型的图像和生成需求。
8. **Meshed-Memory Transformer for Image Captioning (2020)**：基于Transformer架构，引入了多层次结构和记忆向量，对图像区域及其关系进行编码。该方法提升了描述的层次感和一致性，使生成的文本能够更全面地反映图像内容。
9. **Grit: Faster and Better Image Captioning Transformer Using Dual Visual Features (2022)**：结合了网格特征和区域特征，使用双视觉特征，提升了图像描述生成的速度和质量。通过同时利用两种特征，这种方法在生成效率和文本质量上都有显著提升。
10. **Multimodal Neural Language Models (2014)**：该研究提出了多模态神经语言模型，采用编码器-解码器架构，利用CNN提取图像特征，通过RNN生成文本描述。这种方法为多模态融合和描述生成提供了新的思路。
11. **Deep Fragment Embeddings for Bidirectional Image Sentence Mapping (2014)**：该论文提出了一种基于深度片段嵌入的方法，通过双向映射实现图像和句子之间的对齐。这种方法提高了多模态表示的精度，为图像描述生成任务提供了更强的基础。
12. **DEEP CAPTIONING WITH MULTIMODAL RECURRENT NEURAL NETWORKS (M-RNN) (2015)**：这篇论文引入了多模态RNN，通过将图像和文本特征结合在一起，进一步提高了描述生成的效果。该方法在复杂场景的描述生成中表现出色。
13. **Show, Attend and Tell: Neural Image Caption Generation with Visual Attention (2015)**：该论文提出了注意力机制，让模型更好地关注图像中的关键部分。这一机制显著提高了生成描述的准确性和细节捕捉能力。
14. **Areas of Attention for Image Captioning (2016)**：进一步完善了注意力机制的应用，通过更加细致的关注区域，提高了模型的表现和生成文本的质量。
15. **Towards Diverse and Natural Image Descriptions via a Conditional GAN (2018)**：这篇论文使用生成对抗网络（GAN），生成更加多样化和自然的文本描述。通过条件GAN，该方法能够生成风格多样、自然流畅的图像描述文本。

密集字幕生成

密集字幕生成是指为图像中的每个局部区域生成相应的描述，使得每个区域的细节都能被充分表达。这种方法能够捕捉图像中更多的细节信息，提供更加全面和细致的图像描述。其核心困难是如何在不影响整体理解的前提下，生成大量准确且不重复的区域描述。这涉及到图像区域的精确定位、描述生成的多样性以及局部描述与全局信息的一致性。

解决思路是利用全卷积神经网络（Fully Convolutional Networks, FCN）和带有复制机制的长短期记忆网络（LSTM-C）等技术，通过端到端的训练方法，同时实现图像区域的定位和描述生成。注意力机制和复制机制的引入可以帮助模型更好地关注图像的显著部分，并在生成描述时适当引用已知信息，提升描述的多样性和准确性。

主要工作有：

1. **DenseCap: Fully Convolutional Localization Networks for Dense Captioning (2016)** 提出了使用全卷积神经网络同时定位和生成描述的方法。该方法能够在不需要外部proposal的情况下，高效地为图像中的显著区域生成密集描述。
2. **Incorporating Copying Mechanism in Image Captioning for Learning Novel Objects (2017)** 引入了带有复制机制的LSTM，可以描述训练数据中未出现的新图像。这一机制通过在生成过程中引用训练数据中的部分信息，使得模型能够生成更具多样性和新颖性的描述。

全场景字幕生成

全场景字幕生成是指对整个图像场景进行描述，包括图像中的各个对象、动作、背景等综合信息。这种方法不仅关注图像的局部细节，还强调对整体场景的全面理解和描述。其核心困难是如何在生成描述时，兼顾图像的整体场景和局部细节，同时确保生成的描述连贯且有意义。尤其在面对复杂背景和多样化场景时，生成具有连贯性和全面性的描述具有挑战性。

解决思路是构建强大的视觉和语言模型，通过整合场景中所有相关信息，生成全面的描述。使用信心模型来评估描述的质量，确保描述在捕捉细节的同时保持整体连贯性。快速学习新视觉概念的模型则可以帮助系统在面对新场景时迅速适应和生成高质量描述。

主要工作有：

1. **Rich Image Captioning in the Wild (2018)** 解决了自动描述域外数据的挑战，提出了信心模型评估描述质量。通过对复杂和多样化场景的分析，生成更全面和准确的图像描述。
2. **Learning like a Child: Fast Novel Visual Concept Learning from Sentence Descriptions of Images (2019)** 构建了新的学习模型，快速学习新的视觉概念，提升了全场景描述的效果。该方法通过从图像描述中学习新的视觉概念，使得模型在处理未见过的场景时仍能生成高质量的描述。

基于语义概念的图像描述生成

基于语义概念的图像描述生成是指利用图像中的语义信息，如对象、属性、关系等，生成具有语义意义的描述。这种方法强调将图像中的视觉信息转化为高层次的语义概念，以生成更加有意义和连贯的描述。其核心困难是如何将低层次的视觉特征与高层次的语义概念进行有效映射，并在生成描述时保持语义一致性和准确性。特别是在处理复杂场景和多样化对象时，如何生成具有深层次语义的描述是一个主要挑战。

解决思路是引入外部知识和属性信息，通过构建多模态嵌入模型，将图像中的视觉信息和文本中的语义信息映射到共同的表示空间。条件生成对抗网络（Conditional GAN）等技术的应用，可以生成更加自然和多样化的描述，提升语义表达能力。

主要工作有：

1. **Image Captioning and Visual Question Answering Based on Attributes and External Knowledge (2019)** 提出了基于属性和外部知识生成描述的方法。该方法通过利用图像中的属性信息和外部知识库，提高了生成描述的语义深度和准确性。
2. **Towards Diverse and Natural Image Descriptions via a Conditional GAN (2018)** 使用条件GAN生成更加自然和多样化的文本描述，提升了语义表达能力。通过条件GAN模型，可以生成具有多样性和自然流畅的描述文本，使描述更贴近人类语言习惯和理解。

文本到图像生成

文本到图像生成旨在将自然语言描述转换为视觉图像，涉及多模态学习、生成对抗网络（GAN）、自回归模型和扩散模型等技术，目标是生成高质量、符合描述的图像。相比图像到文本生成任务，文本到图像生成在可控性和组合性两个维度上有其特殊性。自然语言指令反映了用户的意图，这需要模型（主要是文本编码器）能够准确地领会用户的意图，并将相关意图传达到图像生成模块，使其遵照指令生成满足用户需求的图像。

文本到图像生成技术的发展经历了几个主要阶段：XXX

其中，核心工作主要有：

1. **Conditional Generative Adversarial Nets (2014)**：该论文引入了条件生成对抗网络（cGAN），通过附加信息指导生成过程。这种方法为文本到图像生成提供了初步尝试，使得生成过程可以根据输入的文本信息进行调整，提高了生成图像的可控性。
2. **StackGAN (2017)**：StackGAN通过将文本到图像生成分为两个阶段，使用两个cGAN逐步生成高质量图像。第一阶段生成一个低分辨率的图像，第二阶段细化图像，提高了图像的清晰度和细节。该方法解决了图像生成中的清晰度问题，并增强了组合性，因为它能够处理复杂的文本描述，并生成细致的图像细节。
3. **StackGAN++ (2018)**：在StackGAN的基础上，StackGAN++采用并列的树状结构，实现了多尺度的图像生成，并引入了条件增强技术。这种结构使得生成图像的质量进一步提升，同时通过多尺度生成过程，增强了模型的组合性，能够生成更符合复杂文本描述的图像。
4. **AttnGAN (2018)**：AttnGAN引入了注意力机制，通过细化文本到图像生成过程，在图像的不同子区域生成细粒度细节。该方法不仅提高了生成图像的质量和一致性，还增强了模型的可控性，使得用户可以通过文本描述精确控制图像的各个部分。
5. **Neural Discrete Representation Learning (2017)**：该论文提出了VQ-VAE，通过离散表示学习，将图像和文本转换为token，然后使用自回归模型进行生成。这种方法通过引入离散表示，提高了生成图像的稳定性和一致性，增强了组合性，能够处理复杂的文本描述并生成高质量图像。
6. **Taming Transformers (2021)**：在VQ-VAE的基础上，该论文引入了Transformer作为自回归编码器，并结合GAN判别器，实现了高分辨率图像生成。通过使用Transformer，该方法显著提高了生成图像的分辨率和质量，同时保持了模型的可控性和组合性。
7. **Zero-Shot Text-to-Image Generation (2021)**：DALL-E提出了一种通过dVAE将图像压缩为token，然后使用自回归Transformer模型进行文本到图像生成的方法。该方法实现了零样本学习，使得模型在没有直接训练样本的情况下也能生成符合描述的图像，极大地提升了模型的可控性和泛化能力。
8. **Diffusion Models Beat GANs on Image Synthesis (2021)**：该论文引入了扩散模型，通过多步加噪和去噪过程，提高了图像生成的稳定性和质量。扩散模型在生成过程中能够逐步还原图像细节，使得生成结果更为稳定和高质量，增强了模型的组合性。
9. **High-Resolution Image Synthesis with Latent Diffusion Models (2022)**：提出的LDM将扩散过程放在隐空间中进行，实现了高分辨率图像生成。通过在隐空间进行扩散，该方法能够生成更高质量的图像，同时保持了生成过程的可控性和灵活性。

图像区域编辑与风格迁移

图像区域编辑涉及对图像的特定区域进行修改或增强，而不影响图像的其他部分。风格迁移是指将一张风格图像中的颜色和纹理风格迁移到另一张图像上，同时保持内容图像的结构。其核心挑战包括：

- **精度与准确性**：在进行图像区域编辑时，确保边界精确无缝连接是一个主要挑战。边界的模糊或错误可能导致编辑结果不自然，特别是在复杂的背景或纹理变化较大的区域。
- **内容感知**：现代技术常常尝试进行内容感知的区域编辑，以保持场景的连续性。然而，准确理解和保留图像中的语义信息仍然具有挑战性。这在处理复杂场景或不同对象之间交互时尤其困难。
- **纹理和颜色的连续性**：在区域编辑中，确保不同区域之间的纹理和颜色保持连续性是一个常见的难题。无缝过渡对于生成自然和真实的结果至关重要。
- **处理复杂对象和场景**：当编辑涉及复杂对象或多层叠加的场景时，准确识别和编辑特定区域变得更加复杂。这可能涉及对图像深度和遮挡的理解。

主要工作包括：

- **SDSD-Language-Guided Global Image Editing via Cross-Modal Cyclic Mechanism (2021)**：提出了跨模态循环机制，通过语言指导图像的全局编辑，初步实现了基于文本的图像区域编辑。

- **DiffEdit (2023)**：基于扩散模型的方法，通过计算不同文本条件下的噪声估计值差异，生成编辑掩码，引导图像编辑过程。
- **CLIPstyler (2022)**：利用对比语言-图像预训练模型（CLIP）进行图像风格迁移，通过轻量级CNN网络生成与文本条件相关的纹理信息，提升风格迁移效果。
- **StyleGAN-NADA (2021)**：在StyleGAN架构上引入CLIP引导的域适应，通过全局目标损失、局部方向损失和嵌入范数损失，实现图像生成器的域迁移和风格迁移。

文本到视频生成

文本到视频生成在文本到图像生成的基础上增加了时间维度，涉及多模态学习、视频处理和生成对抗网络、自回归模型、扩散模型等技术，目标是生成高质量、符合描述的动态视频。由于视频中涵盖的信息量更大，文生视频任务需要使用有限的信息量实现丰富的生成结果。当前，文本到视频生成领域面临以下挑战：

- **计算难度大**：视频生成需要考虑时长、分辨率等因素，需要进行高维特征融合，计算复杂度高，对算力要求大。
- **数据要求高**：缺乏高质量的文本-视频配对数据集，限制了模型的训练和生成效果。
- **技术实现难度大**：在文生图基础上增加时间维度，要求在视频质量、视频时长等方面实现突破，技术难度较高。

以下是几篇关键论文及其贡献：

1. **Video Generation from Text (2018)**：提出了Text2Filter方法，该方法结合了VAE和GAN技术，首先生成视频的背景，再通过过滤器生成动态细节。尽管应用范围有限，但这项研究为文本到视频生成提供了初步的探索方向，通过分离背景和细节生成，提高了生成视频的质量。
2. **Godiva (2021)**：提出了一种自回归的预测过程，通过训练VQ-VAE，将视频表示为离散token。使用稀疏注意力机制生成视频，显著提升了视频生成的质量和一致性。该方法在保证视频生成质量的同时，优化了计算效率，使得生成过程更为稳定。
3. **Make-a-Video (2022)**：解决了缺乏文本-视频数据和动态内容推断能力的问题。通过使用联合的文本-图像数据和未标记视频数据，该方法提出了高帧率、高质量视频生成的技术。通过创新的数据使用方式，Make-a-Video能够生成高质量的动态视频，解决了传统方法中的数据匮乏问题。
4. **Video Diffusion Models (2022)**：首次将扩散模型应用于文本到视频生成，通过修改2D扩散模型的UNet结构，引入时空卷积和注意力机制，成功实现了视频的高效生成。该方法在保证生成质量的同时，提高了计算效率，是扩散模型在视频生成领域的一次突破性应用。
5. **Align Your Latents (2023)**：基于LDM（Latent Diffusion Model）的方法，通过在图像数据集上预训练，然后在视频数据上进行微调，实现了高分辨率视频的生成。该方法利用预训练模型在图像生成中的优势，通过微调适应视频生成，提升了生成视频的分辨率和质量。
6. **Tune-a-Video (2023)**：提出了一种one-shot微调方法，通过给定一个text-video示例来微调模型，降低了训练开销，同时保证了视频生成的时序一致性。该方法通过高效的微调过程，显著减少了训练时间和计算资源的消耗，提升了生成效率。
7. **Text2Video-Zero (2023)**：引入了zero-shot文本到视频生成，通过动态信息丰富生成帧的隐编码，使用跨帧注意力机制保留前景对象的一致性。该方法解决了大规模文本-视频数据需求的问题，使得在没有大量配对数据的情况下，也能生成高质量的视频。
8. **Video GPT (2021)**：使用自回归模型和注意力机制实现了高质量的视频生成。该方法通过逐帧生成视频内容，并利用Transformer处理长序列数据，保证视频帧之间的连续性。
9. **Make-A-Video (2022)**：通过在文本到图像生成基础上扩展到视频生成，利用时间一致性约束和高效的训练方法，实现了从文本描述生成逼真的视频。

10. **TF-T2V (2024)**：其基本思路是利用大规模未标注的视频进行数据扩充，有效丰富视频的动态多样性，使模型能够学习更丰富的运动信息，从而生成连续稳定且高质量的视频。TF-T2V包含两个分支：一个motion分支利用未标注视频学习运动动态，另一个content分支利用大规模的图像-文本配对数据学习表现信息。