人工智能文生视频大模型Sora的核心技术、运行机理及未来场景

朱光辉[1]　王喜文[2]

（1. 北京理工大学 人文与社会科学学院，北京 100081；2. 北京华夏工联网智能技术研究院，北京 100085）

摘　要：Sora的出现对人工智能的发展具有重大意义，如推动人工智能技术的普及和应用，革新人机交互方式，促进跨学科研究和应用。但同时也应注意到，面对人工智能领域的不断革新，会引发一系列伦理和法律问题。政府有关部门应尽快制定相应的战略、规划、政策和标准，引导新一代人工智能技术更好地服务经济社会发展。

关键词：文生视频大模型；Sora；扩散模型；世界模型；ChatGPT

中图分类号：TP18　　**文献标识码：**A　　**文章编号：**1005-9245（2024）04-0149-08

一、Sora的出现

2024年2月16日，OpenAI发布的人工智能文生视频大模型Sora具有里程碑意义。Sora将扩散模型与ChatGPT所用的大型语言模型相融合，使OpenAI在人工智能视频领域实现了与大型语言模型类似的突破（见图1）。

如图1所示，在Sora中输入提示词（Prompt），在其生成的一分钟视频里，实现了多角度的镜头切换，且物体一致。当其他人工智能视频生成算法还在努力实现4秒连贯性时，Sora已实现60秒的人工智能创作，且一个视频画面中有多角度镜头，主体仍能保证完美的一致性，这在以前是无法想象的。Sora正在教授人工智能理解并模拟运动中的物理世界，训练出既能帮助人们解决需要与现实世界互动的世界模型，又能一次性生成整个视频或延长已生成视频时长。通过为模型提供一次多帧的预见性，Sora解决了主体即使暂时消失在视野中也能保持不变的难题（见图2）。

Sora的出现，意味着大型语言模型向多模态

Prompt: A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.

提示：一位时尚女性走在充满温暖霓虹灯和动画城市标牌的东京街道上。她穿着黑色皮夹克、红色长裙和黑色靴子，拎着黑色钱包。她戴着太阳镜，涂着红色口红。她走路自信又随意。街道潮湿且反光，在彩色灯光的照射下形成镜面效果。许多行人走来走去。

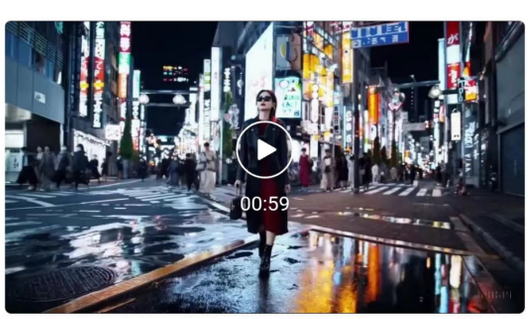

图1　Sora文本生成视频演示图

收稿日期：2024-03-15

作者简介：朱光辉，北京理工大学人文与社会科学学院研究员；王喜文，北京华夏工联网智能技术研究院院长，高级工程师。

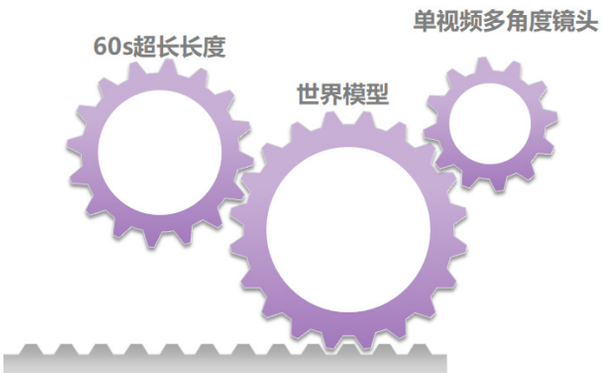

图2　Sora生成视频特征图

升级，能够生产完美视频，其对人工智能整体发展的重大影响主要体现在五个方面。

第一，Sora 能够生成完美视频。这意味着人工智能对人类创造力和创新力的进一步激发。通过文生视频人工智能技术的发展，创意人才和内容创作者将获得更多自由度和灵感来源，有助于其更加轻松地实施自己的想法和创意。文生视频人工智能技术将为创意人才提供更多可能性和工具，推动创意产业的发展与创新。

第二，Sora 的出现将带动人工智能的普及和应用。人工智能文生视频大模型的出现意味着人工智能技术已经能够处理和理解复杂的视觉信息，将进一步推动人工智能技术在实体经济领域的普及和应用，为各行各业提供更加智能和高效的服务，尤其是将深刻改变影视制作和广告营销行业[①]。传统的影视制作和广告营销需要大量人力、物力和财力，文生视频人工智能技术可以帮助企业和创作者降低制作成本、提高效率，并快速生成高质量视频内容。这有望扭转行业格局，使更多小型企业和创业团队进入视频制作和广告营销领域。

第三，Sora 将推动教育和培训领域的变革。通过将文本描述转换为动态视频内容，教育者和培训机构可以提供更加生动、有效的教学内容和教学方式，增强学生的学习兴趣和吸收能力。这将有利于提高教育质量和培训效果，为学生和员工的学习提供更多可能性。此外，文生视频人工智能技术将推动跨学科研究和应用，例如，在视觉与语言的交叉应用领域，为人类提供更加全面深入的理解和认知。

第四，Sora 将引发人机交互的革新。文生视频人工智能技术的成熟将促进人机交互并改善用户体验。通过文生视频人工智能技术，用户可以利用文字描述快速生成个性化的视频内容，实现更加智能化的创作和定制化的体验。这将使人机交互更加自然和智能，提升用户的参与感和满意度。

第五，Sora 将引发有关人工智能的伦理和法律问题。随着技术的不断发展，文生视频人工智能技术可能面临一系列伦理和法律问题，例如，数据隐私保护、知识产权保护，等等。又如，视频内容的真实性、著作权、隐私等，将推动人工智能技术在伦理和法律方面的进一步发展。因此，要加强对上述问题的研究和监管，确保技术的良性发展和社会的可持续发展。

总体而言，Sora 的出现将对社会、产业和个人带来重大影响。通过这项技术的发展，可以预见创意产业的崭新未来、数字化内容的繁荣发展、教育和培训方式的革新以及人机交互的智能化和个性化。但伴随技术的进步，我们要认识到其中蕴含的潜在风险和挑战，并作出积极应对。质言之，文生视频人工智能技术的成熟，将对未来社会和产业产生深远影响。

二、Sora 的运行机理

Sora 是一种将文字描述转换成视频内容的技术，通过深度学习、网络学习、文字描述和视频内容之间的映射关系，生成具有视觉效果的视频。其运行机理包括文本处理、文本编码、视频生成网络、生成器和判别器训练等步骤，通过上述步骤可以实现从文字描述到视频内容的转换。文生视频大模型在虚拟现实、电影制作、游戏开发等领域具有广泛应用前景，可以为用户提供更加生动和沉浸式的体验。其运行机理主要包括六个步骤。

一是文本输入处理。首先，用户输入文字描述，可以是一段文字，也可以是一个或多个关键

① 令小雄、王鼎民、唐铭悦：《ChatGPT到Sora：Sora文生视频大模型对影视创作的机遇、风险及矫治》，《新疆师范大学学报（哲学社会科学版）》，https://kns.cnki.net/kcms2/article/abstract?v=DFdco8SIy0K5Cnm8PN68oQCbipQnOB0ZNME-BNDjHOoFQlnNLErBbXaMZ9I4YyBz-87BqDI15ZKBFTAUIRba68tDD5aMDeQVHvW6GbbvS4XKN-MXHx74IOXmAITVKzH3PvslQb099TY=&uniplatform=NZKPT&language=CHS。

词。这些文字描述是生成视频内容的依据。文本输入处理是文生视频大模型运行的第一步，涉及对输入文本的预处理和分析。其次，模型会对输入文本进行分词，将文本分割成单词或短语。这一步通常使用自然语言处理技术，例如，词性标注和命名实体识别，等等。最后，模型会对分词结果进行词嵌入，将每个单词或短语转换为向量表示，以便后续的文本编码。词嵌入技术通常使用词袋模型或词嵌入模型（例如，Word2Vec 和 GloVe 等）实现。

二是文本编码。输入的文字描述经过文本编码器，将文字转换成向量表示，这种向量是在高维空间中的数学表示，可以捕捉到文字描述的语义和语法信息。文本编码是将处理后的文本转换为计算机可以理解和处理的数字表示。这一步通常使用循环神经网络（Recurrent Neural Network，RNN）或转换器（Transformer）等深度学习技术。RNN 能够对序列数据进行建模，适合处理文本数据。变压器是一种基于自注意力机制的深度学习模型，能够捕捉文本数据中的长距离依赖关系[①]。在文本编码过程中，模型学习文本语法、语义并根据上下文关系，将其转换为向量表示。

三是视频生成网络。文本编码器得到向量表示后，将其输入视频生成网络。视频生成网络通常由生成器和判别器构成，其中，生成器负责生成视频内容，判别器负责评估生成视频的真实性。视频生成网络是文生视频大模型的核心部分，负责将文本编码得到的向量表示转换为视频内容。这一步通常使用生成对抗网络（Generative Adversarial Networks，GAN）或变分自编码器（Variational AutoEncoder，VAE）等深度学习技术。GAN 是一种由生成器和判别器组成的模型，能够生成具有高度竞争力和逼真度的视频内容。VAE 是一种由编码器和解码器组成的模型，能够更好地捕捉视频数据的内在结构。在视频生成网络中，模型会学习视频中的视觉特征和运动规律，并将其转换为视频内容。

四是生成器生成视频。生成器接收文本编码器输出的向量表示，根据向量表示生成对应的视频内容。生成器通常是一个深度神经网络，可以学习不同元素之间的关联，例如，物体的位置、形状、颜色，等等。生成器是视频生成网络中的关键部分，负责根据文本编码得到向量表示生成视频的内容。生成器通常使用 GAN 或 VAE 中的生成器部分实现视频生成。生成器根据输入的文本描述生成相应的视频帧，并将其组合成完整的视频内容。生成器的输出可以是静态图像，也可以是动态视频或三维场景，等等。

五是训练过程。在生成视频的过程中，生成器会不断优化自身参数，使生成的视频内容更贴近真实视频。这一优化过程通过生成器和判别器间的博弈实现，判别器评估生成器生成的视频是否真实，生成器根据反馈调整自身生成的视频内容。训练过程是文生视频大模型的关键步骤，涉及对模型进行训练和学习。训练过程通常包括五个阶段。第一，数据准备阶段，收集大量包含文本描述和对应视频内容的训练数据，便于模型的训练和学习；第二，模型初始化阶段，初始化视频生成网络的参数，包括生成器、判别器或编码器和解码器的权重；第三，训练生成器阶段，通过最小化生成器生成的视频与真实视频之间的差异，训练生成器生成更加逼真的视频内容；第四，训练判别器或编码器和解码器阶段，通过最小化判别器对生成视频和真实视频的区分能力，或最小化编码器和解码器对视频内容的重建误差，训练判别器或编码器和解码器；第五，优化模型阶段，通过调整模型的参数，提高视频生成质量和逼真度。

六是生成视频输出。生成器生成视频内容后，可以将其输出为视频文件，也可以在屏幕上实时展示。用户可以通过这一视频内容观看生成的视频，并将其与输入的文字描述进行对比。生成视频输出是文生视频大模型的最终步骤，涉及将训练好的模型应用于实际任务中，生成符合用户需求和期望的视频内容。具体可分为四个步骤，第一，输入文本描述：用户提供一个文本描述，用于生成视频内容；第二，文本编码：将输入的文本描述进行分词、词嵌入和文本编码，得到文本向量表示；第三，生成视频内容：将文本向量表示输入生成器，生成器根据文本描述生成相应的视频；第四，视频输出：将生成的视频内容输出给用户，供用户观看和评价。

三、Sora 的核心技术

相比大型语言模型，视频模型的生成难点主要集中在六个方面。一是时空复杂性，视频是三维数

① 崔雨萌、王靖亚、闫尚义等：《基于深度学习的警情记录关键信息自动抽取》，《大数据》，2022 年第 6 期。

据，包含时间和空间信息。在生成视频时，不仅要考虑单个帧的内容，而且要考虑帧与帧之间的时序关系和空间关系，这增加了模型理解和生成的复杂性。二是视觉多样性和连续性，真实的视频数据具有较高的视觉多样性，包括不同场景、动作、光照条件，等等。视频模型既要能够处理上述多样性，又要保证生成的视频内容在时间和空间上具有连续性。三是动态范围和细节处理，视频中的动态范围广泛，从静态场景到快速运动的场景都需要模型准确捕捉。同时，视频模型需要生成高分辨率的细节，这对模型的生成能力提出更高要求。四是控制性和交互性，视频模型需要具有一定的可控性，能够根据用户的输入生成相应的视频内容。此外，模型应实现与用户的交互，根据用户反馈进行调整，这要求模型具有较高的智能和适应性。五是数据和计算资源，视频数据通常大于文本数据，这需要更多的数据和计算资源训练视频模型。同时，视频模型需要足够的训练数据学习视频的内在规律和特征。六是真实感与创新性的平衡，视频模型在生成内容时，需要在保证真实性的同时，具有一定的创新性，这需要模型能够理解和模仿现实世界的复杂性，同时，能够创造新的内容。

Sora 有效解决了上述难题和挑战。Sora 与 ChatGPT 幕后的大型语言模型类似，采用 Transformer 架构，解锁了超级扩展性能，并在此基础上融合扩散模型，它通过看似静态噪声的视频开始，并通过多个步骤逐渐去除噪声，最终生成视频。

（一）Transformer

在自然语言处理（Natural Language Processing，NLP）系统中，Transformer 是一种融入注意力机制和神经网络模型领域的主流模型和关键技术。Transformer 具有将所处理的所有文字和句子向量或矢量化、最大限度反映精准意义的能力。从 ChatGPT 的全称“Chat Generative Pre-trained Transformer”可知，其使用的核心技术之一是 Transformer。Transformer 技术是近年来人工智能技术的亮点之一，它是谷歌于 2017 年提出的采用注意力机制的深度学习模型，可以按输入数据各部分重要性的不同，分配不同的权重。Transformer 的精度和性能均优于之前流行的卷积神经网络（Convolutional Neural Networks，CNN）、循环神经网络（RNN）等模型，大幅提升了模型训练的效果，使人工智能得以在更大模型、更多数据、更强算力的基础上进一步增强能力。此外，Transformer 具有较强的跨模态能力，不仅在自然语言理解领域表现优异，而且在语音以及图像方面显示出优异的性能[①]。

（二）Diffusion

扩散（Diffusion）模型的概念最早在 2015 年《利用非均衡热力学的深度非监督学习》（Deep Unsupervised Learning Using Nonequilibrium Thermodynamics）一文中被提出。2020 年，《去噪扩散概率模型》（Denoising Diffusion Probabilistic Models）一文中提出模型可以用来生成图像。从技术角度看，扩散模型是潜在变量（Latent Variable）模型，其通过马尔可夫链（Markov Chain）映射到潜在空间。2022 年，Stable Diffusion 扩散化模型的出现与正式开源，直接推动了生成式人工智能（Artificial Intelligence Generated Content，AIGC）技术的突破性发展。

扩散模型的原理是“先增噪后降噪”。首先，给现有图像逐步施加高斯噪声，直到图像被完全破坏，然后根据给定的高斯噪声，逆向逐步还原图像。当模型训练完成后，随机输入高斯噪声，便能“无中生有”得到一张图像。这样的设计大幅降低了模型训练难度，在逼真的基础上兼具多样性，能够更快、更稳定地生成图片。Sora 选择在大型语言模型基础上融入扩散模型创造文生视频的优势主要表现在三个方面（见图 3）。

一是高逼真度。Sora 擅长直接根据文本指令生成视频，展示了从日常生活到奇幻设置的多样化场景。高逼真度的扩散模型生成的视频画面更自

图3　Sora生成视频优势图

① 王强：《十问ChatGPT：一个新时代正拉开序幕》，《新经济导刊》，2023 年第 1 期。

然、更真实，具有现场感和视觉冲击力，没有明显的模型痕迹，这对于追求真实感的视频生成任务非常重要。

二是高灵活性。Sora 可以生成具有想象力的视频，允许从讲故事到教育内容的广泛创意应用。扩散模型对输入视频的片段长度、分辨率等参数的适应性较强，可适应不同的视频生成需求。与此同时，扩散模型可以灵活地调整扩散系数等参数，以适应不同的视频生成任务，具有较强的自主性。

三是高生成质量。扩散模型在生成视频时，计算复杂度较低、效率较高，可以快速生成大量视频片段。同时，扩散模型可以生成高分辨率、色彩鲜艳且连续的视频片段，在整个持续时间内保持视觉质量，这在传统视频生成模型中并不常见。

四、Sora 的影响分析

（一）Sora 将对多个领域产生影响

一是电影创作。在影视制作行业，Sora 可以帮助电影制作者更好地创建场景和情节，以丰富电影的内容和表现力。此外，Sora 可以为电影制作提供更加真实和自然的角色表现以及更逼真的视觉效果。对影视行业而言，文生视频人工智能技术的成熟将带来革命性的改变。传统的影视制作流程通常需要大量人力、物力、财力，而文生视频人工智能技术可以帮助影视制作公司降低制作成本、提高效率，并快速生成高质量的视频内容。这将促使更多小型制作公司和独立制片人进入影视制作领域，推动影视产业的多元化发展。

二是广告创意。Sora 可以帮助广告制作者更好地表达广告创意，以提供更加生动和吸引人的视觉效果。Sora 还可以帮助广告制作者更好地理解目标受众的需求和喜好，以提供更加个性化的广告体验。对创意产业而言，文生视频人工智能技术的成熟将为创意产业带来革命性变革。创意产业包括广告、设计、文学、艺术等领域，这些领域需要不断创新并以新颖的内容吸引观众和客户。通过文生视频人工智能技术，创意人才和内容创作者将获得更多灵感和工具，可以更加轻松地创作出具有趣味性和高品质的视频内容。这将推动创意产业的发展，激发创意人才的能动性和创造力，进一步丰富数字内容产业的多样性。

三是游戏开发。文生视频人工智能技术可以帮助游戏开发者创建更加真实和自然的游戏场景和角色表现，提供更加沉浸式的游戏体验。此外，文生视频人工智能技术还可以帮助游戏开发者更好地理解游戏玩家的需求和反馈，优化游戏设计和体验。

四是教育和培训。在教育和培训领域，文生视频人工智能技术的成熟将推动教育方式的革新和提升学习体验。通过将文字描述转换为动态视频内容，教育者和培训机构可以为学生提供更加生动、有效的教学内容和教学方式，增强学习者的学习兴趣和吸收能力。这将有助于提高教育质量和培训效果，实现个性化、交互性和深度学习的目标。

值得注意的是，上述影响是否具有颠覆性变革能力或能否给现实行业带来降维打击，取决于诸多因素，包括技术本身的成熟度、应用场景的限制、人类的创造力和想象力，等等。尽管 Sora 可以生成极其逼真的视频，但仍无法完全取代人类的创造力和想象力。因此，我们应以开放和理性的态度看待这项技术的发展，用好它、拥抱它，同时，平衡其潜在风险和利益。

Sora 可能带来某些传统职业的减少或替代某些传统职业，但也会催生新的就业机会，创造新的工种。这种就业变化是科技进步的常态，有助于经济社会持续发展和就业机会转型（见图 4）。

（二）Sora 针对新的职业提供的发展机遇

一是人工智能创作指导人员。这些人员将根据人工智能生成的视频内容进行策划、编辑和优化，确保生成的视频符合特定需求和目标。二是人工智能内容审核和监管人员。伴随人工智能生成视频技术的发展，需要特定人员对生成的视频内容进行审核、监督和管理，确保生成的视频符合伦理、法律和社会规范。三是人工智能故事设计师。这些人员将运用其想象力和创造力，为人工智能生成的视频提供原创的、引人入胜的故事情节和剧本。四是人工智能生成视频数据标注和整理人员。这些人员将负责为训练和优化人工智能生成视频模型提供标注和整理工作，包括对视频中的目标、场景、行为等进行详细的注释和分类。五是人工智能生成视频技术的研发人员。这些人员将致力于不断改进和增强人工智能生成视频技术的性能和功能，包括改进视觉效果、增强模型的创造力和控制性以及提高模型对多样化场景和需求的适应能力，等等。

上述新的职业都需具备人工智能技术的专业知识和创造力。伴随技术的发展和应用场景的拓展，还将涌现更多相关职业。总体而言，尽管 Sora 对传统影视媒体等就业市场造成影响，但同时会创造

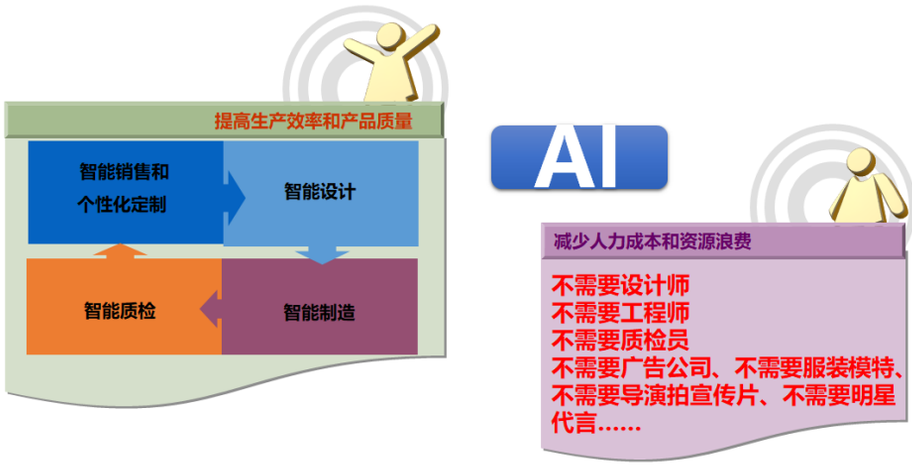

图4　Sora对职业的影响图

新的就业机会和工种。对相关从业人员而言，关键是要持续学习和适应人工智能新技术，以便在未来职场中寻找新的发展机遇。

（三）Sora 对伦理道德产生挑战

一是虚假信息制作与传播。文生视频人工智能技术可能被用于制造虚假信息，包括虚假新闻、虚假广告和虚假政治宣传，甚至可能被滥用以传播虚假信息、不实观点或误导性内容。这些虚假视频可能会误导公众，对社会产生负面影响，破坏信任关系并导致信息混乱，影响社会稳定和民主制度。因此，要制定和落实相关监管政策、技术解决方案和教育措施，以应对虚假信息传播的挑战；同时，要加强对虚假信息的监管和打击，保护公众免受虚假信息的侵害。

二是个人隐私与数据安全。生成视频所需的大量数据可能涉及个人隐私，例如，人脸、身份信息等，这给个人隐私和数据安全带来风险。文生视频人工智能技术还可能被用于制造深度伪造（Deepfakes）视频，伪造他人的身份和言论，侵犯他人的名誉和其他权利。因此，要确保正确的数据使用和保护措施，保护个人隐私，防止数据滥用。

三是著作权与知识产权。使用人工智能生成视频时，可能涉及著作权和知识产权问题。例如，通过生成虚假视频侵犯他人著作权和商标权。因此，要加强对知识产权的保护，建立相应的法律框架和机制，保护原创内容的权益，确保技术创新和创造得到合理的回报和保护，尽可能避免知识产权侵权。

四是社会影响与文化价值。人工智能生成的视频可能对社会和文化产生重大影响，包括透明度、可解释性、公平性和可靠性，等等。应关注和研究人工智能生成视频对社会、文化和媒体的影响，进而更好地应对其可能带来的问题和挑战。

综上所述，人工智能生成视频带来的伦理挑战不可忽视。基于人工智能生成视频的飞速发展，要建立伦理审查和责任追溯机制，确保技术的使用符合伦理规范，并能有效解决潜在的伦理问题。例如，运用法律、技术、政策和教育培训等多种方式，共同应对上述挑战，确保人工智能技术的发展和应用符合经济社会健康发展的整体利益，并在尊重伦理原则的前提下，推动人工智能技术的创新发展。

五、政策建议

算法、算力和数据是人工智能三要素。ChatGPT 和 Sora 的崛起离不开先进算法、规模化算力和规模化数据三要素。打造规模化市场将为我国人工智能产业发展提供满足上述要素的诸多优势。

第一，海量数据资源递增。我国是全球最大的互联网市场之一，拥有庞大的用户基数和丰富的数据资源。2023 年，国家数据局的成立，表明党和政府高度重视数据资源。未来，我国人工智能企业可利用丰富的数据资源训练和优化算法，并为各种应用场景提供更精准的解决方案。这些数据可以用于人脸识别、语音识别、机器翻译、推荐算法等各人工智能领域。

第二，高速网络基础设施建设。我国在 5G 网络基础设施建设方面取得了巨大进展，拥有高速的宽带网络和广泛的覆盖范围，这为人工智能的应用

提供了良好的通信网络基础设施，保证了大规模数据传输和处理方面的高效率和稳定性。这是人工智能算法和模型在实际应用中快速迭代和优化的基础。

第三，集成型技术创新赋能。我国技术创新得到国家和企业的高度重视。我国政府始终致力于人工智能领域的发展，并制定了一系列政策和措施鼓励企业和研究机构进行创新研发。加之企业投资能力强，在人才、资金和技术等方面的支持为人工智能研发和应用的快速发展提供了有力支撑。

第四，多样化应用场景创设。我国市场的多样性为人工智能技术提供了广阔的应用空间。无论制造业、医疗、金融、教育，还是物流、农业等领域，我国都有庞大的市场需求，并提供了大量的实践场景，为人工智能技术的研发和应用提供了宝贵的机会。我国市场变化迅速且多元化，为人工智能算法的改进、产品的创新和应用的推广提供了机遇。

第五，雄厚资本加持。我国政府持续高度关注人工智能领域的发展，通过投资基金、政府补贴和税收优惠等政策为人工智能企业提供了丰富的资本支持。此外，我国风险投资市场较为活跃，许多投资者对人工智能领域寄予厚望，为企业提供了更多融资渠道和资金支持。

第六，国际竞合态势彰显。我国的人工智能产业在国际上积极开展合作与竞争。我国企业从国际先进企业吸纳技术、资本和人才资源，推动了人工智能产业的发展。与此同时，国际企业也将目光投向中国市场，积极与中国企业开展合作，在技术研发、应用创新和市场拓展方面实现共赢。这种国际合作和竞争促进了我国人工智能产业的国际化和标准化，推动整个行业的发展。

综上所述，我国的规模化市场为人工智能产业发展提供了庞大的数据资源，使我国人工智能产业拥有多样化的应用场景、雄厚的资本加持以及广泛的国际合作与竞争机会。上述优势不仅促使我国人工智能产业迅速崛起，为我国在全球人工智能竞争中赢得领先地位，而且促使人工智能技术加快转化为新质生产力，为解决一系列社会、经济和环境问题提供强大的技术支撑和智力支持。目前，在全球人工智能产业加速发展的情况下，我国实现赶超需要从六个方面进行突破。

第一，大模型技术的基础研究与核心技术创新。引爆全球科技圈的 ChatGPT 和 Sora，其背后实际是“大型语言模型”，因此，要实现人工智能产业的赶超，首要任务是加强大模型技术基础研究和核心技术创新。当前，我国在人工智能领域取得了重要研究成果，但与国际先进水平相比仍存在差距。要缩小这一差距，亟须加大在基础研究和核心技术领域的投入，培养更多高水平科研人才，激励科学家进行前沿研究，推动人工智能技术的创新和突破。

第二，复合型跨界人才培养与引进。人工智能领域是高度依赖人才的领域，要实现赶超，就要培养和引进更多人才。我国已建立一批人工智能人才培养项目，但面对快速发展的需求，仍存在巨大的人才缺口，尤其是既懂业务又懂技术的人工智能跨界人才。要吸引人才，就要提供良好的薪酬和福利待遇，创造有利于创新的研发环境，吸引海外优秀人才，同时，积极培养本土人才，加大对各行业人工智能人才的培训力度。

第三，数据资源完善与隐私保护。数据是人工智能的核心驱动力，要实现赶超，就要解决数据资源不足和数据隐私保护等问题。在数据资源方面，要加强数据开放和共享，政府、企业和个人应主动提供数据，并制定相应的数据政策和法规，规范数据的获取、使用和保护。同时，要加强数据隐私保护，加强技术研发，提高数据的安全性和隐私保护能力，保护用户的数据权益，增强用户对人工智能技术的信任和接受度。

第四，产学研结合与产业协同创新。要实现人工智能产业的赶超，就要加强产学研结合和产业协同创新。科技部已出台一系列政策和措施促进产学研结合，鼓励企业与高校和科研机构建立紧密合作关系，提高科研成果的转化和应用能力。同时，要加强产业链的协同创新，加强跨行业和跨部门合作，推动技术、数据和资源共享，形成合力，推动人工智能产业协同发展。

第五，良好政策环境营造与市场机制培育。要实现人工智能产业的赶超，需要提供良好的政策环境和市场机制。政府应制定相关政策和法规，明确支持人工智能产业的发展方向和政策导向，提供相应的资金支持和税收优惠，鼓励企业进行创新研发和技术转化。同时，要加强知识产权保护，提高创新成果的转化和商业化能力。在市场机制方面，要打破垄断，促进公平竞争，鼓励创新创业，形成健康的市场生态环境，激发企业的创新活力和市场竞争力。

第六，国际合作与开放共享。在全球人工智能

产业加速发展的背景下，我国实现赶超需要继续深化国际合作和开放共享。我国可以与国际领先的人工智能企业和研究机构开展合作，共享技术、数据和资源，加速技术的引进和消化吸收。同时，要积极参与国际标准制定，推动人工智能产业的国际化和标准化，提高我国人工智能产业的国际竞争力。

在以 ChatGPT 和 Sora 为代表的新一代人工智能时代到来之际，我国实现人工智能产业的赶超需要加强基础研究和核心技术创新，加强既懂人工智能技术又懂业务的跨界人才的培养和引进力度，完善数据资源和隐私保护，加强产学研结合和产业协同创新，营造良好的政策环境和市场机制，加强国际合作和开放共享。通过努力，我国有望实现人工智能产业的持续发展和赶超。

The Text-to-Video Model Sora: Core Technologies，Operative Mechanism and Application Prospects

ZHU Guang-hui[1]　WANG Xi-wen[2]

（1.School of Humanities and Social Sciences，Beijing Institute of Technology University，Beijing 100081；
2. Beijing Huaxia Institute of Industrial Internet Smart Technologies，Beijing 100085）

Abstract: The emergence of Sora's ability to produce perfect videos has significant implications for the overall development of artificial intelligence. It will drive the popularization and application of AI technology，innovate human-machine interaction methods，promote interdisciplinary research and applications，and also trigger a series of ethical and legal issues. It is urgent for government departments to quickly formulate relevant policies，strategies，plans，and standards to guide the better service of the new generation of AI technology for economic and social development.

Key words: Text-to-video Model；Sora；Diffusion Model；World Model；ChatGPT

[责任编辑：曹晶晶]
[责任校对：王文秋]