视频-BLADE：块稀疏注意力结合步骤蒸馏以实现高效视频生成

Youping Gu[1*] Xiaolong Li[1*] Yuhao Hu[2] Bohan Zhuang[1]

[1]Zhejiang University [2]Central Media Technology Institute, Huawei Technologies
[1] { youpgu71,xiaolong.ziplab, bohan.zhuang } @gmail.com
[2]huyuhao1@h-partners.com

Abstract

扩散变压器目前在高质量视频生成领域处于领先地位，但其缓慢的迭代去噪过程和长序列的高昂二次注意力成本造成了显著的推理瓶颈。虽然步骤蒸馏和稀疏注意机制都已显示出作为独立加速策略的潜力，但有效结合这些方法面临着关键挑战——无训练集成导致次优结果，而在步骤蒸馏后单独训练稀疏注意则需要高昂的高质量视频数据。为克服这些限制，我们提出了 BLADE，一个创新的数据无关联合训练框架，其引入了：(1) 自适应块稀疏注意 (ASA) 机制，用于动态生成内容感知的稀疏性掩码以集中计算在显著的时空特征上，以及 (2) 一个基于轨迹分布匹配 (TDM) 的稀疏性敏感步骤蒸馏范式，其直接将稀疏性纳入蒸馏过程，而非视其为单独的压缩步骤，且具有快速收敛性。我们在如 CogVideoX-5B 和 Wan2.1-1.3B 的文本到视频模型上验证了 BLADE。我们的框架在不同规模上展示了显著的效率提升。在 Wan2.1-1.3B 上，BLADE 实现了比 50 步基线快 14.10 × 的端到端推理加速。此外，在像 CogVideoX-5B 这样的短视频序列长度的模型上，我们的框架提供了强劲的 8.89 × 加速。关键是，这种加速伴随着持续的质量提升。在 VBench-2.0 基准测试中，BLADE 将 CogVideoX-5B 的评分从 0.534 提升至 0.569，将 Wan2.1-1.3B 从 0.563 提升至 0.570，这些结果同时得到了在人类评估中的高评级所进一步证实。我们的代码和模型权重可公开访问：http://ziplab.co/BLADE-Homepage/ 。

1 引言

扩散模型已成为各种生成任务的最新技术，达到了前所未有的图像合成质量，并正推向视频生成这一复杂领域的前沿。通过将生成建模为噪声过程的逐步逆转，这些模型能够生成多样且高保真的内容。然而，对于扩散变压器来说，这种能力伴随着巨大的计算代价。时间维度的引入极大地增加了注意力机制的复杂性，其复杂度随着序列长度以二次方形式增长。再加上去噪过程的迭代性质，导致推理速度极为缓慢，阻碍了实际应用的部署。

为了缓解这一关键的效率瓶颈，两条主要的研究方向已获得显著关注：通过步骤蒸馏 (Song et al., 2023; Salimans & Ho, 2022; Liu et al., 2024; Zheng et al., 2024; Gu et al., 2023; Goodfellow et al., 2014; Yin et al., 2024) 减少推理步骤的数量，以及通过稀疏注意力 (Zhang et al., 2025b; Yuan et al., 2024; Zhang et al., 2025a; Li et al., 2025; Xu et al., 2025; Dao et al., 2022) 降低每步的成本。然而，有效地整合这两种强有力的模式并不容易，并提出了一个关键的难题。一个简单的、无需训练的结合方法，即将预训练的稀疏注意力机制应用于蒸馏模型，产生的结果不是最佳的，因为蒸馏过程与稀疏注意力无关。相反，一个涉及首先进行步骤蒸馏，然后微调模型以实现稀疏性的顺序训练流程同样不切实际，因为这重新引入了对极大且昂贵的高质量视频数据集的需求，抵消了现代无数据蒸馏方法的关键优势 (Gu et al., 2023; Sauer et al., 2024; Luo et al., 2025) 。

在视频领域中，设计合适的稀疏注意力机制的挑战更加严重。许多现有方法依赖于静态的、与内容无关的稀疏模式 (Zhang et al., 2025b; Li et al., 2025; Xi et al., 2025) 。这些固定模

*Equal contribution.

式，例如刚性的局部窗口或预定的跨步，未能适应视频内容的动态和多样化时空特征。因此，它们常常难以保留重要细节和长程依赖关系，导致显著的质量下降，尤其是在需要实现有意义的加速的更高稀疏性水平时。相比之下，另一类研究探索了动态生成的注意力掩码，这些掩码允许稀疏模式适应特定内容的结构。尽管诸如 VSA (Zhang et al., 2025c) 之类的动态掩码方法改善了效率和保真度之间的权衡，它们仅适用于训练环境，并对视频令牌序列的长度有严格要求。另一方面，SpargeAttention (Zhang et al., 2025a) 支持免训练推理，但无法进行训练且表现出有限的稀疏性。这些对使用场景和灵活性的限制阻碍了动态稀疏注意力在真实世界视频生成任务中的广泛使用。

这个领域凸显了对一种稀疏注意机制的明确需求，该机制在计算上是高效的，动态地内容感知，同时足够灵活以支持任意分辨率，并在高稀疏性下不牺牲视觉保真度地支持无需训练和训练感知模式。为此，我们介绍了 ASA，这是一种具有动态令牌选择的无需训练的稀疏注意框架，可以根据输入内容进行适应，同时在各种设置中保持高质量生成。在允许训练的情况下，我们进一步推出了 ASA_GT，这是一种基于蒸馏的变体，通过全局令牌预测来实现端到端训练。ASA 和 ASA_GT 结合在一起，为高效视频生成中的推理和训练场景提供了统一的解决方案。

总体而言，本文认为，一个真正有效的解决方案需要超越将蒸馏和稀疏性视为独立的、事后的优化。我们介绍了 BLADE（BL ock-sparse A ttention Meets step D istillation for E fficient video generation），一种新颖的框架，率先实现了动态稀疏性和步骤蒸馏的协同、无数据和联合训练。我们的方法直接将稀疏性意识融入蒸馏过程，使学生模型能够从教师模型中学习一个简洁高效的轨迹，并基于动态注意模式进行调节。

本研究的主要贡献如下：

- 我们提出了 BLADE，这是一种新颖的无数据联合训练框架，该框架将自适应稀疏注意机制直接结合到一个能感知稀疏性的步骤蒸馏过程中，协同地克服了先前顺序或无训练集成方法的局限性。
- 我们介绍了自适应块稀疏注意力（ASA），这是一种动态、内容感知且硬件友好的注意力机制，它能够即时生成稀疏性掩码，将计算集中在重要特征上，从而优于现有的静态稀疏注意力方法。
- 我们展示了在多样化模型上的显著端到端推理加速，Wan2.1-1.3B 上实现了 14.10 × 的加速，而在较短序列的 CogVideoX-5B 上实现了强劲的 8.89 × 加速。重要的是，这种加速伴随着一致的质量提升，VBench-2.0 的得分在 Wan2.1-1.3B（从 0.563 → 提升至 0.570）和 CogVideoX-5B（从 0.534 → 提升至 0.569）上都有所增加。

2 相关工作

近年来，视频生成领域取得了显著进展，这主要得益于扩散模型的成功。这些模型已成为合成高保真和时间一致性视频内容的事实上的标准，并在各种基准上实现了最先进的成果。

扩散模型的操作原理是学习固定数据损坏过程的逆过程。具体而言，通过使用一个简单的公式 $\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}$ 将干净样本 $\boldsymbol{x}_0 \sim p_{\text{real}}$ 损坏为一个有噪声的样本 $\boldsymbol{x}_t$ ，其中 $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ 是标准高斯噪声。正标量 α_t 和 σ_t 由噪声计划决定，该计划控制在每个时间步长 t (Karras et al., 2022) 的信噪比。

模型的任务是学习这种逆操作。一个网络，通常称为去噪器 $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ ，被训练来从其损坏版本 $\boldsymbol{x}_t$ 预测原始干净样本 $\boldsymbol{x}_0$ 。这个学习得到的去噪器提供了一个评分函数 (Song et al., 2021) 的估计：

$$s_\theta(\boldsymbol{x}_t, t) = \nabla_{\boldsymbol{x}_t} \log p_{\text{real},t}(\boldsymbol{x}_t) \approx -\frac{\boldsymbol{x}_t - \alpha_t \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)}{\sigma_t^2}. \tag{1}$$

然后，通过从纯噪声 $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ 开始，迭代应用学习到的去噪函数来逐步逆转破坏过程，直到获得一个干净样本 $\boldsymbol{x}_0$ ，以此实现生成。

2.1 通过步骤蒸馏加速

步骤蒸馏已成为加速扩散模型 (Song et al., 2023; Salimans & Ho, 2022; Liu et al., 2024; Zheng et al., 2024; Gu et al., 2023; Goodfellow et al., 2014) 的主要策略。其目标是将缓慢的“教师”模型（例如，一个 50 步的采样器）的知识转移到一个更快的“学生”模型，使其

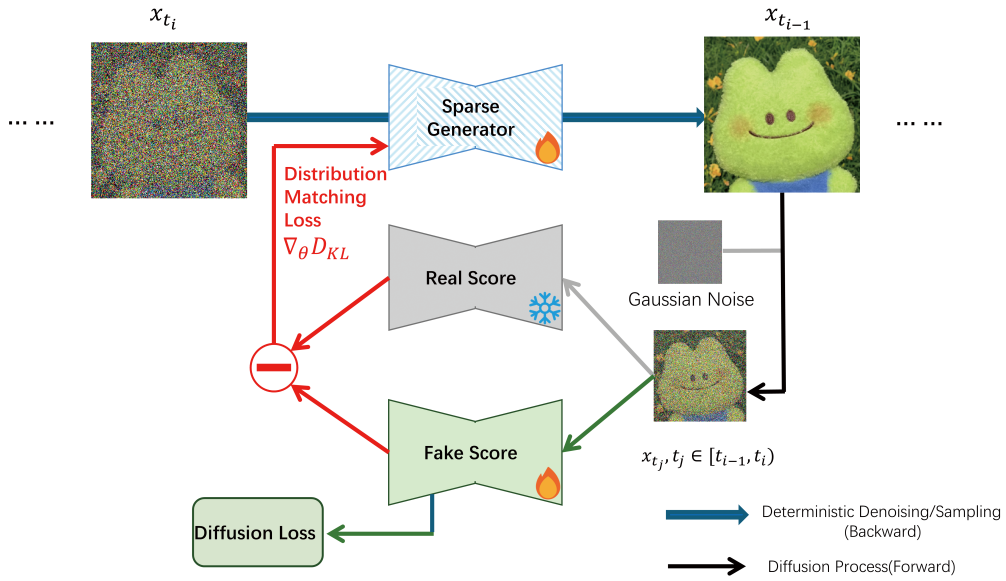


Figure 1: 在单个蒸馏间隔 $[t_{i-1}, t_i)$ 内，Video-BLADE 的训练机制为：稀疏生成器（G_θ）对输入 $\boldsymbol{x}_{t_i}$ 进行去噪以生成样本 $\boldsymbol{x}_{t_{i-1}}$。关键是，这一输出接着被高斯噪声重新破坏以创建中间样本 $\boldsymbol{x}_{t_j}$。一个专门的虚假评分模型对这一再加噪声的样本进行评估。其输出与真实评分模型（即预训练的教师模型）的评分进行对比，以计算分布匹配损失（$\nabla_\theta D_{KL}$）。该损失直接更新学生生成器，迫使其在分布层面上将其生成轨迹与教师模型对齐。

能在很少的步骤（例如，1-8 步）中生成可比的结果。诸如渐进蒸馏 (Salimans & Ho, 2022; Luhman & Luhman, 2021) 之类的早期方法，迭代地减半采样步骤的数量。蒸馏策略可以大致分为输出蒸馏，训练学生去匹配多步教师过程的最终输出，以及轨迹蒸馏 (Luhman & Luhman, 2021; Song et al., 2023)，引导学生追随教师的中间生成路径。轨迹分布匹配（TDM）代表了这一领域最近且复杂的进展 (Luo et al., 2025)。TDM 统一了分布匹配和轨迹匹配的概念。它并不强制严格的实例级轨迹匹配，而是将学生的中间样本的分布与教师在每一步扩散的对应分布对齐。TDM 的一个关键优势是这是一种无需数据的方法；它不需要访问原始的、通常是专有的训练数据集，仅依赖于预训练的教师模型生成指导信号。这使得它成为一个高度实用且多功能的蒸馏框架，我们以此为基础开展工作。

2.2 特定视频的稀疏注意力

已经提出了几种有前景的方法来解决这一挑战，每种方法有不同的机制和取舍。早期的方法如 STA (Zhang et al., 2025b) 和径向注意力 (Li et al., 2025) 主要使用静态注意力掩码。STA 采用固定的局部窗口，这种设计选择使其在特定输入尺寸下最为有效，而径向注意力提出了一种启发式方法，其产生的稀疏性在较短序列上不太明显，限制了其适应性。为了引入更多的动态性，SVG (Xi et al., 2025) 在两个预定义的掩码之间选择，这种二元选择提供了有限的细粒度控制，并可能在质量和稀疏性之间产生取舍。其他方法如 SpargeAttention (Zhang et al., 2025a) 在无训练场景中也显示出潜力。然而，它不适用于训练，其稀疏性水平必须保持适中以保留视频质量。VSA (Zhang et al., 2025c) 引入了训练，并通过固定注意力立方体提供更细致的控制，这种设计影响了可适用分辨率的范围。为了弥合这些多样的取舍，我们提出了自适应块稀疏注意力 (ASA)，这是一种动态的、内容感知的机制，可以即时生成硬件友好的稀疏掩码，为无训练和基于蒸馏的场景提供统一的解决方案。

3 方法

3.1 整体架构

BLADE 是一个通过将动态稀疏性与强大的步骤蒸馏过程相结合来加速视频扩散模型的整体框架。如图 1 所示，我们的架构基于一个学生-教师范式。教师 f_ϕ 是一个预训练的高质量但计算上昂贵的多步扩散模型。学生 G_θ 初始时与教师共享相同的基于 Transformer 的 (DiT) (Peebles & Xie, 2023) 架构和权重。我们设计的关键创新是，为了实现少步骤的生成，用我们提出的自适应块稀疏注意力 (ASA) 机制替换学生模型中的标准自注意力层。训练过程遵循轨迹分布匹配 (TDM) (Luo et al., 2025) 范式。在每次迭代中，稀疏学生模型 G_θ 生成一个中间轨迹。然后，通过无数据的得分蒸馏损失引导该轨迹与教师轨迹的分布匹配。这确保了学生在 ASA 施加的计算限制下学习生成高质量的输出。

3.2 预备知识：轨迹分布匹配（TDM）

轨迹分布匹配 (TDM) (Luo et al., 2025) 是一种先进的蒸馏框架，旨在创建高效的少步扩散模型。其核心思想是将学生模型的整个生成轨迹与教师模型在分布层面对齐，而无需精确的实例层面对齐。这通过一种无数据评分蒸馏过程来实现，该过程依赖于三个关键组件：

1. 预训练的教师模型 f_ϕ ，提供了真实数据分数 s_ϕ 。
2. 学生生成器 G_θ ，能够在几步内生成高保真样本。
3. 伪得分模型 f_ψ ，通过近似学生的难以处理的样本得分提供伪得分 s_ψ 。

训练过程包括两个交织的目标，一个是针对假评分模型的，另一个是针对学生生成器的。

评分蒸馏过程需要学生模型的评分函数 $\nabla_{\boldsymbol{x}_j} \log p_{\theta,j|t_i}(\boldsymbol{x}_j)$ ，这在计算上是棘手的。TDM 通过引入一个假评分模型 f_ψ 来解决这一问题，这是一个同时训练的神经网络，用于逼近学生的评分。为了确保这一逼近是准确的，假评分模型 f_ψ 通过以下去噪目标进行训练：

在这个过程中，先通过学生模型对输入 $\boldsymbol{x}_{t_i}$ 进行去噪来获得干净的目标 $\hat{\boldsymbol{x}}_{t_i}$ 。然后通过扰动这个目标创建一个噪声样本 $\boldsymbol{x}_j$ ，模型学习从这个噪声输入 $\boldsymbol{x}_j$ 预测出干净样本 $\hat{\boldsymbol{x}}_{t_i}$ 。

拥有教师的评分 f_ϕ 和学生自己的评分估计 f_ψ ，学生生成器 G_θ 可以被训练。目标是最小化学生的轨迹分布与教师的轨迹分布之间的 KL 散度。这种对齐是在扩散过程的 K 个阶段中进行的，确保学生能够有效地遵循教师的路径。核心蒸馏损失为：

在实践中，最小化这个 KL 散度是通过匹配评分来实现的。这一目标的梯度通过用假评分模型的输出 s_ψ 替换学生难以处理的真实评分 $\nabla_{x_j} \log p_{\theta,j|t_i}(\boldsymbol{x}_j)$ 来计算，从而得到以下梯度近似：

按照 TDM 框架 (Luo et al., 2025) ，通过两个关键的实现选择，这个过程既实用又节省内存。首先，我们确保蒸馏间隔 $[t_i, t_{i+1})$ 是不重叠的。这种设计允许一个单一的假评分模型 f_ψ 足以应对所有阶段，因为时间步自然地分隔了不同的底层样本分布。其次，为了节省 GPU 内存，反向传播通过学生生成器仅限制在一次 ODE 步骤。

3.3 自适应块稀疏注意力 (ASA)

我们工作的核心设计是自适应块稀疏注意力（ASA）机制，它动态修剪注意力矩阵，以便将计算集中在显著的时空交互上。这种内容感知的方法克服了以往工作中使用的静态掩模的局限性。这个过程包括一个准备步骤，接着是一个动态掩模生成阶段。

预处理：保持局部性令牌重排。输入矩阵 Q 、K 和 V 表示视频令牌的扁平化序列，首先被重组为块。一个关键的初步步骤是重新排列令牌，以保持其固有的空间局部性，这通常会被标准的光栅扫描令牌化扰乱。为此，我们采用 Gilbert 空间填充曲线 (Zhang et al., 2025a) 在分块前重新排序令牌。这确保了生成的块在语义上更具有连贯性，包含空间上连续的信息，这显著增强了后续基于阈值的剪枝效果。

步骤 1：高效的块重要性估计。概念上，可以计算完整的、密集的注意力矩阵 $P = \text{softmax}(QK^\top/\sqrt{d_k})$ ，将其划分为大小为 $b \times b$ 的块，然后对每个块应用最大池化。这将产生一个下采样的重要性矩阵 P_{imp} ，其中每个元素表示对应块的重要性。然后，通过对 P_{imp} 的每一行应用阈值可以生成一个稀疏掩码，使每个查询块仅关注最显著的键值块。然而，完整矩阵 P 的初始计算使得这种方法在实现实际加速时变得不切实际。

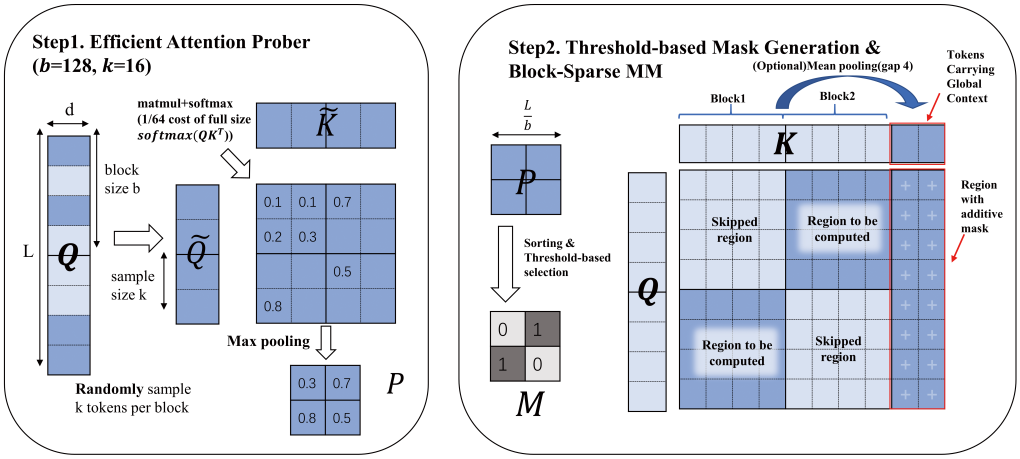


Figure 2: 自适应块稀疏注意力掩码生成的两阶段过程。(1) 高效注意力探针从每个块中采样一些代表性标记（例如，$k = 16$），以计算低成本的最大池化注意力矩阵 P 。(2) 基于阈值的掩码生成器对 P 中的分数进行排序，并选择包含指定阈值（例如，95 %）的顶部块，生成最终的二进制掩码 M 。为了丰富训练的上下文，我们通过将键矩阵 K 与池化版本连接来增强它：$K = \text{Concat}(K, \text{MeanPool}_n(K))$ ，其中 $\text{MeanPool}_n(K)$ 表示对大小为 n 的窗口进行平均池化。在注意力计算期间，原始 K 区域使用二进制块掩码 M ，而池化区域接收固定的加法掩码 $\ln n$ ，软性引导注意力而不破坏稀疏性。

为克服这一限制，我们提出了一种高效的在线近似方法。我们不使用完整的矩阵，而是从 Q 和 K 的每个块中抽取 k 个代表性标记（$k < b$），以形成较小的矩阵 Q_s 和 K_s 。在这些抽取的标记中，我们计算一个更小的低分辨率注意力图 P_{approx} 。由这个近似图导出块重要性矩阵 P_{imp} 。这种方法将掩码生成的复杂度从 $\mathcal{O}(N^2)$ 降低到大约 $\mathcal{O}(N^2 \cdot (k/b)^2)$ ，其中 N 是序列长度。这使得在线掩码生成成为可能。我们的理论结果证明了在适当的 k 和 b 下这种近似的稳定性，表明它产生的掩码可以在显著降低计算成本的同时，紧密地近似基于完整注意力图的标准选择。详细的分析和证明请参见附录 (O'Neill, 2023) 。

步骤 2.1：稀疏掩码构建。一旦获得块重要性矩阵 P_{imp} ，我们就基于阈值掩蔽策略生成最终的稀疏注意力掩码。具体来说，我们将 P_{imp} 的每一行按降序排序，并包括最少数量的关键块，使其累积注意力分数超过指定的阈值（例如，90 %）。这种基于阈值的动态剪枝方法保留了最显著的注意力路径，同时跳过了信息量较少的块，在准确性和效率之间提供了灵活的权衡。

生成的二进制掩码随后用于在训练和推理过程中限制注意力的计算，确保大部分计算资源集中在最相关的交互上。我们在算法 1 中提供了 ASA 的伪代码。

步骤 2.2: 计算。基于这种掩膜生成技术，我们定义了我们的机制的两个变体：

1) 标准 ASA（无训练）：在其主要形式中，生成的二进制稀疏掩码 M 直接与块稀疏注意力内核集成。这个变体可以应用于预训练模型，而无需任何再训练，通过将计算集中在细粒度的显著信息上，实现直接的推理加速。

2) 具有全局 Tokens 的 ASA（用于蒸馏）：为了缓解高稀疏率下潜在的全局信息丢失，我们引入了一种增强变体。我们通过创建一组“全局 tokens”来扩展 Key (K) 和 Value (V) 矩阵。这些全局 tokens 是通过对大小为 n 的窗口进行平均池化生成的，将序列长度减少到原始长度 K 和 V 的 $1/n$ 。增强的矩阵形成如 $K_{\text{aug}} = \text{Concat}(K, \text{MeanPool}_n(K))$ （V 的情况也类似）。在进行注意力计算时，查询与原始 K 区域的交互由二进制稀疏掩码 M 控制，从而保留细粒度的细节。对于增强的“全局 tokens”区域，我们在预 softmax 分数上应用一个固定的加性掩码 $\ln(n)$ 。这种偏置补偿了平均池化的平均效应，确保每个全局 token 都能贡献注意力，好像它代表了其 n 组成部分的全重要性。这种方法柔和地引导每个查询保持对全局上下文的感知，防止在大多数块被剪枝时发生灾难性的信息丢失。

在整篇论文中，我们将标准实现称为 ASA，增强版称为带有全局标记的 ASA（简称 ASA_GT）。

Table 1: 视频质量评估在 VBench-2.0 上。

Model	Method	Sparsity	Total	Creativity	Commonsense	Controllability	Human	Physics	Speedup
CogvideoX-5B	Baseline	-	0.534	0.458	0.523	0.341	<u>0.808</u>	0.539	1 ×
	FA2	-	<u>0.539</u>	0.458	0.498	<u>0.354</u>	0.813	<u>0.570</u>	7.93 ×
	ASA_GT	0.82	0.569	0.546	<u>0.514</u>	0.367	0.802	0.618	8.89 ×
Wan2.1-1.3B	Baseline	-	0.563	<u>0.508</u>	0.549	0.338	0.820	0.600	1 ×
	FA2	-	0.580	0.631	0.485	0.311	0.841	0.631	9.37 ×
	STA	0.74	0.528	0.504	0.471	0.265	<u>0.855</u>	0.543	10.53 ×
	ASA_GT	0.8	<u>0.570</u>	0.472	<u>0.532</u>	<u>0.312</u>	0.918	<u>0.617</u>	14.10 ×

Note: Baseline refers to the official 50 steps baseline. All methods except the Baseline are distilled to 8 steps using TDM.

Algorithm 1 ASA 掩码生成

Require: $Q, K \in \mathbb{R}^{N \times d}$, block size b , sample size k , threshold τ

Rearrange tokens using Gilbert curve
Partition Q, K into $N_b = N/b$ blocks
Randomly sample k tokens from each block to get $Q_s, K_s \in \mathbb{R}^{N_k \times d}$
Compute attention: $\widetilde{P} = \text{softmax}(Q_s K_s^\top / \sqrt{d})$
MaxPool over $k \times k$ blocks to get $P_{\text{imp}} \in \mathbb{R}^{N_b \times N_b}$
for each row i in P_{imp} **do**
    $\tilde{P}_{\text{imp}}(i,j) \leftarrow \dfrac{P_{\text{imp}}(i,j)}{\sum_k P_{\text{imp}}(i,k)}$
    Sort $\tilde{P}_{\text{imp}}[i,:]$ descending $\rightarrow s$
    Find smallest m such that $\sum_{j=1}^{m} s_j \geq \tau$, then clamp m within the range defined by minimum and maximum retention ratios
    Set $M[i,j] = 1$ for top m indices, others $= 0$
end for
return Binary mask M

视频-BLADE 框架的基石之一是稀疏感知蒸馏原则。与先前的方法不同，先前的方法将稀疏性作为训练后的压缩步骤来应用，我们将 ASA 直接整合到 TDM 训练循环中。在每次训练迭代中，学生模型 G_θ 使用 ASA 机制生成其轨迹。然后，分布匹配损失更新学生的权重，以在这些动态稀疏约束条件下提高其输出质量。

这种联合设计强烈地对模型进行正则化，迫使其学习一种稳健的语义表示，通常会产生更优的感知质量。

4 实验

4.1 实验设置

模型。我们在两个文本到视频扩散模型上评估了 BLADE：CogVideoX-5B (Hong et al., 2022) 和 Wan2.1-1.3B (Wan et al., 2025) 。这些模型代表了不同的架构和规模，使我们能够测试我们方法的普适性。

数据集。我们的训练过程以一个包含 10,000 条文本提示的数据集为指导。这些提示从 JourneyDB 基准 (Sun et al., 2023) 中采样，并随后使用 Qwen2.5-3B-Instruct (Team, 2024) 模型进行了质量和多样性上的增强。

指标。我们使用一套标准指标来评估性能：VBench-1.0 (Huang et al., 2024) ， VBench-2.0 (Zheng et al., 2025) ， SSIM & PSNR (Horé & Ziou, 2010) ， 人工评估。

实现细节。除非另有说明，我们使用块大小 $b = 128$ ，每个块的注意力探测器采样 $k = 16$ 个 token。蒸馏通常运行 250-500 次迭代。在 CogVideoX-5B 和 Wan2.1-1.3B 上的实验是在一个由 8 个 A800(80GB) GPU 组成的集群上进行的。

比较方法。ASA_GT、ASA、STA (Zhang et al., 2025b) 和 RaA 分别表示使用我们的自适应注意力，其无需训练的变体，滑动块注意力 (Zhang et al., 2025b) 径向注意力 (Li et al., 2025) 。FA2 指的是 FlashAttention-2 (Dao, 2024) 。

我们的实验表明，Video-BLADE 在不妥协的情况下实现了显著加速，并且经常提高了生成质量。

质量分析。表 1 展示了 CogVideoX-5B 和 Wan2.1-1.3B 在 VBench-2.0 基准测试中的结果，涵盖了多种方法，包括我们提出的 ASA_GT，稀疏基线 STA，FA2，以及 50 步密集基线。

对于 CogVideoX-5B，ASA_GT 在所有主要质量维度上提供了一致且全面的改进。它达到了最高的整体 VBench-2.0 得分（0.569），超过了 50 步基线和 FA2，并在创造性、可控性和物理性方面领先——这些对于生成逼真且吸引人的视频内容至关重要。值得注意的是，ASA_GT 仅通过在一个 17k-token 的短序列上使用 8 步解码，便实现了这一性能，带来了 8.89 × 的速度提升，同时提升了生成质量。这些结果表明，即使在极短的序列长度下，ASA_GT 也能实现稳健的生成质量。

对于 Wan2.1-1.3B，ASA_GT 继续显示出明显的优势。它取得了很高的 VBench-2.0 得分（0.570），最高的人类逼真度（0.918），以及强劲的物理性能，而其运行仅占原始推理时间的 7.09 %（加速 14.10 倍）。相较于具有相似稀疏性的 STA，ASA_GT 在几乎所有指标上都表现得更好。尽管 FA2 在总评分上略胜一筹，但其可控性表现较弱，且计算成本更高。附录中展示了一种画廊风格的视觉比较，展示了不同模型和推理策略的视频样本。

一个引人注目的观察是，尽管 BLADE 具有较高的稀疏性和较少的推理步骤，但它能够超越 50 步密集基线的质量。我们将这种现象归因于我们的联合训练框架所诱发的正则化效应。50 步教师的长迭代轨迹有时可能会积累数值误差或过拟合到嘈杂、不连贯的细节。相比之下，我们的稀疏感知蒸馏迫使学生模型学习更直接和稳定的生成路径（这一原则与过去工作的发现例如 DMD2 (Yin et al., 2024) 相呼应），强制其捕捉最基本的语义，同时隐含地过滤掉教师过程中的“绕道”和噪声。自适应稀疏性通过仅关注最显著的特征进一步帮助了这一点。我们在附录中通过注意力图分析提供了这种效果的视觉证据。因此，生成的模型不仅是一个更快速的近似，而且可以是一个更健壮和连贯的生成器。我们在 VBench-2.0 上评估我们的模型，它更强调语义忠实性——评估生成的视频如何很好地保留高层次意义，而不仅仅是像素级精度。这与我们方法的优势紧密结合。

这些研究结果验证了我们的 ASA_GT 在模型规模和视频长度上的良好泛化性，并通过稀疏感知蒸馏和全局上下文整合在效率和感知质量之间实现了良好的平衡。

效率分析。在内核级别，我们的 ASA 实现相比于 8 步 FA2 基线中使用的标准密集注意力实现了 3.30 × 的加速（22.21 毫秒对比 73.25 毫秒），受益于有效的稀疏率 0.798。该低级别的增益直接转化为显著的端到端加速：我们的基于 ASA 的模型在 24.00 秒内完成生成，而其密集对应版本则需 36.11 秒——实现了 1.504 × 的 E2E 加速。

Table 2: 在 H20 上的 Wan2.1-1.3B 的效率分析。

Metric	FA2-50	FA2-8	ASA-8 (Ours)
Kernel Time (ms)	73.25	73.25	22.21
Kernel Speedup	1.00 ×	1.00 ×	3.30 ×
E2E Time (s)	338.41	36.11	24.00
E2E Speedup	1.00 ×	9.37 ×	14.10 ×

Note: The number suffix (e.g., FA2-50 or FA2-8) indicates the number of inference steps used in each model.

值得注意的是，虽然内核加速超过了 3 × ，但端到端的增益是次线性的。这表明在蒸馏模型中，注意力不再是主要的瓶颈；相反，其他操作（例如，VAE 编码器/解码器和变压器中的非注意力层）开始主导运行时间。这一转变验证了我们针对内核优化以最小化现代扩散管道中注意力开销的有效性。

为了单独评估 ASA 机制的性能，我们在 Wan2.1-1.3B 的无训练推理环境下，将其与其他稀疏注意方法进行比较。表 ?? 显示，在 0.75 的可比稀疏性水平下，ASA 在 PSNR 和 SSIM 方面显著优于 STA 和 SVG，确立了其作为动态注意机制的优越性。不同方法采样的视频展示在图 3 中。更多的消融研究，包括人类评估结果，提供在附录中。

在本文中，我们介绍了 BLADE，这是一个有效解决视频扩散模型中关键效率挑战的新框架。通过协同共同设计一个动态的、内容感知的自适应块稀疏注意（ASA）机制以及一个无数据的轨迹分布匹配（TDM）蒸馏过程，我们的方法在不牺牲生成质量的情况下实现了显著的推理加速。事实上，我们的结果表明，通过使模型在训练过程中感知稀疏性，我们常常

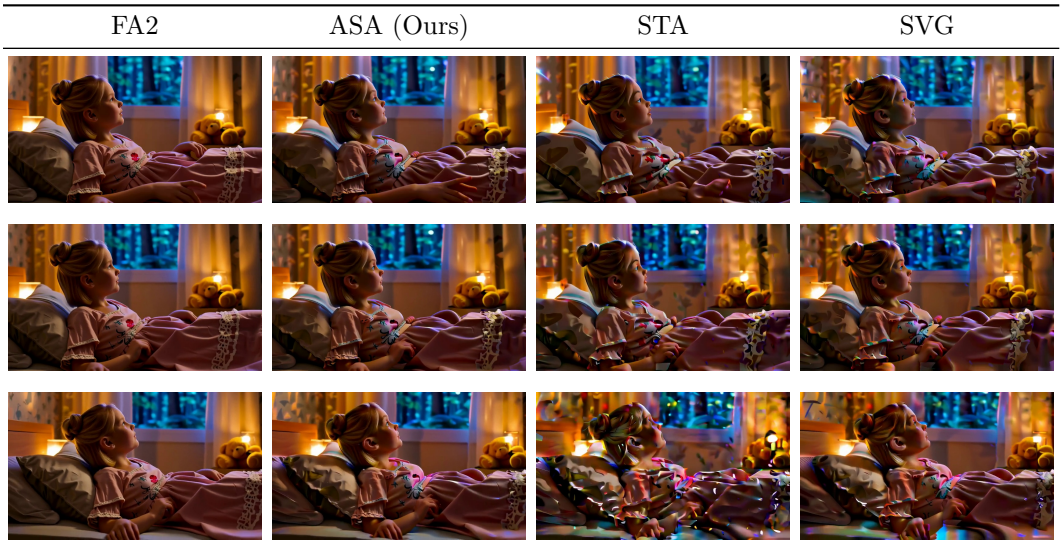


Figure 3: 对于提示“卧室的宁静画面”，比较在第 0，40，80 帧生成的视频。每行显示跨越 4 种方法的相同帧索引。所有视频均使用 8 步采样方法生成。

可以在视觉质量和内在忠实度方面取得比原始多步教师模型和密集蒸馏学生模型更好的结果。

我们的贡献通过对各种视频模型的大量实验得到验证，实验结果显示在内核级效率、端到端推理速度和生成质量方面有显著改善，生成质量通过自动化基准测试（VBench-2.0）和人工评估进行测量。

限制和未来的工作。尽管 Video-BLADE 表现出色，我们也认识到一些限制，这些限制为未来的研究指明了有前景的方向。首先，我们当前的实验仅限于中等长度的视频序列。扩展和验证 ASA 机制以生成时长达到数分钟且包含数十万令牌的视频仍然是下一个重要步骤。此外，我们当前的 ASA 核出于简化目的采用 Triton 实现，这限制了其完全实现理论加速的能力。在未来的工作中，我们计划开发一个更优化的 CUDA 实现，以更好地利用 ASA 的效率潜力。这些方向强调了在更高要求的环境中评估 ASA 的重要性，并探索进一步的架构增强。最后，稀疏感知训练作为一种正则化形式展现了潜力，且可以扩展到视频合成之外的其他生成领域。

5 致谢

本工作由中央媒体技术研究院华为技术通过技术合作支持。

References

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In ICLR, 2024.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness, 2022. URL https://arxiv.org/abs/2205.14135.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Lingjie Liu, and Josh Susskind. Boot: Data-free distillation of denoising diffusion models with bootstrapping, 2023. URL https://arxiv.org/abs/2306.05544.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868, 2022.

Alain Horé and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In 2010 20th International Conference on Pattern Recognition, pp. 2366–2369, 2010. doi: 10.1109/ICPR.2010.579.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. URL https://arxiv.org/abs/2206.00364.

Xingyang Li, Muyang Li, Tianle Cai, Haocheng Xi, Shuo Yang, Yujun Lin, Lvmin Zhang, Songlin Yang, Jinbo Hu, Kelly Peng, Maneesh Agrawala, Ion Stoica, Kurt Keutzer, and Song Han. Radial attention: $o(n \log n)$ sparse attention with energy decay for long video generation, 2025. URL https://arxiv.org/abs/2506.19852.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation, 2024. URL https://arxiv.org/abs/2309.06380.

Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed, 2021. URL https://arxiv.org/abs/2101.02388.

Yihong Luo, Tianyang Hu, Jiacheng Sun, Yujun Cai, and Jing Tang. Learning few-step diffusion models by trajectory distribution matching, 2025. URL https://arxiv.org/abs/2503.06674.

Ben O'Neill. The distribution of order statistics under sampling without replacement, 2023. URL https://arxiv.org/abs/2207.00270.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models, 2022. URL https://arxiv.org/abs/2202.00512.

Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation, 2024. URL https://arxiv.org/abs/2403.12015.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. arXiv preprint arXiv:2303.01469, 2023.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, Limin Wang, and Hongsheng Li. Journeydb: A benchmark for generative image understanding, 2023. URL https://arxiv.org/abs/2307.00716.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314, 2025.

Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, Jianfei Chen, Ion Stoica, Kurt Keutzer, and Song Han. Sparse videogen: Accelerating video diffusion transformers with spatial-temporal sparsity, 2025. URL https://arxiv.org/abs/2502.01776.

Ruyi Xu, Guangxuan Xiao, Haofeng Huang, Junxian Guo, and Song Han. Xattention: Block sparse attention with antidiagonal scoring, 2025. URL https://arxiv.org/abs/2503.16428.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In NeurIPS, 2024.

Zhihang Yuan, Hanling Zhang, Pu Lu, Xuefei Ning, Linfeng Zhang, Tianchen Zhao, Shengen Yan, Guohao Dai, and Yu Wang. Ditfastattn: Attention compression for diffusion transformer models, 2024. URL https://arxiv.org/abs/2406.08552.

Jintao Zhang, Chendong Xiang, Haofeng Huang, Jia Wei, Haocheng Xi, Jun Zhu, and Jianfei Chen. Spargeattention: Accurate and training-free sparse attention accelerating any model inference, 2025a. URL https://arxiv.org/abs/2502.18137.

Peiyuan Zhang, Yongqi Chen, Runlong Su, Hangliang Ding, Ion Stoica, Zhengzhong Liu, and Hao Zhang. Fast video generation with sliding tile attention, 2025b. URL https://arxiv.org/abs/2502.04507.

Peiyuan Zhang, Haofeng Huang, Yongqi Chen, Will Lin, Zhengzhong Liu, Ion Stoica, Eric Xing, and Hao Zhang. Vsa: Faster video diffusion with trainable sparse attention, 2025c. URL https://arxiv.org/abs/2505.13389.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. arXiv preprint arXiv:2503.21755, 2025.

Jianbin Zheng, Minghui Hu, Zhongyi Fan, Chaoyue Wang, Changxing Ding, Dacheng Tao, and Tat-Jen Cham. Trajectory consistency distillation: Improved latent consistency distillation by semi-linear consistency function with trajectory mapping, 2024. URL https://arxiv.org/abs/2402.19159.

Appendix

A 附加实验

我们开展了一系列消融研究，以剖析 Video-BLADE 每个组件的贡献。

A .1 重排策略的影响

Table 3: 使用 VBench-1.0 质量评分评估的标记重排策略的消融结果。

Configuration	Quality Score
Without Rearrange	0.779
With Rearrange (Ours)	0.788

我们验证了 Gilbert 重排策略的重要性。如表 3 所示，使用该策略并通过 ASA 精简的 CogVideoX-5B 模型相比未使用该策略的模型，得到更高的 VBench-1.0 质量评分（0.788 对 0.779），这证实了其在保持空间局部性从而更有效地进行分块剪枝方面的作用。

A .2 ASA_GT 中附加掩码和全局令牌的影响

Table 4: 全局标记（GT）和加性掩码（AM）在 ASA 中对 CogVideoX-5B（VBench-2.0）的影响。

Config	Sparsity (%)	VBench-2.0
ASA	0.8	0.539
ASA-GT	0.82	0.569
ASA-GT_w/o_AM	0.82	0.559
Baseline-50	-	0.534

Note: GT = Global Token, AM = Additive Mask. Baseline-50 is the original 50-step FA2 model.

我们在 VBench-2.0 上进行了消融研究，以验证我们关键设计的有效性：全局标记（GT）和附加掩码（AM）。如表 4 所示，我们的基础模型 ASA 已经超越了基线（0.539 对 0.534）。在整合 GT 后，我们的模型 ASA-GT 的性能显著跃升到 0.569。这个显著的提升突出表明 GT 在聚合全局时空信息中起到了关键作用。此外，从完整模型中移除 AM（即 ASA-GT_w/o_AM）导致性能显著下降到 0.559，这证实了 AM 在稀疏注意力机制下保持模型完整性的必要性。总的来说，这些结果表明 GT 和 AM 都是不可或缺的组件，它们协同作用使得我们最终模型的性能更加优越。

A .3 人工评估结果

Table 5: 人类偏好：8 步模型与 50 步基线。

Comparison		Win	Lose	Tie
CogVideoX-5B				
ASA_GT (Ours) vs.	Baseline	16	10	24
Wan2.1-1.3B				
ASA_GT (Ours) vs.	Baseline	10	12	28
STA vs.	Baseline	0	26	24

我们进行了一个人类偏好研究，以评估我们的高效 8 步 ASA_GT 模型（稀疏率 0.8）和 8 步 STA 模型（稀疏率 0.74）相对于标准 50 步基线的表现。评估使用了 50 个不同的视频提示进行。汇总结果如表 5 所示。

对于 CogVideoX-5B 模型，ASA_GT 在 80 % 的比较中被更倾向选择或评级相等，同时实现了推理时间的 8.89 倍加速。对于 Wan2.1-1.3B，ASA_GT 获得了 56 % 的平局率，总体

上达到了 76 % 的不劣性率，同时将推理时间减少到基线的 7.09 %。相比之下，STA 始终被基线超越，0 次获胜，损失率为 52 %。这些结果强调了 ASA_GT 在尽管积极加速的情况下仍然保持高视觉保真度，验证了其实际部署的有效性。

B P_{imp} 矩阵相似性的证明

我们证明，ASA 的块重要性矩阵与完整注意力块重要性矩阵在比例上是等价的，从而在行归一化后确保获得相同的稀疏掩码。

1: 设置和符号

问题设置：令 $Q, K \in \mathbb{R}^{N\times d}$ 的序列长度为 $N = 32k$ 。分为 $N_b \times N_b$ 块，每块大小为 128×128 。对于块 (i, j) ，定义 $\mathcal{B}_{ij}$ 为完整索引集，并定义 $\mathcal{S}_{ij} \subset \mathcal{B}_{ij}$ 为 $|\mathcal{S}_{ij}| = 16 \times 16 = 256$ 个元素的均匀采样子集。

全注意探测器：

$$E_{st}^{\text{full}} = \exp(Q_s \cdot K_t/\sqrt{d}) \tag{2}$$

$$A_s^{\text{full}} = \sum_{t=1}^{N} E_{st}^{\text{full}} \tag{3}$$

$$P_{st}^{\text{full}} = E_{st}^{\text{full}}/A_s^{\text{full}} \tag{4}$$

ASA 稀疏注意力探测器：

$$E_{st}^{\text{sparse}} = \begin{cases} E_{st}^{\text{full}} & \text{if } (s,t) \in \cup_{i,j}\mathcal{S}_{ij} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$A_s^{\text{sparse}} = \sum_{t:(s,t)\in\cup_j\mathcal{S}_{sj}} E_{st}^{\text{sparse}} \tag{6}$$

$$P_{st}^{\text{sparse}} = E_{st}^{\text{sparse}}/A_s^{\text{sparse}} \tag{7}$$

块重要性矩阵：

$$P_{\text{imp}}^{\text{full}}[i, j] = \max_{(s,t)\in\mathcal{B}_{ij}} P_{st}^{\text{full}} \tag{8}$$

$$P_{\text{imp}}^{\text{sparse}}[i, j] = \max_{(s,t)\in\mathcal{S}_{ij}} P_{st}^{\text{sparse}} \tag{9}$$

2: 理论假设

Assumption 1 (High Quantile Approximation). 对于任何 128×128 注意力模块，高百分位数近似于最大值：

$$\text{percentile}_{99}(\text{block}) \approx \max(\text{block})$$

Assumption 2 (Softmax Scaling Consistency). 对于任何查询 s ，预期的稀疏注意力和按比例缩放：

$$\mathbb{E}[A_s^{\text{sparse}}] = \frac{16}{128}A_s^{\text{full}} = \frac{1}{8}A_s^{\text{full}}$$

3: 顺序统计分析

采样配置：每个块包含 $|\mathcal{B}_{ij}| = 128^2 = 16384$ 个元素，从中我们均匀采样 $|\mathcal{S}_{ij}| = 16^2 = 256$ 个元素。

期望排名：根据不放回的均匀抽样的顺序统计理论，样本最大值在所有 n 个元素中的期望排名为：

$$\mathbb{E}[\text{Rank}] = \frac{n+1}{k+1}$$

对于我们的情况：

$$\mathbb{E}[\text{Rank}] = \frac{16385}{257} \approx 63.74$$

这对应于 $\approx 99.6\%$ 百分位。

方差和置信区间：秩的方差为：

$$\text{Var}[\text{Rank}] = \frac{k(n-k)(n+1)}{(k+1)^2(k+2)} \approx 3970 \tag{10}$$

标准差：$\sigma \approx 63$ 。

置信分析：对于较大的 n, k ，使用正态近似：

- 68 % 置信度：排位 $\in [1, 127]$ （99.2 % -100 % 百分位数）
- 95 % 置信度：等级 $\in [1, 187]$ （98.9 % -100 % 百分位）
- 99 % 置信区间：排名 $\in [1, 226]$ （98.6 % 至 100 % 百分位数）

这表明我们的采样一致地捕获了非常高的百分位数。

4: 主要结果

Theorem 1 (Block Importance Matrix Proportionality). 在假设 1 和 ?? 下：对所有块 (i, j) ，

$$P_{\text{imp}}^{\text{sparse}}[i, j] \approx 8 \times P_{\text{imp}}^{\text{full}}[i, j]$$

成立。

Proof. 步骤 1（局部缩放）：对于 $(s, t) \in \mathcal{S}_{ij}$ ：

$$P_{st}^{\text{sparse}} = \frac{E_{st}^{\text{sparse}}}{A_s^{\text{sparse}}} = \frac{E_{st}^{\text{full}}}{A_s^{\text{sparse}}} \tag{11}$$

$$\overset{\text{Ass. ??}}{\approx} \frac{E_{st}^{\text{full}}}{\frac{1}{8}A_s^{\text{full}}} = 8P_{st}^{\text{full}} \tag{12}$$

步骤 2（样本最大分析）：根据顺序统计，$\max_{(s,t)\in\mathcal{S}_{ij}} P_{st}^{\text{full}}$ 在 $\mathcal{B}_{ij}$ 中所有元素中，期望排名是 ≈ 64 。步骤 3（高分位数近似）：根据假设 1 和我们的置信分析，排名第 16 的值（99.6 % 分位数）$\approx$ 排名第 1 的值：

$$\max_{(s,t)\in\mathcal{S}_{ij}} P_{st}^{\text{full}} \approx \max_{(s,t)\in\mathcal{B}_{ij}} P_{st}^{\text{full}}$$

步骤 4（最终结果）：

$$P_{\text{imp}}^{\text{sparse}}[i, j] = \max_{(s,t)\in\mathcal{S}_{ij}} P_{st}^{\text{sparse}} \approx 8 \max_{(s,t)\in\mathcal{S}_{ij}} P_{st}^{\text{full}} \approx 8 \max_{(s,t)\in\mathcal{B}_{ij}} P_{st}^{\text{full}} = 8P_{\text{imp}}^{\text{full}}[i, j] \tag{13}$$

□

5: 稀疏掩码一致性

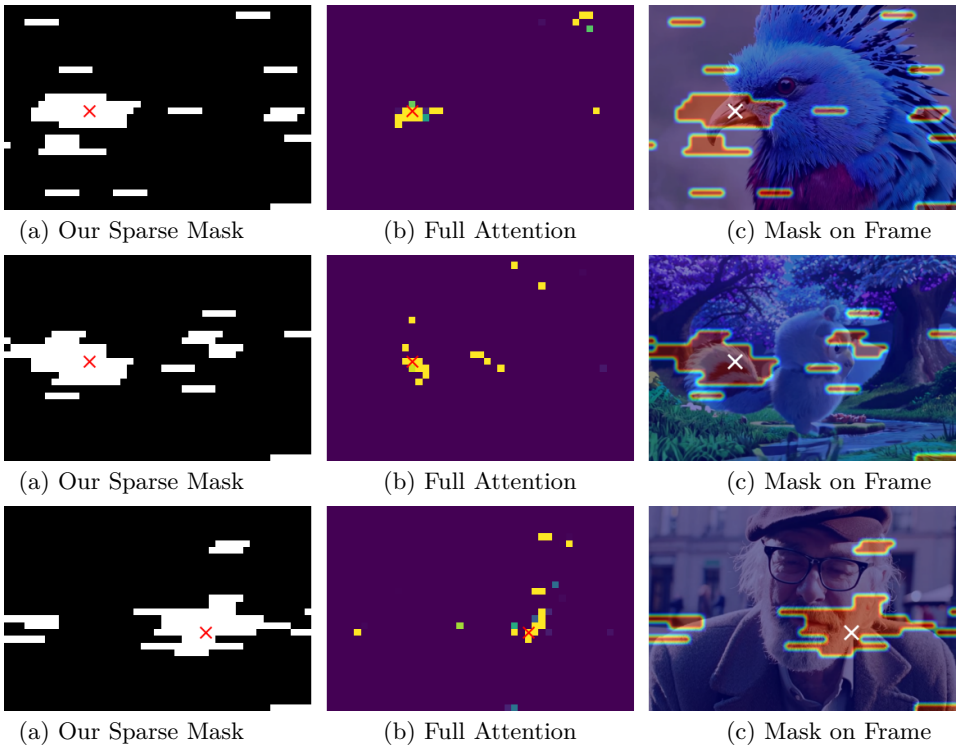
经过行归一化以生成基于阈值的掩码：

$$\frac{P_{\text{imp}}^{\text{sparse}}[i, j]}{\sum_k P_{\text{imp}}^{\text{sparse}}[i, k]} \approx \frac{8P_{\text{imp}}^{\text{full}}[i, j]}{8\sum_k P_{\text{imp}}^{\text{full}}[i, k]} = \frac{P_{\text{imp}}^{\text{full}}[i, j]}{\sum_k P_{\text{imp}}^{\text{full}}[i, k]} \tag{14}$$

由于归一化的块重要性矩阵是相似的，基于阈值的稀疏掩码生成会产生相似的结果，这确立了 ASA 与完整注意计算的一致性。

6: 鲁棒性分析

对高分位数假设的敏感性：即使假设 1 仅近似成立，我们的置信区间显示，以 99 % 的概率，抽样的最大值落在最高 1.4 % 的值范围内，提供了稳健的近似质量。



(a) Our Sparse Mask (b) Full Attention (c) Mask on Frame

(a) Our Sparse Mask (b) Full Attention (c) Mask on Frame

(a) Our Sparse Mask (b) Full Attention (c) Mask on Frame

Figure 4: 注意力掩码的可视化和分析。我们稀疏的方法 (a) 显示能够捕捉到全部注意力 (b) 所识别的最显著区域，有效地聚焦于框架内的关键语义对象 (c)。

C 遮罩可视化

为了阐明我们提出的稀疏注意力方法在视频生成中实现显著加速和增强质量的机制，我们对模型的内部注意力模式进行可视化分析。我们假设，对模型进行受限的计算预算使其不得不忽略低信息、冗余区域（例如静态背景），并更有效地将其注意力集中在场景中的核心语义对象上。

图 4 提供了对该假设的直接且直观的验证。每一行展示了单个样本的注意力行为，将我们的方法与标准的全注意力基线在各种场景下进行比较。具体来说，每个合成图像中的三个子图对应于：

(a) 我们的稀疏掩码：这展示了我们的稀疏方法为单个查询块 Q 生成的注意力掩码。白色区域表示保留用于计算注意力得分的关键块 K 的空间位置。相反，大片的黑色区域是我们的方法修剪掉的位置，在这些位置不会计算注意力，实际上在 softmax 操作之前屏蔽掉了非显著信息。

(b) 完整注意力图：作为基准，此图显示了同一查询补丁在没有稀疏性约束下的原始注意力权重分布。颜色越亮（例如黄色）表示注意力分数越高。

(c) 覆盖在帧上的掩膜：稀疏掩膜作为半透明红色覆盖层突出显示，叠加在实际视频帧上，以直观地显示注意力的空间位置。

在三个不同的场景（鸟类、卡通和老年人）中，可以明显看出，尽管我们的方法修剪了大量计算（如列 a 所示），保留的注意力区域准确覆盖了核心语义对象，例如鸟的喙、卡通角色的尾巴和老年人的胡须。值得注意的是，我们稀疏掩码选择的区域与全注意力图中得分最高的区域高度重叠。这提供了强有力的证据，表明我们的稀疏性策略有效地识别并保留了最显著的语义信息，同时过滤掉多余的背景噪音。这种聚焦机制为模型生成质量的意外提升提供了合理的解释。

D 模型配置细节

Table 6: Wan2.1-1.3B 和 CogVideoX-5B 模型的详细配置参数。

Category	Parameter	Wan2.1-1.3B	CogVideoX-5B
Model Architecture	Model Class	WanTransformer3DModel	CogVideoXTransformer3DModel
	Number of Layers	30	42
	Number of Attention Heads	12	48
	Attention Head Dimension	128	64
	In/Out Channels	16	16
	Temporal Compression Ratio	4	4
	Prediction Dtype	flow	velocity
	Sequence Length	32760	17550
	Text Dimension	4096	4096
	Patch Size	[1,2,2]	[1,2,2]
	Vocab Size	256384	32128
	Number of Timesteps	1000	1000
Training & Inference	Student learning rate	1e-4	1e-4
	Fake model learning rate	5e-4	5e-4
	LoRA Enabled	True	True
	LoRA alpha	64	64
	Optimizer	AdamW	AdamW
	Adam Beta1	0	0
	Adam Beta2	0.95	0.95
	Gradient Clipping	1.0	1.0
	Seed	42	42
	CFG	5	6
	Video Resolution	480 × 832	480 × 720
	Sample FPS	16	8
	Gradient Checkpointing	True	True
	Training Mode	Zero2	Zero2

E 伪代码

受 SeerAttention 中的伪代码启发，我们调整了它以适应我们的实现，从下采样的 Q , K 输入中获取注意力图 P 的最大池化。该过程详细见算法 2 和 3 ：

Algorithm 2 计算块重要性得分

Input: Query Q , Key $K \in \mathbb{R}^{H\times S\times d}$, block size $b = 128$, tokens per block $k = 16$, scale factor s

Output: $P \in \mathbb{R}^{H\times T_r\times T_r}$ where $T_r = \lceil S/b \rceil$

1: Make length divisible by b : $Q_p \leftarrow \text{Pad}(Q, b)$, $K_p \leftarrow \text{Pad}(K, b)$

2: Sample k tokens per block:

3: $\widetilde{Q} \leftarrow \text{BlockSample}(Q_p, b, k)$ $\widetilde{K} \leftarrow \text{BlockSample}(K_p, b, k)$

4: $P \leftarrow \text{GetMaxPooledAttnMap}(\widetilde{Q}, \widetilde{K}, k, s)$

F 视觉比较图库

本节展示了基线模型和我们的 ASA_GT 蒸馏 8 步模型之间的定性比较。对于每个比较，第一行显示来自基线 50 步模型的结果，而第二行显示我们 ASA_GT 方法仅使用 8 步的结果（Wan2.1-1.3B 的稀疏比为 0.8，CogVideoX-5B 为 0.82）。每行展示了从生成的视频序列中采样的 4 帧，展示了不同提示下的时间一致性和视觉质量。

Algorithm 3 获取最大池化注意力图

Input: $Q, K \in \mathbb{R}^{H \times \hat{S} \times d}$, pooling size $\hat{b}$, scale factor s
Output: $A \in \mathbb{R}^{H \times T_r \times T_r}$, $T_r = \lceil \hat{S}/\hat{b} \rceil$

Initialize M as $-\infty$ with shape $(H, \hat{S})$
Initialize ℓ as 0 with shape $(H, \hat{S})$
Initialize R as $-\infty$ with shape $(H, \hat{S}, T_r)$
for each head h do
  Split Q_h K_h into T_r blocks: $Q_1, \ldots, Q_{T_r}$ $K_1, \ldots, K_{T_r}$
  for $i \leftarrow 1$ to T_r do
    $\widetilde{M} \leftarrow M[h, (i-1) * \hat{b} : i * \hat{b}]$
    $\widetilde{\ell} \leftarrow \ell[h, (i-1) * \hat{b} : i * \hat{b}]$
    $\widetilde{R} \leftarrow R[h, (i-1) * \hat{b} : i * \hat{b}, :]$
    for $j \leftarrow 1$ to T_r do
      $s_{ij} \leftarrow Q_i \cdot K_j^\top \cdot s$
      $m_{ij} \leftarrow \text{rowmax}(s_{ij})$, $\tilde{P}_{ij} \leftarrow \exp(s_{ij} - m_{ij})$
      $\tilde{\ell}_{ij} \leftarrow \text{rowsum}(\tilde{P}_{ij})$, $m_{\text{new}} \leftarrow \max(\widetilde{M}, m_{ij})$
      $\widetilde{\ell} \leftarrow e^{\widetilde{M} - m_{\text{new}}} \cdot \widetilde{\ell} + e^{m_{ij} - m_{\text{new}}} \cdot \tilde{\ell}_{ij}$
      $\widetilde{M} \leftarrow m_{\text{new}}$, $\widetilde{R}[:, j] \leftarrow m_{ij}$
    end for
    for $j \leftarrow 1$ to T_r do
      $s_{ij} \leftarrow e^{\widetilde{R}[:,j] - \widetilde{M}}$, $s_{ij} \leftarrow s_{ij} / \widetilde{\ell}$
      $A[h, i, j] \leftarrow \max(s_{ij})$
    end for
  end for
end for