arXiv:2305.03614v5 [cs.CV] 1 Dec 2024

面向连续手语识别的去噪对比对齐

Leming Guo、Wanli Xue*，*IEEE* 会员，以及 Shengyong Chen，*IEEE* 高级会员，

摘要—连续手语识别（CSLR）旨在从未经剪辑的手语视频中识别手语，并将其转化为文本手语词。CSLR 的一项关键挑战是实现视频与手语 gloss 序列之间的有效跨模态对齐，以增强视频表征。然而，当前的跨模态对齐范式往往忽视文本语法在引导视频表征学习全局时序上下文方面的作用，从而对识别性能产生不利影响。为克服这一局限，我们提出了去噪对比对齐（DCA）范式。DCA 创新性地利用文本语法，通过两种互补方式增强视频表示：从判别视角对手语与 gloss 之间的实例对应关系进行建模，并从生成视角对齐其全局上下文。具体而言，DCA 使用对比损失，实现手语与 gloss 之间灵活的实例级对应。在此基础上，DCA 以视频表征为引导，通过从噪声中去噪手语词表征，对齐视频与手语 gloss 序列之间的全局上下文。此外，DCA 引入梯度调制来优化对齐梯度与识别梯度，从而确保更有效的学习过程。通过融合 gloss 层面与全局上下文知识，DCA 显著增强了面向 CSLR 任务的视频表征。在公开基准数据集上的实验结果验证了 DCA 的有效性，并证实了其增强视频表征的可行性。

索引术语—连续手语识别、去噪扩散、全局上下文对齐、实例对应。

一、引言

手语是听障人士之间一种便捷的交流方式，也是一种典型的视觉语言。如果听力正常的人能够理解手语，将有助于推动听障人士积极融入社会，但这是一项艰巨的挑战。连续手语识别（CSLR）研究旨在应对这一挑战。作为人工智能服务社会公益的代表性研究，连续手语识别是指识别手语。

在未剪辑视频中转换为文本手语词[1]，，从而帮助听人理解手语。

目前，连续手语识别通常被建模为典型的跨模态识别任务。因此，需要有效的跨模态对齐范式来增强视频表征 [1]–[4]，，这会显著影响识别

本研究部分得到了国家自然科学基金委员会的资助，项目编号为62376197、62020106004和92048301；部分得到了天津市科学技术计划的资助，项目编号为23JCYBJC00360。L. Guo、W. Xue和S. Chen均就职于天津理工大学计算机科学与工程学院（中国天津市，邮编300384）。*W. Xue为通讯作者（xuewanli@email.tjut.edu.cn）。

[1] 手语 gloss：与手语视频中对应的文本句子里的每个单词。

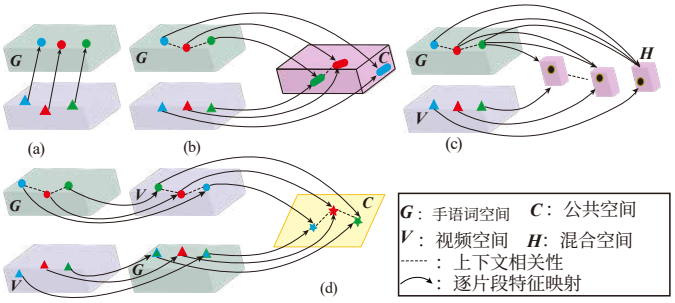


图1. 连续手语识别中研究的跨模态对齐范式。(a) 手语词空间中的视频片段→与单个手语 gloss映射。(b) 两种模态的分布在高维公共潜在空间中彼此接近。(c) 视频片段与前几个时间步的 gloss在多重混合空间中的映射。(d) 两种模态的片段级特征相互映射，并投影到低维公共空间中。

结果。这些对齐范式主要分为以下几类：❶ 使用连接主义时序分类（CTC）将视频表示映射到手语词空间 [5]–[7]（图1(a)）。❷ 使用语言模型学习手语词上下文，并提取手语词表征。随后，将视频表示和手语词表征映射到一个共同的高维空间，以缩小二者的分布差异 [3]，[4]，[8]，[9] （图1(b)）。❸ 将每个时间步的视频表示与所有先前时间步的手语词表征（由预训练语言模型提取）映射到多个高维混合空间 [10]（图1(c)）。

具体而言，对于范式❶，视频表示被映射为与对应手语 gloss 的独热表示相近的表示，因此视频编码器无法学习手语语义来优化自身。尽管范式❷和❸的视频编码器都具有手语语义先验分布，但它们只学习相邻手语词上下文，忽略了用于学习文本语法的全局上下文对齐。因此，它们的视频编码器无法获取全局手语词上下文来优化自身。

受自然语言处理的启发 [11]，我们认为，文本语法指导需要满足以下三个关键要点：*(1)* 建立语言模型，以捕捉手语词序列的文本语法，从而了解每个手语 *gloss* 的语义以及手语词序列的全局上下文。*(2)* 建立手势与 *gloss* 之间的实例对应，引导视频编码器学习手势语义。*(3)* 建立视频序列与手语词序列之间的全局上下文对齐，引导视频编码器学习全局时序上下文。

为了有效学习文本语法，我们提出了一种新颖的**D**去噪-**C**对比**A**对齐（DCA）范式。该范式对手语 gloss 实例对应关系进行建模，从

从对比视角出发，同时从生成视角对齐视频与手语 gloss 序列之间的全局上下文。具体而言，DCA 首先使用语言模型微调来捕捉 gloss 的语义含义以及手语 gloss 序列的全局上下文。随后，DCA 执行对比实例对齐，促使每个手势与其语义上最相关的 gloss 进行匹配，使二者更加接近。这一过程使视频编码器能够学习每个手势的语义。在此基础上，DCA 提出了去噪扩散对齐，促使视频序列表示在低维潜在空间中从噪声生成手语 gloss 序列表示。生成能力被认为是学习 [12], [13] 的最高级体现之一。在生成式去噪过程中，这种对齐使视频编码器能够隐式地将视频序列表示与手语 gloss 序列表示对齐。此外，这种去噪过程有助于视频编码器深入理解生成准确手语 gloss 序列表示所必需的关键因素。这些因素包括 gloss 的顺序和全局上下文知识，它们引导视频编码器重建所有视频片段的全局时序上下文，如图 1(d) 所示。最后，我们引入梯度调制，以优化对齐梯度与识别梯度之间的角度，从而避免优化冲突。

总而言之，本研究着重于文本语法学习，通过解决两个关键问题来增强视频表示：手语与 gloss 之间的实例对应关系学习，以及视频与手语 gloss 序列之间的全局上下文对齐。在公开的连续手语识别基准数据集上的实验结果表明，所提出的 DCA 方法效果显著，取得了最先进的性能，并在全局时序上下文学习方面表现出色。

主要贡献总结如下：

- 不同于连续手语识别中的跨模态对齐范式，我们提出了一种新颖的去噪-对比对齐（DCA）方法。DCA 融合文本语法指导，并将其分解为两个部分：实例对应关系学习和全局上下文对齐，这两者均首次应用于连续手语识别。
- 我们提出了对比实例对齐，以判别式方式建模手语-gloss实例对应关系。此外，我们提出了去噪扩散对齐，根据视频表征从噪声中生成手语词表征，从而对齐视频与手语 gloss 序列之间的全局上下文。这些技术相辅相成，实现了协同学习，增强了局部和全局对齐。
- 实验结果表明，DCA 达到了最先进的识别性能，提升了视频表示的泛化能力，并实现了令人满意的全局时序上下文学习。

II. 相关工作

A. 连续手语识别

连续手语识别旨在识别视频中与多个 gloss 对应的手语，且 gloss 的顺序与手语的顺序一致。由于句子级标注较弱（缺少每个手语的分割真值），且当前连续手语识别基准数据集所提供的数据规模较小，许多最先进的方法 [7], [9], [10] 利用连接时序分类（CTC）[14]，通过最大化视频片段与 gloss 之间所有对齐路径的概率，将每个视频片段准确映射到对应的 gloss。在此基础上，一些方法利用预先提取的姿态热图 [15]或人体关键点 [16], 对运动轨迹进行建模 [6],[17]，或构建帧间图结构 [7]，以进一步增强视频表征。然而，这些方法忽视了 CTC 条件独立性假设带来的问题：该假设只能实现视频片段→与单个 gloss 的映射，缺乏对 gloss 上下文的学习。为解决这一问题，[3],[4], [18] 进一步利用语言模型或 VAE 模型，或提供广泛的手语上下文监督，随后将视觉表征映射到高维手语词空间，并通过约束使其分布接近。与此同时，C^2ST [10] 会将所有先前时间步的手语词表征与当前时间步的视觉表征反复融合，从而构建多个混合空间，将两种模态的上下文注入其中。然而，这些方法仅专注于引导视频片段表征学习手语上下文，却忽视了利用文本语法引导视频片段学习其全局时间上下文对齐。

在本研究中，DCA 利用视频片段之间灵活的对应关系，引导视频编码器学习手语语义。此外，还可以通过基于视频对带噪声的 gloss 序列表征进行去噪，引导视频编码器学习手语词序列的全局上下文。

B. 跨模态对齐

在连续手语识别中，由于基准数据集的标注较弱，几乎所有方法都采用 CTC损失函数，将视觉表示映射到文本手语空间以实现跨模态对齐 [3], [4], [8], [16], [19]。基于 CTC，一些方法还采用交叉注意力操作 [9],、动态时间规整（DTW） [8],和对比学习方法 [3]来提供手语上下文监督。目前先进的跨模态对齐方法包括对比学习方法 [20], [21]以及扩散模型 [22], [23]。具体而言，CLIP [21] 采用对比学习范式来对齐每个词与图像。DiffDis [22] 则利用强大的扩散模型实现卓越的图文对齐。受上述方法启发，我们采用对比学习实现实例级跨模态对齐，并采用去噪扩散范式实现全局跨模态对齐。

C. 去噪扩散模型

去噪扩散模型（DDM）包含一个正向高斯扩散加噪过程和一个反向去噪生成过程，可从高斯噪声开始迭代地优化生成对象。近期研究 [22], [23]指出，DDM 的潜在空间能够呈现跨模态对齐。具体而言，DiffusionRet [23]将文本-视频检索建模为从噪声逐渐生成联合分布的过程。l-DAE [24]将 DDM 转换为

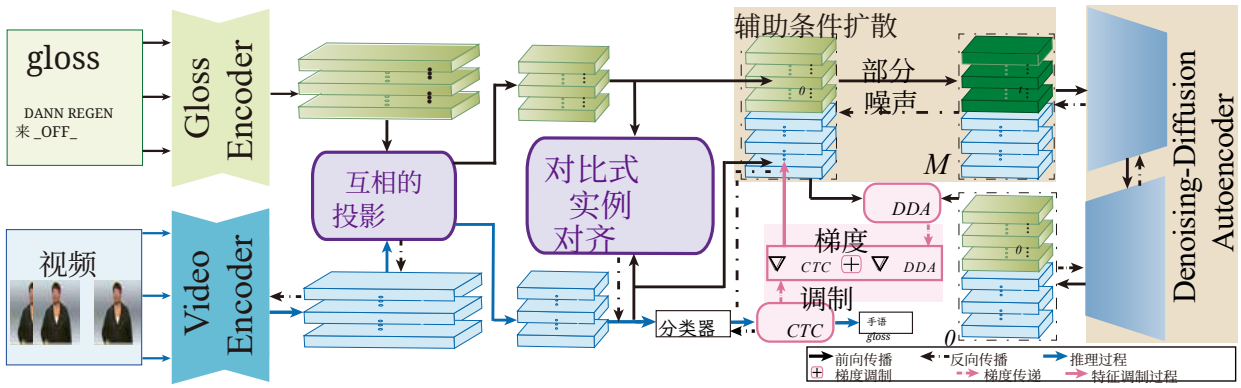


图 2. 所提出的去噪对比对齐（DCA）示意图。首先，我们采用对比实例对齐来建模手语- gloss实例对应关系，从而学习手语语义。该方法鼓励每个手语与语义上最相关的 gloss 匹配。接下来，我们提出去噪扩散对齐，以对齐视频与手语 gloss 序列之间的全局上下文。该技术引导视频序列表示从噪声中重建手语 gloss 序列表示。随后，学习到的实例级和全局手语 gloss 语义将监督视频编码器，帮助完善视频表示的全局时序上下文。最后，我们引入梯度调制，以调整对齐梯度与识别梯度之间的优化角度，从而避免优化冲突。

去噪自编码器，以针对面向识别的表示进行自监督学习，并指出 DDM 的表示能力主要来自去噪过程，而非扩散过程。此外，DIVA [25] 利用文本到图像扩散模型的生成式反馈，提升 CLIP 的细粒度视觉能力。在这项工作中，我们将视频与手语 gloss 序列之间的全局对齐构建为生成式去噪过程，并利用这种对齐来监督视频表征。

III. 方法

本节首先介绍通用的连续手语识别框架。随后，我们在第 III-C 节提出对比实例对齐（CIA），用于建模灵活的手语- gloss实例对应关系，以学习手语语义。之后，我们在第 III-D 节提出去噪扩散对齐（DDA），通过生成式去噪的方式对齐视频与手语 gloss 序列之间的全局上下文。此外，我们在第 III-F 节构建目标函数，以优化 CIA、DDA 和连接时序分类损失函数之间的梯度优化方向，从而调节对齐与识别。最后，在第 III-E 节分析去噪扩散对齐与连续手语识别任务之间的关联。图 2 展示了所提出的去噪对比对齐（DCA）示意图。

A. 通用连续手语识别框架

连续手语识别（CSLR）框架 [3]–[10] 采用由视频编码器、分类模块和跨模态对齐函数组成的范式。**视频编码器**。视频编码器 Φ_{VEnc} 包含空间感知模块和时间感知模块。形式上，给定一个包含 T 帧的手语视频 $\mathcal{X}=\{x_i\}_{i=1}^{T}$，空间感知模块首先从 $\mathcal{X}$ 中提取空间特征 $F_{sp}=\{f_{sp}^{i}\}_{i=1}^{T}$。随后，时间感知模块学习手语特定知识和上下文相关性，以提取视频表示 $F_V=\{f_V^i\}_{i=1}^{T'}\in\mathbb{R}^{(T')\times d}$，这就是视频编码器的输出。此外，$F_V$ 将被输入分类器 Φ_Z 以预测相应的逻辑值 $Z_V=\{z_V^i\}_{i=1}^{T'}$。最后，跨模态对齐函数学习映射 $p(g_i|clip;\theta)$。片段=$\{x_i\}_{i=1}^{T'}$，$T'<T$是视频中的视频片段，θ 表示视频编码器的参数。

连接主义时序分类。手语视频 $\mathcal{X}=\{x_i\}_{i=1}^{T}$ 及其对应的手语词序列 $\mathcal{G}=\{g_i\}_{i=1}^{L}$，采用连接主义时序分类 [14]作为 $\mathcal{H}$，通过对所有可行的对齐路径 π的概率求和，将未分段的视频片段 $\{x_i\}_{i=1}^{T'}$ 映射为 gloss 序列 $\{g^i\}_{i=1}^{L}$：

$$\mathcal{H}=-\log p(\mathcal{G}|x_i;\theta)=-\log\left(\sum_{\pi}p(\pi|x_i;\theta)\right),\quad(1)$$

其中，$p(\pi|x_i;\theta)$ 由连接时序分类计算得出：$p(\pi|x_i;\theta)=\prod p(\pi_i|x_i;\theta)$，概率 $P_\theta=softmax(Z_V)$ 可通过对视频编码器的逻辑值Z_V应用 *Softmax* 函数计算。连接时序分类已被验证能够实现优异的视频片段→单个手语词对齐。

基线方法。值得注意的是，本工作中配备分类器和连接时序分类对齐函数 $\mathcal{H}$的视频编码器构成了基线，这与其他连续手语识别方法 [6], [9], [10]相同。

B. 文本语法捕获

受自然语言处理 [11], [26], [27],的启发，我们将文本语法学习分为两个部分：学习每个手语 gloss 的语义，以及学习手语词序列的全局上下文。为此，我们首先使用手语数据 [3], [16] 对 mBART 模型[2],进行微调，以提取每个手语 gloss 的语义。这个在 CC25 语料库[3]上预训练的 mBART，积累了丰富的句子级全局上下文知识，以及对单个词语的区分性语义知识。经过微调后，mBART 可以将词语的语义迁移到对应的手语 gloss 上（手语 gloss 是自然语言词语的一个子集），同时保留全局上下文学习先验分布。

[2]https://huggingface.co/facebook/mbart-large-cc25
3https://commoncrawl.org/

为进一步学习手语词序列的全局上下文，我们采用类似于 GPGN [4],的掩码语言模型（MLM）方案，将句子中约 50% 的 gloss 遮蔽，以预测其余 gloss。经过微调后优化的 mBART 模型被用作我们的 Gloss 编码器。给定手语 gloss 序列 $\mathcal{G} = \{g^i\}_{i=1}^{L}$，mBART 将提取手语词序列表示 $F_G \in \mathbb{R}^{L\times d}$。

C. 对比实例对齐

视频-手语注释互投影。 当前的连续手语识别方法在高维空间中实现视频片段→gloss映射（如图1(b)和(c)所示）。然而，这种映射方式面临较大的模态偏差，难以有效传递信息 [28]。为解决这一问题，我们提出视频-手语注释互投影，以缩小偏差，促进两种模态之间更顺畅的上下文迁移。给定彼此无法充分捕捉对方语义的视觉序列表示 F_V 和手语词序列表示 F_G，我们首先使用V-L Mapper [18]将每种表示投影到另一种模态的特征空间中。随后，通过两个线性层将两种表示投影到低维潜在空间（如图1(d)所示）。

$$\begin{aligned} F_G &= VL(F_G), F_V = VL(F_V), \\ F_G &= Linear(F_G), F_V = Linear(F_V). \end{aligned} \tag{2}$$

经过上述投影后，视频和 gloss 模态之间的偏差得以减小，并且两种模态的语义可以在低维空间中得到抽象，从而更容易交换上下文和细粒度知识，提高二者的一致性。

对比实例对齐。 为了确保学习手语语义，基于所提出的互投影方法，我们采用对比实例对齐（CIA）来建模灵活的手语-gloss实例对应关系。从数学上讲，给定一对小批次视频表示及其对应的 gloss 表示，我们可以自然地构建视频片段与 gloss 的配对 $\{(F_V, F_G)\}$。随后，CIA 首先计算每个视频片段与 gloss 之间的相似度矩阵， $E^{(i,j)} = F_V \cdot F_{\mathrm{T}_G} \in R^{T'\times L}$。然后计算每一行的 Softmax 分数，对相似度进行重新加权，从而获得手语与 gloss 之间近似实例对应关系 $S^{(i,j)}$。最后，对比实例对齐的形式化定义如下：

$$\mathcal{L}_{CIA} = -\frac{1}{2B}\sum_{i=1}^{B}\log\frac{\exp(S^{(i,j)}/\tau)}{\sum_{j=1}^{B}\exp(S^{(i,j)}/\tau)}, \tag{3}$$

其中， $\tau = 0.07$ 是一个平滑的温度系数， B 表示视频片段与 gloss 的配对。将优化 $\mathcal{L}_{CIA}$ 函数以最小化其值，从而增大手语与 gloss 之间的相似度。因此，视频编码器将被促使为视频片段的表征找到最匹配的手语词表征，并使二者更加接近，从而实现手语与→gloss实例对齐。这种设计通过使 gloss 序列的手语语义得以体现，增强了视频编码器的表征能力。

D. 去噪扩散对齐

辅助条件扩散。经过视频-手语注释互投影后， F_G 将用作提示，并与 F_V 拼接以生成视频-gloss双模态表示 $M = [F_G; F_V] = [f_G^1, ..., f_G^{T'}, f_V^1, ..., f_V^L] \in \mathbb{R}^{(T'+L)\times d}$。其中， T'和 L 分别表示视觉表征和手语序列表征的序列长度。因此，M可视为将 gloss 序列的全局上下文融入近似后验 $q_\phi(M|\mathcal{X}, \mathcal{G})$，其中$\mathcal{X}$、 $\mathcal{G}$ 表示视频与手语 gloss 序列。接下来，我们采用 DDM [29], [30] 前向过程，以马尔可夫链的方式$F_{G0}, ..., F_{GT}$逐步向双模态表示$M_0 = M$中的 gloss 部分添加多级高斯噪声 $F_{G0} = F_G$。从F_{Gt-1} 到F_{Gt}的步骤如下所示：

$$q\left(F_{Gt}|F_{Gt-1}\right) = \mathcal{N}\left(F_{Gt}; \sqrt{1-\beta_t}F_{Gt-1}, \beta_t \mathrm{I}\right). \tag{4}$$

因此，给定 F_{G0}, F_{Gt} ，可以通过经验方式重新定义为：

$$F_{Gt} = \sqrt{\alpha_t}F_{G0} + \sqrt{1-\overline{\alpha}_t}\epsilon, \tag{5}$$

其中， ε 表示高斯噪声，我们通过经验将扩散步数设为包括 T、 $\beta_t \in (0,1)_{t=1}^{t}$、 $\alpha_t := 1-\beta_t$;$\alpha_t := \prod_s \alpha_s$。此外，采用反向噪声调度 [31]。该调度侧重于使用适度的噪声级别，有助于识别表征学习。具体而言，$M_t = [F_{Gt}; F_V]$ 表示与视频表征相连的带噪手语词序列表征。

去噪扩散自编码器。为对视频与手语 gloss 序列之间的全局上下文对齐进行建模，我们采用去噪扩散自编码器 [24] ，根据视频序列表示，从噪声中生成手语词序列表示。该去噪扩散自编码器包含扩散编码器 Φ_{DEnc} 和扩散解码器 Φ_{DDec}。受 l-DAE [24], 框架的启发，编码器和解码器采用相同的架构，且解码器比编码器更深。给定 gloss 部分加噪后的双模态表征 M_t，扩散编码器 $\Phi_{DEnc}(M_t, t)$对加噪的手语词序列表示进行嵌入，并从视频序列表示中学习其上下文，输出潜在特征 F_H。 Φ_{DEnc} 基于视频序列表示，对 M_t 中的噪声部分进行去噪，同时保留初始双模态表征 M_0 与去噪结果之间最相关的信息。随后，扩散解码器 $\Phi_{DDec}(M_t, t)$将 F_H 解码为去噪输出 $M_0 = \Phi_{DDec}(M_t, t)$。该模型经过优化，使可处理的变分下界 $\mathcal{L}_{VLB} \leq \mathbb{E}[-log p_\theta(M_0)]$，能够确保对 M_t 进行有效去噪，并近似还原干净的初始双模态表征 M_0 [29], [30]。$\mathcal{L}_{VLB}$ 是负证据下界：$-KL(q_\theta(M_0|\mathcal{X}, \mathcal{G})||p_\theta(M_0|\mathcal{X})) + E_{q_\theta(M_0|\mathcal{X},\mathcal{G})}[\log p_\theta(\mathcal{G}|M_0, \mathcal{X})] \leq \log p_\theta(\mathcal{G}|\mathcal{X})$，其中 G和X 分别表示手语词序列和视频序列。我们遵循 [29] ，将 $\mathcal{L}_{VLB} \leq \mathbb{E}[-log p_\theta(M_0)]$ 简化为均方误差。简化后的目标定义为 $\mathcal{L}_{DDA}$：

$$\mathcal{L}_{DDA} = \mathbb{E}_{M_0\sim q(M_0), \hat{M}_0, t\sim[1,T]}\left\|M_0 - \hat{M}_0\right\|^2. \tag{6}$$

优化公式 6 时，基于视频序列表示对噪声进行去噪，以生成手语词序列

表示。去噪扩散自编码器隐式地将视频序列表示与手语 gloss 对齐，从而促进联合视频-文本特征空间的学习，这与其他视频-gloss 对齐方法类似。在这一联合特征空间中，自编码器和视频编码器都必须推断手语词序列的全局上下文。通过这些方法，自编码器帮助视频编码器深入理解手语注释顺序，以及定义手语词序列表示的全局上下文。因此，视频编码器经过隐式优化，能够完善其全局时序上下文，使其与手语 gloss 的全局上下文对齐。

E. 去噪扩散对齐与连续手语识别之间的相关性。

形式上，给定一个包含 T 帧的手语视频 $\mathcal{X} = \{x_i\}_{i=1}^{T}$，以及对应的、包含 L 个 gloss 的手语词序列 $\mathcal{G} = \{g_i\}_{i=1}^{L}$，我们引入视频表征 F_V（见第 3.1 节）作为潜变量 z，以帮助建模连续手语识别的条件概率：

$$p\left(\mathcal{G}|\mathcal{X}\right) = \int_{z} p\left(\mathcal{G}, z|\mathcal{X}\right)dz = \int_{z} p\left(\mathcal{G}|z, \mathcal{X}\right)p\left(z|\mathcal{X}\right)dz. \quad (7)$$

在式 7 中，我们发现潜变量 z 是预测 gloss 的关键，它也有助于缩小视频和 gloss 模态之间的偏差。此外，我们尝试从理论上优化真实后验概率 $p_\theta(z|\mathcal{X},\mathcal{G})$，以获得理想的 z。由于真实后验分布 $p_\theta(z|\mathcal{X},\mathcal{G})$ 难以求解，因此采用变分近似后验分布 $q_\theta(z|\mathcal{X},\mathcal{G})$ 来训练连续手语识别模型，并通过证据下界进行优化：$-KL(q_\theta(z|\mathcal{X},\mathcal{G})||p_\theta(z|\mathcal{X}))$ $+ E_{q_\theta(z|\mathcal{X},\mathcal{G})}[\log p_\theta(\mathcal{G}|z,\mathcal{X})] \leq \log p_\theta(\mathcal{G}|\mathcal{X})$。请注意，证据下界优化与第 III-D 节中的负可计算变分下界 $\mathcal{L}_{VLB} \leq \mathbb{E}[-logp_\theta(M_0)]$ 相同，可通过优化去噪扩散自编码器实现。通过优化 $\mathcal{L}_{VLB}$，先验 $p_\theta(z|\mathcal{X})$ 能够有效捕获后验分布 $q_\phi(z|\mathcal{X},\mathcal{G})$ 中的双模态信息，从而增强潜变量 z。

F. 优化

我们方法 $\mathcal{L}_{DCA}$ 的目标包含目标函数 $\mathcal{L}_{DDA}$ 和 $\mathcal{L}_{CIA}$ 以及 $\mathcal{L}_{CTC}$。具体而言，结合 $\mathcal{L}_{DDA}$ 和 $\mathcal{L}_{CIA}$ 可确保视频编码器有效捕获文本 gloss 模态中的手语语义和全局上下文。此外，$\mathcal{L}_{CTC}$连接主义时序分类能够建立多个视频帧与一个手语 gloss 之间的多对一对应关系，从而实现 gloss 分类。总之，我们的目标$\mathcal{L}_{DCA}$ 定义如下：

$$\mathcal{L}_{DCA} = \mathcal{L}_{CTC} + \mathcal{L}_{CIA} + \gamma_1\mathcal{L}_{DDA}, \quad (8)$$

其中，γ_1 是用于平衡贡献的超参数。

梯度调制。在 $\mathcal{L}_{DCA}$ 优化过程中，视频编码器会接收手语-gloss 相关性梯度以及由去噪-扩散自编码器反馈提供的全局上下文知识梯度，因此需要进行调制以优化其与连接时序分类梯度之间的角度。我们认为，视频-手语 gloss 对齐以及全局上下文知识ˆ梯度均来自生成式表示 F_V。受领域自适应任务 [32],的启发，我们调制 F_V ˆ的连接时序分类梯度优化方向，迫使其适应 F_V，具体如下：

$$\begin{aligned} g_o = g_o + g_a - \delta({g_o}^{\mathrm{T}} g_a < 0)\frac{{g_o}^{\mathrm{T}} g_a}{\|g_a\|^2} g_a \\ -\delta({g_a}^{\mathrm{T}} g_o < 0)\frac{{g_a}^{\mathrm{T}} g_o}{\|g_o\|^2} g_o, \end{aligned} \quad (9)$$

其中，g_o 表示 F_V的梯度，g_a 表示 F_V的梯度ˆ。如果 δ 为真，则其值为1，否则为0。

IV. 实验

A. 数据集与评估

PHOENIX-2014 [33]。该基准数据集采集自为天气预报提供手语翻译的公众手语译员，包含由 9 名手语者翻译的 6,842 个句子，手语 gloss 词汇量为 1,295。此外，PHOENIX-2014 数据集官方划分为训练集、开发集和测试集，分别包含 5,672、540 和 629 个视频。

PHOENIX-2014T [34]。该基准广泛用于连续手语识别（CSLR）任务和手语翻译（SLT）任务。它包含 1,085 个用于 CSLR 任务的手语 gloss 词汇，所有视频均由官方划分为训练集、开发集和测试集，分别包含 7,096、519 和 642 个视频。

CSL-Daily [35]。该基准是一个大型中文 CSLR 数据集，适用于连续手语识别（CSLR）任务和手语翻译（SLT）任务。具体而言，CSLR 任务记录了 2000 个手语词汇，官方划分为 18,401 个训练集视频、 1,077 个开发集视频和 1,176 个测试集视频。

评估指标。本文采用词错误率（WER）指标评估连续手语识别（CSLR），并使用权重中存储信息的压缩（IIW）[36]来表示神经网络的泛化误差（即测试集与训练数据集上的准确率差异）。具体而言，WER 属于编辑距离，用于衡量将预测的手语词序列转换为相应的参考手语词序列所需的最少替换（#sub）、删除（#del）和插入（#ins）操作次数。WER 的计算方法如下：

$$WER = \frac{\#sub + \#del + \#ins}{L}, \quad (10)$$

其中，$\#sub$、$\#del$和$\#ins$ 分别表示替换、删除和插入操作的次数。因此，WER 值越低，意味着识别性能越好。对于 IIW，在使用相同训练数据集的情况下，神经网络的 IIW 值越低，意味着该网络对测试数据的泛化能力越强。

B. 实现细节

网络架构。具体而言，在视频编码器中，采用经过 CLIP [21] 预训练的 Resnet50 [37] 作为

表 I 与基线方法在 PHOENIX-2014 上的比较。“GROUP1-4”分别对应图 1(A)、(B)、(C) 和 (D) 所示的跨模态对齐范式。标记为“∗”的条目使用了额外线索（关键点）。

组别	方法	开发集（%）↓		测试集（%）↓	
		删除/插入	词错误率	删除/插入	词错误率
第1组	VAC [19]	7.9/2.5	21.2	8.4/2.6	22.3
	SMKD [42]	6.8/2.5	20.8	6.3/2.3	21.0
	SEN [17]	5.8/2.6	19.5	7.3/4.0	21.0
	TLP [38]	6.3/2.8	19.7	6.1/2.9	20.8
	C^2SLR^* [15]	-	20.5	-	20.4
	SGN [43]	5.1/3.1	19.5	5.3/2.8	20.2
	RadialCTC [5]	6.5/2.7	19.4	6.1/2.6	20.2
	CoSign [44]	-	19.7	-	20.1
	SignBERT+ [45]	4.8/3.7	19.9	4.2/3.8	20.0
	CorrNet [6]	5.6/2.8	18.8	5.7/2.3	19.4
	TwoStream-SLR* [16]	-	18.4	-	18.8
	TB-Net [46]	-	18.9	-	19.6
	SignGraph [7]	-	18.2	-	19.1
Group2	CMA [8]	7.3/2.7	21.3	7.3/2.4	21.9
	CVT-SLR [3]	6.4/2.6	19.8	6.1/2.3	20.1
	GPGN [4]	5.8/3.6	19.9	6.3/2.8	20.4
	CTCA [9]	6.2/2.9	19.5	6.1/2.6	20.1
组3	C^2ST [10]	4.2/3.0	17.5	4.2/3.0	17.7
组4	DCA（我们的方法）	4.5/2.5	**17.3**	4.4/2.8	**17.7**

空间感知模块和时间感知模块在经验上设置为配备双层双向长短期记忆模块的 TLP [38] 。TLP 和双向长短期记忆模块的特征维度均设为 1024。对于去噪扩散自编码器，我们采用手语 gloss 编码器（见第 III-B 节）的双层 Transformer 作为扩散编码器，并采用手语 gloss 编码器的三层 Transformer 作为扩散解码器。浅层可最大限度地减少训练过程中先验分布的损失。SimMIM [39] 和掩码自编码器 [40] 表明，较浅的解码器层足以完成解码。

参数设置。对于辅助条件扩散，按照 DDIM [30]和 [24]，，扩散加噪时间步 T 设为 600，并将 β_t 从 $\beta_1 = 0.0001$ 设为 $\beta_T = 0.99$。协方差 Σ_θ 固定为 $\Sigma_\theta = \beta_t\mathbb{I}$。式 8 中的超参数 γ_1 设为 5。0。

训练与推理过程。我们使用批量大小为 4 对 DCA 进行训练，采用 Adam优化器 [41] ，其中 Resnet50 的初始学习率为 $1e-5$，其他模型的初始学习率为 $1e-4$，权重衰减因子为 $1e-4$，动量设置为 0.9 和 0.99，共训练 80 轮。学习率分别在第 31 轮和第 61 轮衰减（0.2）。超参数 γ_2 设置为 10。所有实验均使用 PyTorch 在两块 A100 GPU 上实现。推理阶段移除去噪扩散自编码器。DCA 的推理过程首先将测试视频 $\mathcal{X}$输入视频编码器，以生成视频表示。随后，将这些视频表示输入分类器，获得分类概率分数。最后，通过 CTC波束搜索对输出进行解码，生成预测的手语词序列。波束宽度设置为 2。

表II PHOENIX-2014T 上与基线方法的对比（%）。标记为“∗”的条目使用了额外线索（关键点）。

方法	WER	
	Dev%	测试集%
V-L Mapper [18]	21.9	22.5
TLP [38]	19.4	21.2
SEN [17]	19.3	20.7
CorrNet [6]	18.9	20.5
C^2SLR^* [15]	20.2	20.4
CVT-SLR [47]	19.4	20.3
CTCA [9]	19.3	20.3
CoSign [44]	19.5	20.1
SignBERT+ [45]	18.8	19.9
TwoStream-SLR* [16]	17.7	19.3
C^2ST [10]	17.3	18.9
DCA（我们的方法）	**17.0**	**18.5**

C. 与当前最先进方法的比较

我们在三个公开基准数据集 PHOENIX-2014 [33], PHOENIX- 2014T [34], 和 CSL-Daily [35] 上，将我们的 DCA 与几种基线方法进行了比较。如表 I、表II、表III所示，我们提出的 DCA 仅对手语视频的 RGB 线索及其对应的手语 gloss 序列进行建模，便在全部三个公开基准数据集上取得了最先进的识别性能。这一结果凸显了文本语法指导在提升视频编码器性能方面的作用。表 I 展示了四种跨模态对齐范式（从组1到组4，分别对应图 1(a) 至图 1(d)）的性能。我们观察到，基于 CTC 的范式（组 1）虽然通过捕捉手语动作 [6], [7], [38], [46]、关键点监督 [15], [16], 或自监督约束 [19], [42], [43], 等策略实现了显著的识别效果，却忽视了手语 gloss 语义和全局上下文学习的重要性。此外，从组 2 转向组 3 时，我们观察到，对手语 gloss 语义进行建模，并采用跨模态对齐方案来指导视频编码器学习手语 gloss 语义，能够有效提升识别性能。然而，其中一些方法的性能甚至不如组 1 中的方法，这可能是由于视频编码器质量方面的局限性所致。这凸显了强大的视频编码器表征对于取得良好识别结果的重要性。具体而言，与基于 CTC 对齐的 TwoStream-SLR [16] 相比，我们的 DCA 在 PHOENIX-2014 的开发集和测试集上分别将词错误率降低了 1.1% 和 1.1%，在 PHOENIX-2014T 的开发集和测试集上分别降低了 0.7% 和 0.8%（见表II）。DCA 与基于 CTC 的方法（组 1）之间的这些性能差距，凸显了手语与 gloss 之间的实例对应关系，以及视频与手语 gloss 序列之间的全局上下文对齐所产生的作用。

此外，我们的 DCA 相较于局部手语词上下文对齐范式（ Group2 和 Group3）也取得了显著的性能提升。例如，C^2ST [10] 将局部手语词上下文注入标签预测中，以建立视频片段与手语 gloss 之间的映射，

表III CSL-DAILY 上与基线方法的比较（%）。标注为“∗”的条目使用了额外线索（关键点）。

方法	Dev%		测试集%	
	删除/插入	WER	删除/插入	WER
BN-TIN+Transf [35]	13.9/3.4	33.6	13.5/3.0	33.1
SEN [17]	-	31.1	-	30.7
SGN [43]	10.8/3.1	30.4	10.7/2.9	30.3
CorrNet [6]	-	30.6	-	30.1
CTCA [9]	9.2/2.5	31.3	8.1/2.3	29.4
TB-Net [46]	-	28.4	-	28.2
CoSign [44]	-	28.1	-	27.2
TwoStream-SLR* [16]	-	**25.4**	-	**25.3**
C^2ST [10]	9.3/2.7	25.9	9.0/2.7	25.8
DCA（我们的方法）	9.1/2.8	25.6	9.0/2.1	**25.3**

随后对视频表征进行细化。值得注意的是，在 PHOENIX-2014 的开发集上，我们的 DCA 性能优于 C^2ST 0.2%；在 PHOENIX-2014T 的开发集和测试集上，词错误率分别优于 C^2 ST 0.3% 和 0.4%（见表II）。此外，在 CSL-DAILY 的开发集和测试集上，DCA 的性能分别优于 C^2ST 0.3% 和 0.5%（见表III）。这些性能差距凸显了全局上下文对齐的有效性：它引导视频编码器根据手语词序列的全局上下文细化其全局时序上下文。

表IV PHOENIX-2014 上 DCA 的消融研究（词错误率%）。

方法	MP	CIA	DDA	开发集 (%) ↓	测试集 (%) ↓
基线	-	-	-	19.5	20.4
	✓	✓	-	18.0 (1.5% ↓)	18.8 (1.6% ↓)
	✓	-	✓	18.0 (1.5% ↓)	18.6 (1.8% ↓)
	-	✓	✓	17.5 (2.0% ↓)	18.2 (2.2% ↓)
	✓	✓	✓	**17.3** (2.2% ↓)	**17.7** (2.7% ↓)

D. 消融研究

对所提出的 DCA 进行消融分析。我们展示了所提出的 $\mathcal{L}_{DCA}$ 的效果。表 IV 对 PHOENIX-2014 基准上的去噪对比对齐（DCA）进行了消融研究。可以看到，基线方法（见 III-A）在开发集和测试集上的词错误率分别为 19.5 % 和 20.4%。值得注意的是，所提出的对比实例对齐（CIA）带来了显著提升，在开发集和测试集上的词错误率分别达到 18.0% 和 18.8%。引入 CIA 使基线模型能够灵活地将每个手势映射到对应的手语 gloss，从而帮助视频编码器学习手语 gloss 语义并优化其表征。

通过加入去噪扩散对齐（DDA），性能进一步提升，在开发集和测试集上的词错误率分别降低了1.5%和1.8%。这些结果凸显了通过生成式去噪过程实现视频与手语 gloss 序列之间全局上下文对齐的优越性。该方法隐式地将视频

表 V PHOENIX-2014 上去噪扩散自编码器的消融研究。“自编码器”和“解码器”分别表示去噪扩散自编码器的编码器和解码器。“完整加噪过程”表示完整的加噪过程。

变体	开发集 (%) ↓	测试集 (%) ↓
DCA（我们的方法）	17.3	17.7
不含解码器的 DCA	18.6（1.3%↑）	18.8 (1.1%↑)
DCA（无自动 EEC）	18.2 (0.9%↑)	18.6（0.9%↑）
DCA（含完整加噪过程）	18.1 (0.8%↑)	18.6（0.9%↑）

序列与手语 gloss 对齐，并帮助视频编码器加深对手语注释顺序和全局上下文的理解。此外，“MP”指第III-C节所述的视频-手语注释互投影。省略“MP”时，我们遵循图 1(b)所示范式，将视频与手语 gloss 序列投影到高维公共潜在空间中。如表IV所示，采用“MP”可显著提升识别性能，在开发集和测试集上的词错误率分别降低0.2%和0.5 %。这表明，将视频特征和手语注释特征分别投影到彼此的特征空间，有助于缩小模态偏差，从而在低维空间中更有效地捕捉两种模态的语义。

表VI PHOENIX-2014 上不同分布对齐方法的性能比较。

方法	对齐	开发集 (%) ↓	测试集 (%) ↓
基线	最大均值差异 [48]	18.6	19.0
	联合最大均值差异 [49]	18.3	18.6
	噪声对比估计 [21]	17.9	18.1
	SimMIM [39]	17.9	17.9
	掩码自编码器 [40]	17.5	17.8
	DCA（我们的方法）	**17.3**	**17.7**

去噪扩散自编码器的消融实验。如表 V 所示，当 DCA 移除扩散解码器（“Auto- Dec”）时，剩余的扩散编码器会经过优化，以学习 DDM 前向过程中每个时间步添加的噪声。这导致识别性能下降：开发集和测试集的词错误率分别增加 1.3% 和 1.1%。此外，当 DCA 移除扩散编码器（“Auto- Enc”）时，剩余的扩散解码器通过$\mathcal{L}_{DCA}$ 进行优化，以近似 M_0，此时性能略有下降（开发集和测试集的词错误率均增加 0.9%）。这两项实验展现了扩散解码器的优势，并验证了去噪驱动过程对于有效表征学习至关重要，这一点已在 l-DAE [24] 中得到证明。此外，“FN”指完整加噪过程，即对双模态表征 M中的视频和手语词序列部分都添加噪声。将部分加噪过程替换为“FN”后，识别性能下降，开发集和测试集的词错误率分别增加 0.8% 和 0.9%。这一结果证明，部分加噪过程有助于自然模态上下文学习，并能促进全局上下文对齐

视频与手语 gloss 序列之间。

不同分布对齐目标的消融实验。我们将所提出的 DCA 与其他旨在实现视频与手语 gloss 序列之间局部或全局上下文对齐的分布对齐方法进行比较，以探究 DCA 的优势。如表VI所示，在考虑文本语法指导的情况下，DCA 的性能显著优于 MMD [48], 联合最大均值差异 [49],噪声对比估计损失 [21],SimMIM [39], 和 MAE [40]。具体而言，虽然噪声对比估计损失能够使分布更接近共同句子，并将不同上下文的分布区分开来，但由于忽略了全局对齐，其性能受到限制。此外，尽管 SimMIM 和 MAE 与 DDM 类似，都是生成方法，但它们无法像采用多级高斯噪声添加的 DDM 那样，同时学习多级掩码比例。此外，SimMIM 和 MAE 都无法很好地建模手语- gloss 实例相关性。这些实验表明，在去噪扩散自编码器中构建全局时间上下文对齐是可行的，并且能够实现良好的对齐能力。

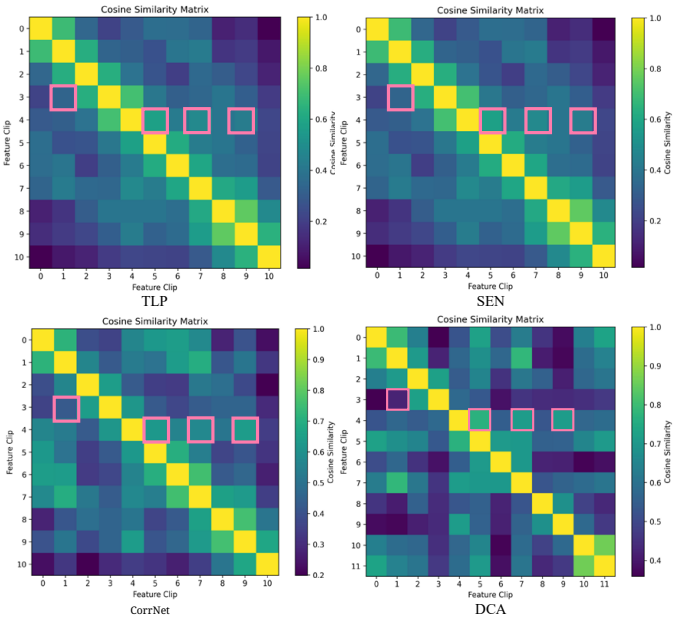


图 3. 在 PHOENIX-2014 测试集上，对 DCA 及其他最先进连续手语识别方法进行全局时序上下文学习的评估。

表VII PHOENIX-2014 上自编码器的不同网络性能比较。

方法	网络	开发集 (%) ↓	测试集 (%) ↓
自编码器	预训练的 mBART	**17.3**	**17.7**
	mBART	17.9	18.2
	双向长短期记忆网络	17.5	17.8
	一维U-Net	17.6	17.9

针对去噪扩散自编码器中不同网络的消融实验。我们研究了不同扩散编码器和解码器网络的影响，所有这些网络都能够自然地获取上下文知识。“预训练 mBART”表示在手语上进行预训练的 mBART语言模型（见第 III-B 节）。“mBART”表示 mBART语言模型 [3], [18]。具体而言，“预训练 mBART”模型具有手语语法的先验分布，而“mBART”模型仅具有自然语言先验知识。如表VII所示，所有网络均取得了优异的性能，表明学习视频与手语 gloss 序列之间全局语义相关性的能力有助于提升 DDA 的有效性。值得注意的是，保留了手语注释序列上下文先验分布的预训练 mBART 模型优于其他模型，突出了更稳健的上下文学习先验能够提升 DDA 的性能。

表VIII
PHOENIX-2014 上时间步 T 因子的消融研究。

时间步 T	1000	800	600	200	100	50
开发集（%）↓	18.1	17.5	**17.3**	17.9	18.6	18.8
测试集 (%) ↓	18.3	17.9	**17.7**	18.2	18.8	19.4

扩散加噪时间步消融实验 T。本节逐步减小加噪过程的扩散步数（从 1000 减至 50）。当时间步 T 较小时，添加到手语词序列表征中的噪声较少，反之亦然。噪声步数较少时，去噪过程会更容易，但扩散解码器学到的手语词上下文知识也会减少。表 VIII 显示，最优时间步为 600。值得注意的是，即使时间步 T 为 50，模型仍能取得出色的性能。这表明，在去噪过程中仅去除少量噪声时，视频表示仍能与手语词序列有效对齐。手语词上下文知识的有限反馈也有利于视频编码器。此外，对低级潜在表征施加噪声似乎不需要大量噪声。相反，当时间步 T 过大时，扩散解码器难以根据视频推断手语词序列表征。

E. 其他评估

全局时序上下文学习的评估。在本节中，我们旨在评估 DCA 的全局时序上下文学习能力。具体而言，我们首先从 Phoenix-2014 测试集中的两个随机视频中提取视频序列表示。随后，我们选择一个窗口大小，以近似视频序列中视频片段相对于 gloss 的长度。例如，如果一个视频包含 55 个视频片段，对应 16 个 gloss，则窗口大小设置为 (55 mod 16 =3)。接下来，我们计算裁剪后的视频片段表示之间的余弦相似度。这样，每个视频片段便可以大致对应一个 gloss（其中“***”表示无法识别的 gloss），而余弦相似度结果则反映了全局时序上下文的相关性。如图 3 所示，我们的 DCA 能够有效地为相似的 gloss 建立较高的相似度，为不同的 gloss 产生更大的相似度差异，并且几乎实现了与 gloss 的正确对应。具体而言，片段 4、5、7、9 对应同一个 gloss（参见

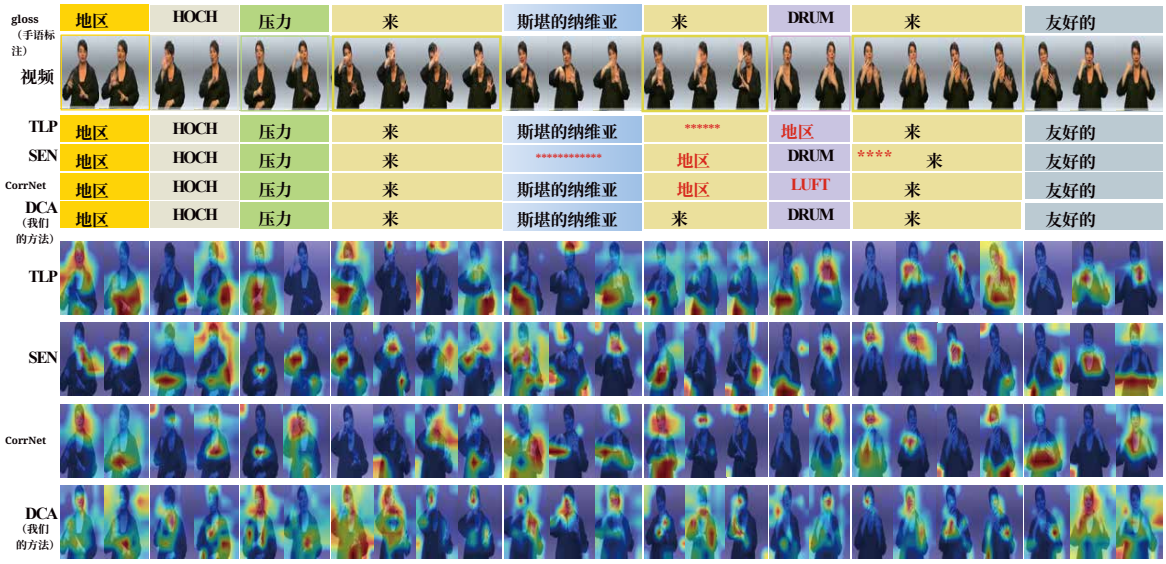


图 4. 在 PHOENIX-2014 测试视频上，可视化 TLP [38], SEN [17], CorrNet [6] 以及所提出的 DCA 的识别结果和 Grad-CAM 结果。带有红色符号的 gloss 表示预测错误的 gloss。区域的颜色深浅（蓝色、黄色、红色、深红色）表示模型对手语空间区域的注意力由弱到强。

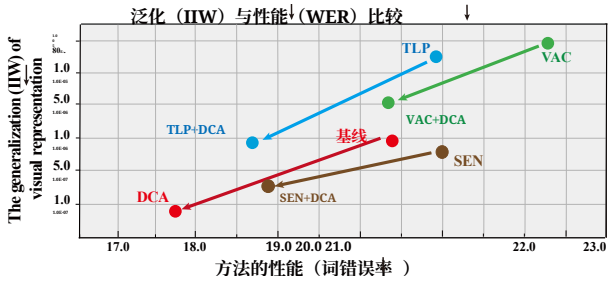


图 5. 评估 DCA 在其他 SOTACSLR 方法上的泛化能力，以及其在 PHOENIX-2014 测试集上的识别准确率。

表 IX 我们的 DCA 与其他连续手语识别方法在 PHOENIX-2014 上的效率比较。所有实验均在批量大小为 1 的 A100 GPU 上测量。

方法	参数量	GFLOPs	吞吐量	开发集（%）↓	测试集（%）↓
VAC	34.3	567	17.0	21.2	22.3
TLP	59.5	573	17.0	19.7	20.2
SEN	34.5	578	15.5	19.5	21.0
C^2ST_{SW+WE}	78.2	1368	4.4	17.6	18.3
DCA（我们的方法）	70.1	1655	10.5	**17.3**	**17.7**

红色方框），与其他方法相比，我们的 DCA 为它们分配了更高的相似度。此外，片段 3 与片段 1 的手语语义不同，而 DCA 显示这两个片段之间的不相似度比其他方法更大。这些结果表明，我们的 DCA 能够有效学习全局时序上下文。

评估 DCA 对其他 CSLR 方法的泛化能力。 我们采用权重中存储信息的压缩（IIW） [36]，以定量衡量 DCA 的泛化能力。如图 5 所示，我们观察到，在进行 DCA 优化后，所有方法的性能和泛化能力均得到持续提升，词错误率和 IIW 值显著下降。具体而言，VAC [19] 和 TLP [38] 在性能和泛化能力方面均取得了显著提升。尤其是，TLP 和 SEN 在 Phoenix-2014 测试集上取得了出色的识别结果（词错误率分别为 18.6% 和 18.8%）。

方法效率比较。 在本节中，我们使用 THOP [51] 工具评估连续手语识别方法的参数量和 GFLOPs，并评估其推理过程中的吞吐量（视频数/秒）。我们采用 140 帧作为默认设置。如表 IX 所示，我们的 DCA 在识别性能与推理速度（见吞吐量）之间实现了平衡。

对手语动作捕捉能力的评估。 为进一步定性评估我们的 DCA 捕捉手语动作的能力，我们在图 4 中使用 PHOENIX-2014 测试集中的一个示例视频，可视化了识别结果以及 TLP [38], SEN [17], CorrNet [6],和我们提出的 DCA 的类别激活映射（CAM） [50]。在这个例子中，动词 gloss“来”在句子中出现了三次，每次分别表达“来了”“来自”和“将来”的含义。“来”的解释会随上下文而改变，因此需要结合手语 gloss 的全局上下文来调整全局时序上下文，才能准确理解。我们的 DCA 成功识别了所有“来”，而 TLP、SEN 和 CorrNet 未能正确理解第二次出现的“来”，因为它们没有利用文本语法来调整全局时序上下文。在类别激活映射方面，我们发现，与其他方法相比，DCA 能够更敏锐、准确地捕捉视频特定区域中对应的手语动作，展现出更优的识别能力。

五、结论

本研究探讨了利用文本语法改进连续手语识别（CSLR）视频表征的效果。我们提出了一种名为**去**噪**对**比**齐**（DCA）的新方法，通过对齐视频与手语 gloss 序列，实现实例对应和全局上下文对齐，以建模文本语法。DCA 以对比方式近似手语与 gloss 之间的实例对应关系，为视频编码器提供手语语义，以改进其对每个手语动作的表征。此外，DCA 还通过去噪扩散对齐来对齐视频与手语 gloss 序列表示。该过程使模型能够根据视频，通过对来自噪声的手语 gloss 序列表示进行生成式去噪，学习手语 gloss 的全局上下文。这种对齐有助于视频编码器改进对全局时序上下文的理解。在公开的连续手语识别基准数据集上的实验结果表明，DCA 达到了优异的识别性能。其他消融研究结果不仅验证了所提出的去噪扩散对齐的有效性，也凸显了去噪扩散模型在提升视觉表示学习方面的潜力。

局限性。与基线方法相比，DCA 每个训练周期将增加约 8 分钟的训练时间开销，面临过高的计算开销。我们仍可能需要长期研究，才能使去噪扩散模型在视觉特征提取中稳定运行，并替代连续手语识别中的对齐过程。

参考文献

[1] H. Hu、J. Pu、W. Zhou 和 H. Li，“协作式多语种连续手语识别：统一框架”，*TMM*，第 1–12 页，2022年。[2] W. Xue、J. Liu、S. Yan、Y. Zhou、T. Yuan 和 Q. Guo，“缓解中国手语识别的数据不足问题”，*Visual Intelligence*，第 1 卷，第 1–9 页，2023年。[3] J. Zheng、Y. Wang、C. Tan、S. Li、G. Wang、J. Xia、Y. Chen 和 S. Z. Li，“CVT-SLR：通过变分对齐实现手语识别的对比视觉文本变换”，载于*CVPR*，2023年。[4] L. Guo、W. Xue、B. Liu、K. Zhang、T. Yuan 和 D. Metaxas，“手语 gloss 先验引导的连续手语识别视觉特征学习”，*IEEE*图像处理汇刊，2024年。[5] Y. Min、P. Jiao、Y. Li、X. Wang、L. Lei、X. Chai 和 X. Chen，“用于视觉序列学习的深度径向嵌入”，载于*ECCV*，2022年。[6] L. Hu、L. Gao、Z. Liu 和 W. Feng，“基于相关性网络的连续手语识别”，载于*CVPR*，2023年。[7] S. Gan、Y. Yin、Z. Jiang、H. Wen、L. Xie 和 S. Lu，“SignGraph：一个手语序列胜过节点图”，载于*CVPR*，2024年，第 13 470–13 479 页。[8] J. Pu、W. Zhou、H. Hu 和 H. Li，“通过跨模态增强提升连续手语识别”，载于*ACMMM*，2020年。[9] L. Guo、W. Xue、Q. Guo、B. Liu、K. Zhang、T. Yuan 和 S. Chen，“为连续手语识别提炼跨时间上下文”，载于*CVPR*，2023年。[10] H. Zhang、Z. Guo、Y. Yang、X. Liu 和 D. Hu，“C2ST：用于连续手语识别的跨模态上下文化序列转导”，载于*ICCV*，2023年。[11] K. Clark、U. Khandelwal、O. Levy 和 C. D. Manning，“BERT关注什么？对BERT注意力的分析”，载于*ACL*研讨会，2019年8月，第 276–286 页。[12] J. Johnson、A. Gupta 和 L. Fei-Fei，“从场景图生成图像”，载于*IEEE*计算机视觉与模式识别会议论文集，2018年，第 1219–1228 页。[13] D. A. Hudson、D. Zoran、M. Malinowski、A. K. Lampinen、A. Jaegle、J. L. McClelland、L. Matthey、F. Hill 和 A. Lerchner，“SODA：用于表征学习的瓶颈扩散模型”，载于*CVPR*，2024年，第 23 115– 23 127 页。[14] A. Graves、S. Fernández、F. Gomez 和 J. Schmidhuber，“连接主义时序分类：利用循环神经网络标注未分段序列数据”，载于国际机器学习大会，2006年。[15] R. Zuo 和 B. Mak，“C2SLR：一致性增强的连续手语识别”，载于 CVPR，2022年。[16] Y. Chen、R. Zuo、F. Wei、Y. Wu、S. LIU 和 B. Mak，“用于手语识别与翻译的双流网络”，载于神经信息处理系统大会，2022年。[17] L. Hu、L. Gao、Z. Liu 和 W. Feng，“用于连续手语识别的自强调网络”，载于人工智能促进协会年会，2023年。[18] Y. Chen、F. Wei、X. Sun、Z. Wu 和 S. Lin，“用于手语翻译的简易多模态迁移学习基线”，载于 CVPR，2022年。[19] Y. Min、A. Hao、X. Chai 和 X. Chen，“用于连续手语识别的视觉对齐约束”，载于 ICCV，2021年。[20] W. Li、C. Gao、G. Niu、X. Xiao、H. Liu、J. Liu、H. Wu 和 H. Wang，“UNIMO：通过跨模态对比学习实现统一模态理解与生成”，载于计算语言学协会年会，2021年。[21] A. Radford、J. W. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark、G. Krueger 和 I. Sutskever，“通过自然语言监督学习可迁移视觉模型”，载于国际机器学习大会，2021年。[22] R. Huang、J. Han、G. Lu、X. Liang、Y. Zeng、W. Zhang 和 H. Xu，“DiffDis：赋予生成式扩散模型跨模态判别能力”，载于 ICCV，2023年，第 15 667–15 677 页。[23] P. Jin、H. Li、Z. Cheng、K. Li、X. Ji、C. Liu、L. Ming Yuan 和 J. Chen，“DiffusionRet：基于扩散模型的生成式文本-视频检索”，载于 ICCV，2023年，第 2470–2481 页。[24] X. Chen、Z. Liu、S. Xie 和 K. He，“解构去噪扩散模型以实现自监督学习”，arXiv，2024年。[25] W. Wang、Q. Sun、F. Zhang、Y. Tang、J. Liu 和 X. Wang，“扩散反馈有助于提升 CLIP 的视觉能力”，arXiv，第 abs/2407.20171 卷，2024年。[26] D. Marcheggiani 和 I. Titov，“利用图卷积网络编码句子以进行语义角色标注”，载于自然语言处理经验方法会议，2017年9月，第 1506–1515 页。[27] J. D. M.-W. C. Kenton 和 L. K. Toutanova，“BERT：面向语言理解的深度双向变换器预训练”，载于 NACCL-HLT，第 1 卷，2019年，第 2 页。[28] S. Chen、Z. Hong、G. Xie、W. Wang、Q. Peng、K. Wang、J. Jun Zhao 和 X. You，“MSDN：用于零样本学习的互语义蒸馏网络”，载于 CVPR，2022年，第 7602–7611 页。[29] J. Ho、A. Jain 和 P. Abbeel，“去噪扩散概率模型”，载于神经信息处理系统大会，2020年。[30] J. Song、C. Meng 和 S. Ermon，“去噪扩散隐式模型”，载于国际学习表征会议，2021年。[31] D. A. Hudson、D. Zoran、M. Malinowski、A. K. Lampinen、A. Jaegle、J. L. McClelland、L. Matthey、F. Hill 和 A. Lerchner，“SODA：用于表征学习的瓶颈扩散模型”，载于 CVPR，2024年6月，第 23 115–23 127 页。[32] F. Huang、S. Song 和 L. Zhang，“ 无监督域自适应中的梯度协调 ”，IEEE 模式分析与机器智能汇刊，第 46 卷，10 336，2024年12月，10 319 号，第 – 页。[33] O. Koller、J. Forster 和 H. Ney，“连续手语识别：面向可处理多个手语者的大词汇量统计识别系统”，《计算机视觉与图像理解》，2015年。[34] N. C. Camgoz、S. Hadfield、O. Koller、H. Ney 和 R. Bowden，“神经网络手语翻译”，载于 CVPR，2018年。[35] H. Zhou、W. Zhou、W. Qi、J. Pu 和 H. Li，“通过手语反向翻译，利用单语数据改进手语翻译”，载于 CVPR，2021年。[36] Z. Wang、S. Huang、E. E. Kuruoglu、J. Sun、X. Chen 和 Y. Zheng，“PAC-贝叶斯信息瓶颈”，载于国际学习表征会议，2022年。[37] K. He、X. Zhang、S. Ren 和 J. Sun，“用于图像识别的深度残差学习”，载于 CVPR，2016年。[38] Z. Lianyu Hu、Liqing Gao 和 W. Feng，“用于连续手语识别的时序提升池化”，载于 ECCV，2022年。[39] Z. Xie、Z. Zhang、Y. Cao、Y. Lin、J. Bao、Z. Yao、Q. Dai 和 H. Hu，“SimMIM：一种简易的掩码图像建模框架”，载于 CVPR，2022年6月，第 9653–9663 页。[40] K. He、X. Chen、S. Xie、Y. Li、P. Doll´ar 和 R. Girshick，“掩码自编码器是可扩展的视觉学习器”，载于 CVPR，2022年，第 15 979–15 988 页。[41] D. P. Kingma 和 J. Ba，“Adam：一种随机优化方法”，载于国际学习表征会议，2015年。[42] A. Hao、Y. Min 和 X. Chen，“用于连续手语识别的自互蒸馏学习”，载于 ICCV，2021年。[43] W. Xue、Z. Kang、L. Guo、S. Yang、T. Yuan 和 S. Chen，“通过自引导网络实现面向听障消费者交流的连续手语识别”，IEEE消费电子汇刊，2023年。

[44] P. Jiao, Y. Min, Y. Li, X. Wang, L. Lei, 和 X. Chen，“CoSign：探索基于骨架的连续手语识别中的共现信号”，载于 *ICCV*，2023年。[45] H. Hu、W. Zhao、W. Zhou 和 H. Li，“SignBERT+：面向手语语言理解的手部模型感知自监督预训练”，*IEEE*模式分析与机器智能汇刊，第1–20页，2023年。[46] J. Liu、W. Xue、K. Zhang、T. Yuan 和 S. Chen，“TB-Net：用于连续手语识别的帧内与帧间视频相关性学习”，*Information Fusion*，第109卷，第102438页，2024年。[47] J. Zheng、Y. Wang、C. Tan、S. Li、G. Wang、J. Xia、Y. Chen 和 S. Z. Li，“CVT-SLR：通过变分对齐实现用于手语识别的对比视觉文本变换”，载于 *CVPR*，2023年。[48] A. Gretton、K. M. Borgwardt、M. J. Rasch、B. Scholkopf 和 A. Smola，“核双样本测试”，*J. Mach. Learn. Res.*，第13卷，773，2012年 723页。– 。[49] M. Long、H. Zhu、J. Wang 和 M. I. Jordan，“通过联合适应网络进行深度迁移学习”，载于 国际机器学习大会，第70卷，2017年，第2208–2217页。[50] R. R. Selvaraju、M. Cogswell、A. Das、R. Vedantam、D. Parikh 和 D. Batra，“Grad-CAM：通过基于梯度的定位实现深度网络的视觉解释”，载于 *ICCV*，2017年。[51] Lyken17、HaoKang-Timmy、lvmingzhe 和 ttumiel，“THOP：PyTorch-opcounter”。 [在线]。可用地址：https://github.com/Lyken17/ pytorch-OpCounter，2019年.8



Leming Guo 于2024年获得天津理工大学计算机科学与技术博士学位。他目前是天津理工大学计算机科学与工程学院讲师。其研究兴趣包括手语识别、动作识别、领域自适应和多媒体分析。



Wanli Xue 是天津理工大学计算机科学与工程学院教授。他于2019年获得天津大学计算机应用技术博士学位。其研究兴趣包括视觉跟踪、图像拼接和手语识别。



Shengyong Chen 于2003年获得香港城市大学机器人视觉博士学位。2006年至2007年，他在德国汉堡的汉堡大学工作。目前，他是中国天津天津理工大学的教授。他目前的研究兴趣包括计算机视觉、机器人学和图像分析。