

基于深度模型的视频缩略动图生成

July 1, 2022

摘要: 当今社会, 视频逐渐成为互联网传播信息的一个主要媒介。因此, 如何高效地产生视频缩略动图, 从而为人们搜索信息提供助力成为了研究的一个热门方向。本篇文章中提出了一个基于深度模型的缩略动图生成方法, 并且在此基础上, 针对讲座类型的视频设计了特殊的关键帧提取方式和评分方法。这种方法可以将时间复杂度控制在视频总时长的 2% 之内, 具有很高的计算效率。

关键词: 深度神经网络; 视频缩略动图; 图像处理;

1 引言

据统计, 每分钟都有长达几百小时的视频被上传到视频分享网站, 并被数亿人观看和分享。为了更好地从大量的视频中选取有效信息, 人们常常依靠视频缩略动图来作出判断。视频缩略动图由视频最具有代表性的片段构成, 能够在最短的时间内吸引人们的目光 [1], 并且在很大程度上影响人们浏览时的行为: 点击观看或者继续搜索 [2]。一个好的缩略动图能够提高视频的浏览量, 因此如何为每一个视频配备一个优秀的缩略动图一直是视频创作者们需要解决的问题。

缩略动图的评价准则具有高度的主观性, 因此在过去, 缩略动图常常由视频创作者们自行剪辑产生。但是随着视频量的增长, 这种方法会消耗大量的创作精力; 同时, 由于视频的来源多样, 网站很难保证缩略动图的质量。因此, 本文中提出了一个能够自动产生缩略动图的通用算法, 通过深度模型提取视频的核心特征, 由此组合出 60 帧左右的动图, 在保证缩略动图质量的同时大大减小了视频处理时间。

本文中提出的通用算法由关键帧提取、特征向量提取、聚类三个阶段组成, 其最大优点是具有很

低的时间复杂度, 这尤其适合于长视频的处理。利用单核 CPU, 本方法可以在 2 分钟内完成对 100 分钟视频的处理, 并且产生相对良好的缩略动图; 同时, 基于成熟的图像处理算法, 我们能够提取出最高质量的帧作为缩略动图的候选, 这保证了最终结果在帧水平上的质量。除此之外, 本文首创性地利用预训练模型来提取视频的深度特征, 这是大规模预训练模型在视频缩略动图领域的创新性应用。

基于通用算法, 本文创新性地针对课程讲座设计了专用的关键帧提取方法, 并且引入了视频标题作为生成过程中的监督信息。新的算法在数据集上取得了较好的生成效果。

2 相关工作

2.1 基于传统图像处理的方法

Gao et al. [3] 首先提出基于主题的视频缩略动图生成方法; 具体而言, 他们运用了互联网的庞大数据库, 选取与具有某个特征主题(通常是视频标题)的网络图片具有很高的相似度的视频片段作为缩略动图。Song et al.[4] 运用聚类的方式, 通过多次的分类操作筛选获得最具有吸引力的关键

帧作为缩略动图的候选。

这些方法无论是在算法的设计还是运用上都能实现最低的时间复杂度，这说明传统的图像处理方法不失为一种在计算量和质量之间的折中选择。但是传统的方法通常过多依赖图像的统计特性，如 HSV 空间中的统计量和灰度共生矩阵 (GLCM[5]) 等来构成图像的特征向量，这种方法难以描述图像的高级特性（如图像中物体的位置，是否有人出现等），从而难以对生成的缩略动图进行更高层次上的控制与选择。

2.2 基于深度学习的方法

考虑到传统图像处理方法在特征提取方面的不足，人们开始引入深度学习模型，以期待获得更好的图像表征从而促进关键帧的提取。利用视频所特有的序惯性，Zhang et al.[7] 使用了序列至序列的监督学习方法，具体架构为 LSTM 模型。另一种类似的方法 [8] 中使用了卷积神经网络，但总体架构基本相似。在无监督学习方面，Xu et al.[6] 利用 VAE 模型来学习缩略动图的特性，这种特性在文章中被定义为给定视频时的后验概率分布。[9] 使用了 Attention 模型来处理视频的总结问题，并且证明了完全依靠 Attention 的方法能够实现视频边缘的检测。

以上几种方法在传统方法的基础之上引入了深度学习的思想，在关键帧识别与特征提取方面达到了优良的性能。但是这些方法可以看作是图像处理的扩展应用，并没有利用视频的特殊性质，如多通道中的声音信息等。

2.3 多模态信息的运用

利用视频中的音轨，可以更好的判定视频内容，从而保证视频缩略动图的内容连贯性与主题相关性。[12] 考虑将每一段语音与视觉物体对齐，从而获得该物体的特征向量；但是这种方法只适用于特殊的应用场景，且对运算的要求较高，并不适用于通用方法的构建。

在视频网站的应用场景下，也可以将文字作为监督的信息，利用人们搜索的输入来产生特定的缩

略动图。这个问题可以建模为“判断并提取出视频中最符合输入文字的片段”，即匹配问题。Liu et al.[10] 利用人们的输入作为问题矩阵与视频片段做 Attention 操作，保证了生成结果与输入文字的相关性。[11] 在此基础上使用了图卷积神经网络来保证缩略动图之间的连贯关系，兼顾了多模态之间的对应与不同模态内部的一致性。

2.4 特殊场景下的专用视频缩略动图生成

授课、讲座、vlog 等视频类型具有非常特殊的性质，如背景较为单一、人物数目较少且出现频率固定等。因此可以针对这些特殊的视频类型设计专用的缩略动图生成方法，提高缩略动图对于内容的概括性。Xu et al.[12] 针对“讲座”类型的视频，在视频帧处理的阶段训练了一个分类器来判别每一帧的类型，从而可以利用 OCR 等技术进一步拆分授课的内容，生成内容更丰富的缩略动图。针对 vlog 类型，[13] 显式地对每一帧进行人脸与情感识别，从而选取最为生动有趣的图像作为关键帧。

3 通用方法

对任意长度的输入视频，通用方法都能够产生长度固定为 60 帧的缩略动图，其产生的结果可以作为基准进行比较分析。特别注意，在强制限制缩略动图帧数为 60 (2 秒) 的情况下，我们几乎不可能通过截取视频片段的方法来构成缩略图，因此最终产生的结果为 60 帧离散画面，通过定格的方式逐一播放。

3.1 第一阶段：提取关键帧

流媒体中的视频帧率一般为 24，即每秒由 24 帧不同的画面构成，利用人眼的视觉暂留现象形成动态的画面。在这 24 帧之中，绝大多数帧都是所谓的“过渡帧”，即上一个动作至下一个动作的过度，或者场景之间的切换画面。这些帧常常具有模糊、昏暗的特点，并且难以体现出视频的核心思想，因此在进行特征提取之前首先需要对数据集进行清洗，过滤掉这些无效的输入。

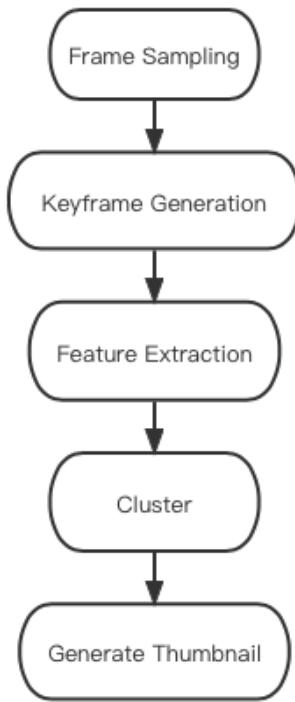


Figure 1: 通用算法流程图

传统的图像处理方法提供了许多评价图片的统计标准，比如亮度、锐度等。Redi et al.[14] 对这些统计特性进行了回归分析，最终发现图像的明亮度对图像的质量影响较大。除此之外，将两帧图像进行差分，得到图像的平均像素强度（帧间差分强度）可以用来衡量两帧图像的变化大小。由于关键帧通常具有较大的信息熵率（含有的新信息较多），可以认为帧间差分强度的大小与帧的重要程度成正相关。

各个判决准则如下：

$$\begin{aligned} \text{Luminance}(I_{rgb}) &= 0.2126I_r + 0.7152I_g + 0.0722I_b \\ \text{Difference}(I_{gray}) &= \sum(I_i \text{ gray} - I_{i-1} \text{ gray}) \end{aligned}$$

利用超参数 α 确定两个判决准则的相对重要程度。最终的评分 $Score$ 如下定义，其中 α 取经验值 0.2。

$$Score(I) = \alpha \text{ Luminance}(I_{rgb}) + \text{Difference}(I_{gray})$$

实际操作中，考虑到计算的复杂度，在对数据进行清洗之前首先需要对视频的所有帧进行采样。

为了控制处理时间在实际时间的 2% 之内，视频首先需要被截短成为 5 秒的短片段集合，每一秒的视频只随机采样 5 帧来计算评分。如果短片段的个数超过一个参数（实验中设定为 400），则只选取评分为前 $5 * 400 = 2000$ 帧进行下一步的特征提取。这主要是因为第一阶段对视频每一帧的遍历耗时与视频长度成正相关，而第二阶段的特征提取耗时只与关键帧的长度正相关。因此，是否能够有效地提取关键帧是限制算法总体效率的瓶颈之一。在第五部分中，本文对不同的关键帧筛选方法进行了评价，最终发现限制关键帧的长度是最能够兼顾生成缩略图质量和减小运算时间的方法。

不同的关键帧提取方法能够直接影响到生成缩略图的特性。针对不同的视频类别，我们可以设定特殊的评分标准，这一部分将在第三部分中进行分析。

3.2 第二阶段：特征向量的提取

实验中使用了两种大规模的预训练模型来作为特征向量的提取器。CLIP[15] 是在大规模视频-文本语料集上训练得到的模型，并且在诸多下游应用，如图片分类、图片问题回答中取得了 SOTA 效果。通过利用 CLIP 模型的预训练权重，我们可以将图片的浅层特征（如像素的概率分布）转化为 512 维度的深层特征，并且在这个特征空间中进行接下来的聚类分析。同样的，另一个预训练模型（Pre-trained ImageNet Classification Model）也可以用作特征提取器。

两个预训练模型之间存在一些区别，这涉及到多模态信息的使用。CLIP 模型在提取图像特征的同时，还能够将对应的自然文本映射到相同的特征空间，这为文本作为监督信息的引入提供了方便；反之，ImageNet 模型并不涉及到除标签之外的文本，因此难以获得自然语言文本的相应特征表示。这两种方法在图片特征提取的领域都具有非常优良的性能，由于通用算法并不涉及到文本的使用，因此可以任意选取一个模型作为特征提取器。

在性能评价方面，CLIP 模型可以识别图像形式的文字 [18]，这与其大规模的训练数据有关。因

此实验中首选利用 CLIP 模型，以求获得良好的泛化性。

3.3 第三阶段：聚类与筛选

特征空间具有许多特殊的性质，如低维不可分的数据在高维度通常是可分的。因此通过将数据映射到高维空间（如 CLIP 输出向量的 512 维），人们可以运用简单的聚类方法对图像进行分析。考虑到计算的复杂性，聚类的族数选择为 4，这也与故事发展的四个阶段-起承转合相对应。聚类之后，每个关键帧会拥有一个对应的标签，通过选择与聚类质心最接近的关键帧便可以简便地构建出视频的缩略动图。

4 针对讲座的视频缩略图生成

“讲座”(Lecture) 是一种非常特殊的视频类型。从视觉的角度而言，“讲座”通常由两种画面构成：授课的教师与讲稿（或者黑板）。第一种画面的组成通常为人、讲台、黑板，结构相对固定，易于分辨。第二种画面则是高度结构化的 PPT 等，可以利用光学 OCR 的方式进行物体的拆分。总之，针对这种特殊的视频结构设计专用的缩略图生成模型具有很高的应用价值。

4.1 第一阶段：基于预训练分类器的关键帧提取

缩略动图的一个特性是含有有趣的物体，如人、夸张的表情、文字等。在通用算法中并未考虑引入对特定物体的限制，因为这种高层语义高度依赖于视频的类型。但是在“讲座”的视频环境下，我们可以考虑引入帧的类型来作为关键帧的选取原则。

借鉴 [12] 的实现思路，本方法中采用预训练的分类器来判定帧的类型。预训练分类器采用 ResNet-34 作为基础架构，可以直接用在端到端的模型中，实现输入帧的分类判决。为了提高判定的准确度并降低计算的复杂度，实验中采用了简化分类模型，可以将输入帧分为以下三类：

- others
- presenter_slide
- slide

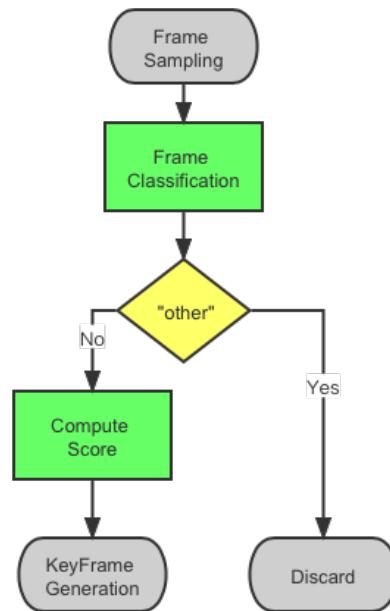


Figure 2: 基于分类器的关键帧提取 流程图

“other”类别中包含了所有我们不感兴趣的类型，如观众、观众与教师等与主题无关的场景。因此在关键帧筛选中可以首先对采样的帧进行分类判决：如果判决得到的帧类型为 other，则直接丢弃；若不是，再进入下一步的关键帧选取流程。这个方法能够进一步降低所生成的关键帧数量，从而降低了计算时间。

4.2 第二阶段：特征向量提取

CLIP 模型中提供了具有良好特性的自然语言编码器，因此在针对讲座的特殊算法中，我们采用 CLIP 模型作为特征提取器。注意，此时的特征向量仍旧是图像编码器 (Image Encoder) 的直接输出，在这个特征空间中即可完成对于帧的聚类操作。

4.3 第三阶段：多模态信息的聚类与判决

“讲座”类视频通常具有丰富的文字信息。考虑到利用 OCR 提取全部 PPT 中的文字再进行分析会大大增加计算的复杂度，我们可以利用视频标题作为监督的文字信息。

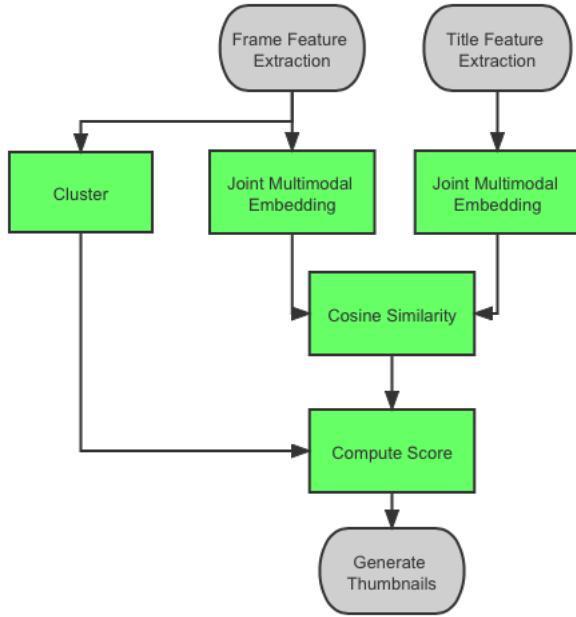


Figure 3: 多模态信息的聚类与判决 流程图

与通用算法不同，在聚类之后，关键帧的选取不再仅以距离质心的距离为评价准则。利用 CLIP 模型中提供的联合多模态编码方法，在通用算法中提取出的帧特征向量和标题的特征向量可以被映射到同一个空间。在这个空间里借鉴 Cosine 相似度的思想，可以定义帧与视频标题的相似度为：

$$\text{Similarity}(F(I), F(\text{text})) = \frac{F(I) \cdot F(\text{text})}{\|F(I)\| \|F(\text{text})\|}$$

因此，最终产生缩略动图的评分可以定义为：

$$\begin{aligned} \text{Score}(I, \text{text}) &= \beta \cdot \text{Similarity}(F(I), F(\text{text})) \\ &\quad + \text{Distance}(I) \end{aligned}$$

其中 Smimilarity 表示帧与标题的相似程度， Distance 表示帧到聚类质心的距离， $F(\cdot)$ 表示联合多模态编码， β 为经验值，实验中取 1。

5 结果与讨论

5.1 通用算法时间复杂度

算法对视频的处理用时高度依赖于视频的长短。正如 3.1 中所介绍的，为了限制处理时间的线性增长趋势，在视频的关键帧提取步骤中本算法采用了硬限制的方式，即关键帧的数目超过一定阈值之后便不再增长。利用长为 60min 的视频，可以绘制出如下所示的处理时间-视频长度图。图 4 中 (a) 为关键帧提取、筛选所用的时间，总结为 initiation time；(b) 为以利用 CLIP 模型为例，从关键帧中抽取特征向量的用时 (feature extraction time)；图 (c) 为总耗时。

可以看到，关键帧的抽取过程耗时近似与视频长度成正比，这主要是因为这个过程需要遍历所有的帧并进行计算。特征提取的过程在视频长度较短的时候与之成正比，但是在视频长度大于某一个临界值时（实验中设定为 15min）处理时间便不再增长。这直接导致总处理时间的增长趋势在这个阈值附近出现降低，线性增长的系数得到抑制。实验中 100min 的视频处理时间约为 118.3 秒，可以认为实现了 0.02^* 视频长度的时间复杂度要求。

5.2 针对讲座的生成算法评价

为了评判针对讲座的生成算法是否能够捕捉到讲座中与主题相关的特定内容，我们分别利用通用算法和特殊算法对同一段讲座视频进行处理，并得到如下所示的结果。以下图片是按照等距的原则，在生成缩略动图的第 15 帧、30 帧、45 帧位置处抽取的图样，视频 [16] 来源于 Youtube，标题为：“Lecture 1 | Introduction to Convolutional Neural Networks for Visual Recognition”。

可以看出，虽然两种方法都能够产生来自于不同视频片段的关键帧，但是通用算法产生的结果有大量的时间关注在演讲者身上，这与演讲的题目并没有很强的相关性。但是作为通用算法，它捕捉到了有人物存在的场景，这个特性能使其胜任大部分视频类型的生成任务。针对于讲座的专用算法，反之，能够更好的捕捉到与视频主题相关的帧，并

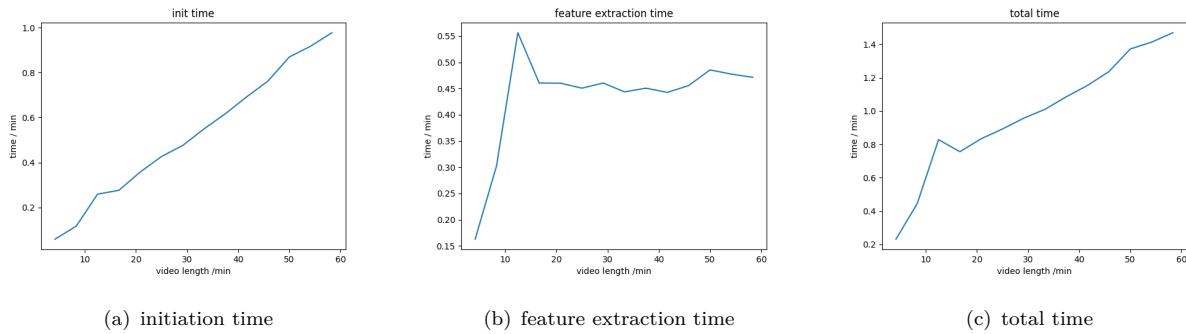


Figure 4: 通用算法耗时

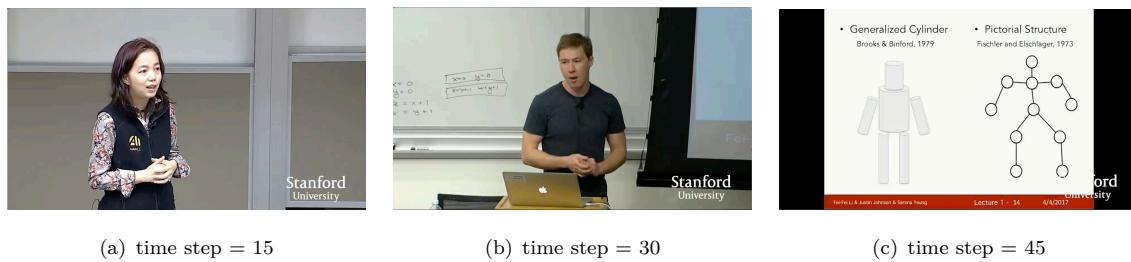


Figure 5: 通用算法：关键帧



Figure 6: 专用算法：关键帧

且在高层语义上实现对视频的概括。因此可以认为专用算法在对于讲座类型的视频处理中表现更为出色。

6 总结

本文专注于视频缩略图的生成，提出了基于深度神经网络的视频缩略图生成方法，并且针对特殊类型的视频设定了不同的算法。除此之外，本文分析了所提出算法的时间特性和特殊算法的有效性，证明了这种算法能够在生成结果的质量与处理时间两个特点之间取得较好的平衡。虽然这种算法可能具有一定的缺点，如生成的缩略图时间较短、难以考虑内容的连贯性等，但是这种方法不失为解决缩略图产生问题的一次很好的尝试。

References

- [1] G. Buscher, E. Cutrell, and M. R. Morris. What do you see when you're surfing?: using eye tracking to predict salient regions of web pages. In CHI, 2009.
- [2] S. J. Cunningham and D. M. Nichols. How people find videos. In JCDL, 2008.
- [3] Y. Gao, T. Zhang, and J. Xiao. Thematic video thumbnail selection. In ICIP, 2009.
- [4] Song Y, Redi M, Vallmitjana J, et al. To click or not to click: Automatic selection of beautiful thumbnails from videos[C]//Proceedings of the 25th ACM international conference on information and knowledge management. 2016: 659-668.
- [5] Hall-Beyer M. GLCM texture: a tutorial[J]. National Council on Geographic Information and Analysis Remote Sensing Core Curriculum, 2000, 3(1): 75.
- [6] Xu Y, Bai F, Shi Y, et al. GIF Thumbnails: Attract More Clicks to Your Videos[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(4): 3074-3082.
- [7] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman, "Video summarization with long short-term memory," in Proceedings of the European Conference on Computer Vision (ECCV), 2016
- [8] Mrigank Rochan, Linwei Ye, and Yang Wang, "Video summarization using fully convolutional sequence networks," in Proceedings of the European Conference on Computer Vision (ECCV), 2018.
- [9] Vasileios Argyriou Dorothy Monekosso Jiri Fajtl, Hajar Sadeghi Sokeh and Paolo Remagnino, "Summarizing videos with attention," Proceedings of the AAAI Conference on Artificial Intelligence Workshops (AAAI workshops), 2018.
- [10] Liu M, Wang X, Nie L, et al. Attentive moment retrieval in videos[C]//The 41st international ACM SIGIR conference on research & development in information retrieval. 2018: 15-24.
- [11] Yuan Y, Ma L, Zhu W. Sentence specified dynamic video thumbnail generation[C]//Proceedings of the 27th ACM International Conference on Multimedia. 2019: 2332-2340.
- [12] Xu C, Wang R, Lin S, et al. Lecture2Note: Automatic Generation of Lecture Notes from Slide-Based Educational Videos[C]//2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2019: 898-903.
- [13] Akari Shimono, Yuki Kakui, and Toshihiko Yamasaki. 2020. Automatic YouTube-

- Thumbnail Generation and Its Evaluation.
In Proceedings of the 2020 Joint Workshop
on Multimedia Artworks Analysis and At-
tractiveness Computing in Multimedia
- [14] Redi M, O'Hare N, Schifanella R, et al. 6
seconds of sound and vision: Creativity in
micro-videos[C]//Proceedings of the IEEE
Conference on Computer Vision and Pattern
Recognition. 2014: 4272-4279.
- [15] Radford A, Kim J W, Hallacy C, et al. Learn-
ing transferable visual models from natural
language supervision[C]//International Con-
ference on Machine Learning. PMLR, 2021:
8748-8763.
- [16] Youtube. "Lecture 1 | Introduction to Convo-
lutional Neural Networks for Visual Recogni-
tion".Online video clip.
- [17] Yang W, Tsai M H. Improving YouTube
video thumbnails with deep neural nets[J].
Google Research Blog, Oct, 2015, 8: 5.
- [18] Sandhini Agarwal and Gretchen Krueger and
Jack Clark and Alec Radford and Jong Wook
Kim and Miles Brundage, Evaluating CLIP:
Towards Characterization of Broader Capa-
bilities and Downstream Implications