ICS 35.240.01
CCS L 72

中华人民共和国国家标准

GB/T XXXXX—XXXX

信息技术　多场景、多字体藏文文字识别数据标注

Information technology—Data labeling for multi-scenario and multi-font Tibetan text recognition

（本草案完成时间：2026 年 11 月 18 日，征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX-XX-XX 发布　　XXXX-XX-XX 实施

国家市场监督管理总局
国家标准化管理委员会　发布

目　　次

前言 III
1 范围 1
2 规范性引用文件 1
3 术语和定义 1
4 图文数据采集标注流程 2
5 标注任务前期准备要求 3
5.1 待标注数据的采集及确定 3
5.2 图文数据标注任务 3
5.3 场景标注规则 4
5.4 文本框标注规则 4
5.5 字体类型标注规则 6
5.6 文字内容标注规则 7
5.7 标注格式示例 8
5.8 标注需求变更流程 9
5.9 标注环境要求 9
6 标注任务执行过程的要求 9
6.1 标注任务创建要求 9
6.2 标注任务分发要求 10
附录A （资料性）标注人员管理要求 11
A.1 标注人员范围 11
A.2 标注人员培训 11
A.3 标注人员能力档案 11
参考文献 12

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

本文件由全国信息技术标准化技术委员会（SAC/TC 28）提出并归口。

本文件起草单位：西藏大学、电子科技大学、中国电子技术标准化研究院、青海师范大学、北京邮电大学。

本文件主要起草人：尼玛扎西、于永斌、仁青东主、黄姗姗、陈晓研、王向向、班玛宝、仁增多杰、头旦才让、洛桑嘎登、嘎玛扎西、拥措、群诺、范满平、陈安龙、道吉扎西、刘羽童、冯箫、沈嘉润、盛天颖、张子玥、次仁白玛、李臣波、郑智文、格桑加措、李蕾、李彝利、官却才让。

信息技术　多场景、多字体藏文文字识别数据标注

1　范围

本文件规定了信息技术领域多场景、多字体藏文文字识别数据标注的要求。

本文件适用于藏文图像文字识别、文本检测等处理。

2　规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 13000—2025　信息技术 通用编码字符集（UCS）

GB/T 45674—2025　网络安全技术 生成式人工智能数据标注安全规范

3　术语和定义

下列术语和定义用于本文件。

3.1

图文数据标注　graphic text data labeling

为图文数据样本指定目标变量和赋值的过程。

3.2

图文数据标注任务　graphic text data labeling task

按图文数据标注说明对数据进行标注的活动。

3.3

图文数据标注方　graphic text data labeler

承担图文数据标注任务的人员或机构。

3.4

图文数据需求方　graphic text data user

提出图文数据标注需求的人员或机构。

3.5

图文数据标注管理方　graphic text data labeling administrator

管理图文数据标注任务评估、分发、交付、验收以及质量控制的人员或机构。

3.6

图文标注工具　graphic text labeling tool

具备文种分类、文本框标注（能实现标注框的创建、调整和位置信息获取）、标准藏文编辑（能实现藏文字符的规范录入与正确渲染显示）、专用标签录入、任务分配与日志管理、多角色权限分级及安全管控等功能，并能用于标注结果验收的软件工具或辅助系统。

3.7

图文标注任务说明　graphic text labeling task description

图文数据需求方用于向图文数据标注管理方以及图文数据标注方明确标注任务的文字表达。

注：标注任务说明通常包含对要执行的标注任务的描述、标注方法、正反示例、验收方法与验收指标等内容。

4 图文数据采集标注流程

图文数据采集标注流程包括图文数据采集、图像数据清洗、数据标注、标注数据检验与审核、质量评估五个核心环节，具体流程见图1。首先，采集原始图文资料，并进行数据去重、图像噪声处理和图像旋转裁剪。随后，由经过专业培训的标注人员按照统一的标注规范完成标注（标注规范依据本文件规定的标注规则和数据需求方的具体要求制定）。所有标注数据均应进行文本框检验和文本检验，并接受审核。审核不合格的，应将问题反馈给标注人员，重新标注并检验，直至审核合格。审核合格的数据应由数据质量评估团队进行质量评估，评估通过后方可交付给数据需求方。

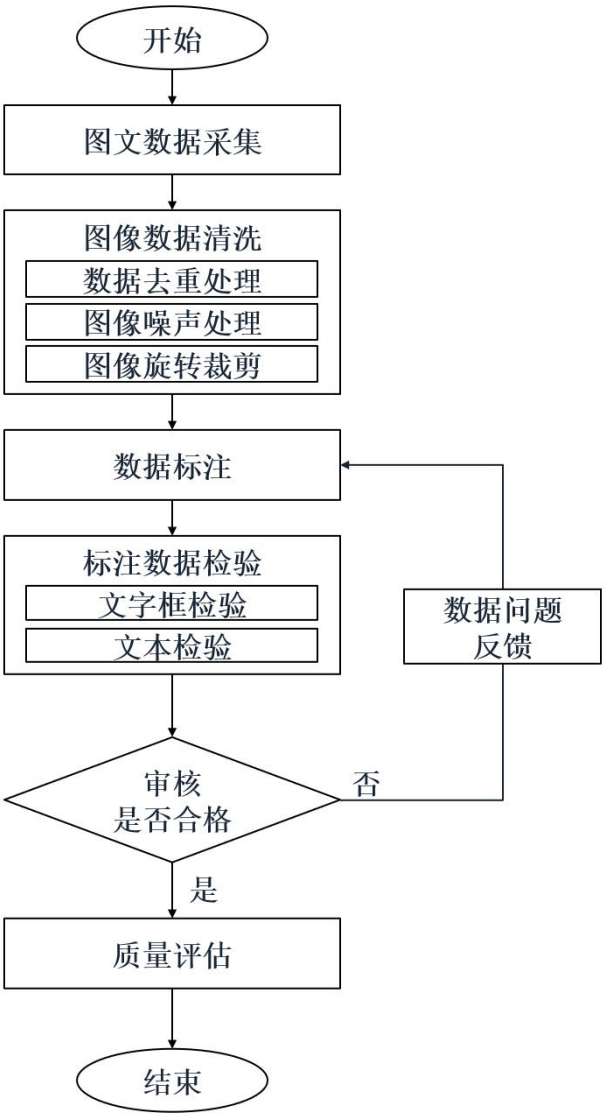


图 1　图文数据采集标注流程

5 标注任务前期准备要求

5.1 待标注数据的采集及确定

待标注数据按以下要求采集及确定。

a) 数据采集：标注管理方按标注需求方的要求采集待标注数据，宜覆盖多样化的场景和类型，如笔记、典籍、书法、路牌、广告、篆刻等。

b) 待标注数据确定。包括：

1) 待标注数据整理：数据需求方应对已采集的待标注数据进行详细分析，包括数据内容、格式、清晰度等方面的核查。对数据进行整理和分类，根据内容特征（如场景、字体、文种等）进行聚类，为后续标注任务提供结构化支持。对于不符合标注要求的数据（如模糊、残缺或重复数据），应及时剔除或补充采集。

2) 数据安全确定：数据需求方、标注管理方及数据标注方应按照 GB/T 45674—2025 给出的要求，检查数据在采集、传输、存储和标注过程中的安全性。对涉及敏感信息的数据（如个人隐私、商业机密等），需进行脱敏处理或加密保护。标注任务中应明确数据安全责任，各方需定期检查数据内容及存储环境，防止数据泄露或滥用。

5.2 图文数据标注任务

以SunOCR标注工具为例，图文数据标注包括四种主要任务：

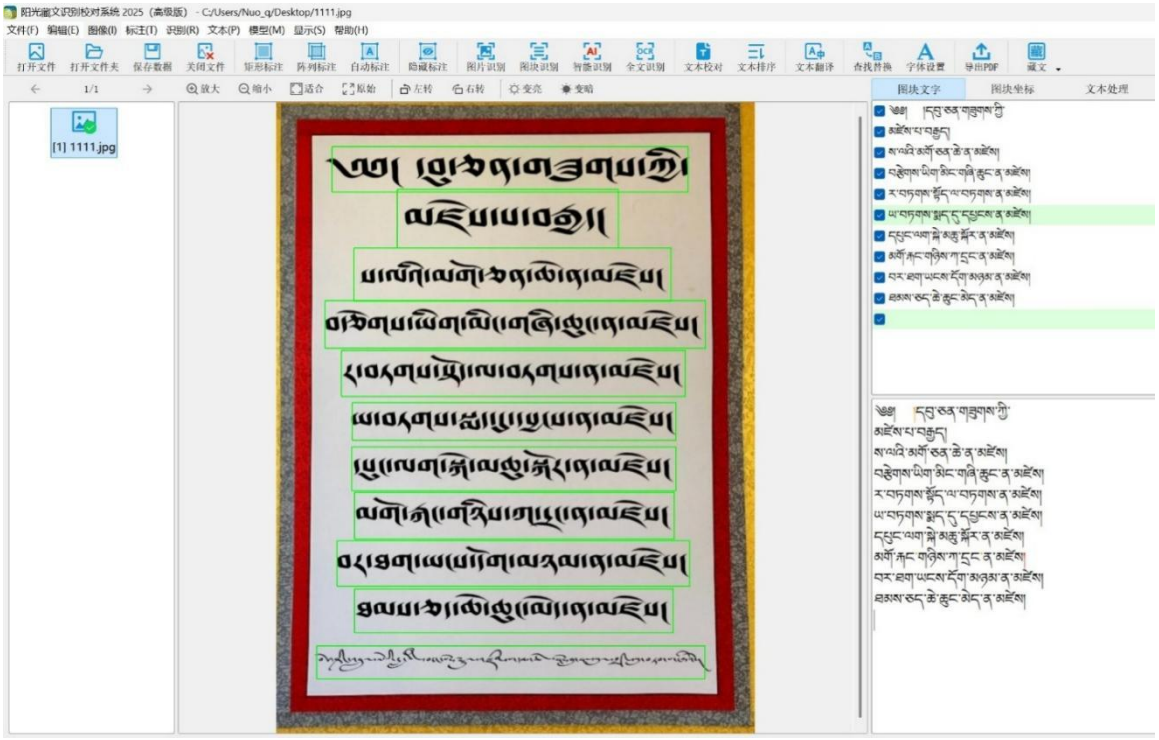
a) 场景标注：标注图像中的场景（如书籍、手写笔记、石碑等）；

b) 文本框标注：标注文本框左上角像素坐标，以及长宽像素值；

c) 字体类型标注：标注每个文本框中文本所使用的字体类型，如乌金体、簇仁体、簇通体、朱匝体、柏簇体、丘伊体等；

d) 文字内容标注：标注每个文本框中对应的文字内容。

示例：



5.3 场景标注规则

5.3.1 场景标签

5.3.1.1 标注时应明确文本出现的具体场景类别，包括但不限于以下类型：

a) 书籍：古籍、历史档案文献、教科书、杂志等；
b) 手写笔记：个人日记、课堂笔记、信件等；
c) 石碑铭文：历史遗迹、纪念碑等；
d) 数字化媒体：网页、移动通讯终端应用、电子书等；
e) 自然场景：户外标识、交通工具标识、建筑标识、选单与包装等场景中的文字。

5.3.1.2 若一张图像中出现有多类场景，先将图像裁剪为多张单场景图像，再进行标注。裁剪过程中宜使各个场景完整。

5.3.1.3 采用汉文和英文进行标注。

5.3.2 场景标签用于文件管理

分别对不同场景的图文数据进行标注。以场景标签为文件夹名，用于管理不同场景的图文标签。

示例：

```
标注数据/
    ├── 笔记/
    │   ├── 笔记_001.jpg
    │   ├── 笔记_002.jpg
    │   └── label.json
    ├── 典籍/
    │   ├── 典籍_001.jpg
    │   ├── 典籍_002.jpg
    │   └── label.json
    ├── 书法/
    │   ├── 书法_001.jpg
    │   ├── 书法_002.jpg
    │   └── label.json
    └── 路牌/
        ├── 路牌_001.jpg
        ├── 路牌_002.jpg
        └── label.json
```

5.4 文本框标注规则

5.4.1 标注方法

按住鼠标左键，从文本框左上角拖动，将鼠标移动到文本框右下角后松开左键，标出文本区域。

5.4.2 标注格式

通过标注文本框左上角像素坐标（x轴坐标、y轴坐标）和文本框长宽，标注文本框的区域。

示例：[x, y, x+w, y+h]。其中，(x, y)是左上角坐标，w 和 h 分别是文本框长和宽。

5.4.3 注意事项

对文本框标注时，注意事项如下。

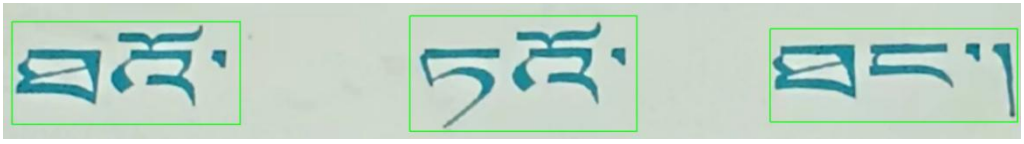
a) 若图像不是正向的，应先将图像进行旋转矫正，然后再标注。

b) 文本框是包括文字的最小外边框，不应过大，也不应压到文字。

c) 一个框中只应包含一行文字，不应包含多行文字。

d) 以文本行高度为标准，如果文字与文字之间的左右间距大于 2 倍高度，则文本行应分成两个框标注。

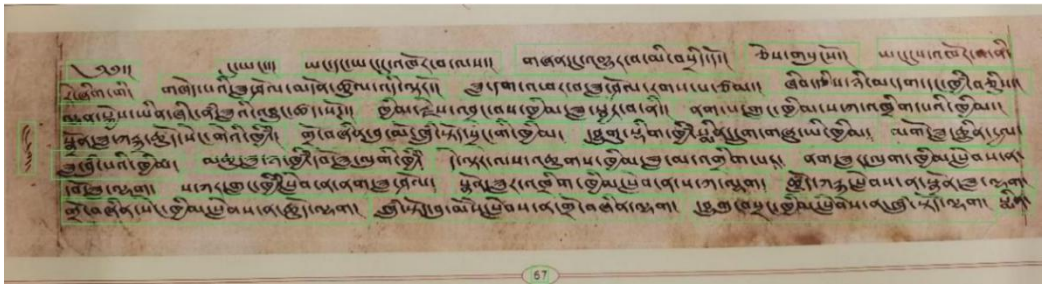
示例 1：



e) 一个文本行中，如果存在高度明显不同的文字，即使离得很近，也应标注为两个文本框。

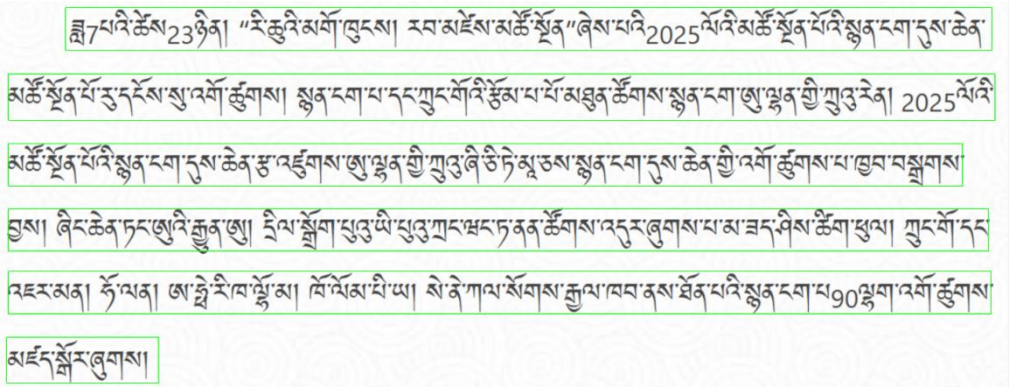
f) 一个文本框中只应包含一种字体类型的文字，不同字体类型应分开框。

示例 2：



g) 阿拉伯数字可当藏文，取决于阿拉伯数字所在的文本行类型，不必将阿拉伯数字单独标注出来。

示例 3：



h) 目前只考虑横排的情况，所有标注按照横排标注。如果样本中存在竖排样本，将样本区域裁剪后旋转再进行标注。

示例 4：



i） 针对重叠文字的标注：重叠的标上层的文字。有些样本如果分不出上下层，则样本先整体不标，挑出来单独放。

5.5 字体类型标注规则

5.5.1 字体类型

对字体类型标注时，应从以下给出的藏文字体类型中选取。

a） 乌金体（Uchen）：乌金体是藏文最常用的印刷体，字形端庄、笔画清晰，适用于正式文档、书籍印刷等场景。标注时需检查乌金体的字形完整，笔画粗细均匀。

b） 簇仁体（Tsugring）：簇仁体是一种手写风格字体，笔画流畅，常用于手抄文献或艺术设计。标注时需注意其连笔特征，检查连笔部分的连贯性。

c） 簇通体（Tsugtung）：簇通体是簇仁体的变体，笔画更为简洁，适用于快速书写或非正式场景。标注时需注意其简化笔画，避免误标。

d） 朱匝体（Drutsa）：朱匝体是一种草书风格字体，笔画简练，常用于快速书写或非正式场景。标注时需注意其草书特征，检查连笔部分的连贯性。

e） 柏簇体（Betsu）：包括净足柏簇体、毛足柏簇体、朱簇体。所谓净足柏簇体是指这类字体字腿末端像刀切的一样不带毛刺，形成末端齐茬。所谓毛足柏簇体是指其字长腿末端略带向下挑的墨迹。所谓朱簇体即柏簇体和朱匝体的混合体。在刻板印经技术出现之前，柏簇体很盛行。珍藏于萨迦寺内的许多写卷都是用柏簇体书写的。后世写卷也常用柏簇体。

f） 丘伊体（Chuyig）：“伊体”是噶丹颇章时期产生的一种字体。经过长期实践，形成了具有独特风格的“伊体”，常用于历史档案的书写。

g） 其它字体：若图文使用的字体不属于上述六种标准字体，则统一标注为“其它”。

5.5.2 标注方法

点击藏文字体对应的按钮，用相应颜色的文本框框出文字。

示例：



5.6 文字内容标注规则

5.6.1 藏文文本规则

对文字内容标注时，藏文文本的要求如下。

a) 编码要求：藏文文本标注应采用 GB/T 13000—2025 编码，不应使用其他藏文编码字符集。标注工具应具有显示与编辑 GB/T 13000—2025 藏文区字符的功能。标注时应检查藏文字符的书写顺序和拼写规则符合藏文语法规则。

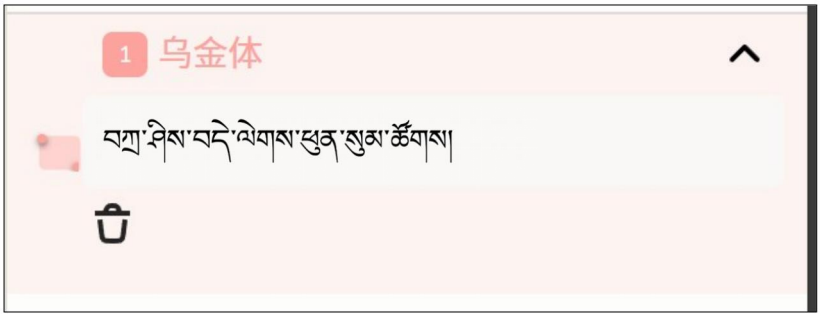
b) 字体显示：标注工具应正确渲染藏文字体，不出现乱码或显示错误。标注过程中应使用标准藏文字体，检查字形是否清晰、规范。

c) 多音节词处理：对于藏文多音节词，应检查每个音节的标注完整且符合拼写规则。若多音节词中存在无法辨认的部分，按整体标注为“&#”。

5.6.2 标注方法

在文本框对应的输入位置中填写对应的藏文文本。

示例：



5.6.3 注意事项

对文字内容标注时，注意事项如下。

a) 模糊文本处理：对于模糊不清、人眼无法辨识的文本，统一标注为“&#”。若文本部分模糊但可辨认，应根据可辨认部分进行标注，并在备注中说明模糊情况。
b) 残缺文本处理：对于文字残缺但可辨认的文本，正常标注可辨认部分。若残缺部分较多（超过 50%），则统一标注为“&#”。
c) 无法辨认字符处理：一个文本块中若存在无法辨认的字符，则该块的文字内容统一标注为“&#”。若文本块中仅个别字符无法辨认，但整体内容可推断，则正常标注并在备注中说明。
d) 特殊符号处理：若文本中包含标点符号、数字或其他非藏文字符，应按原文标注。对于无法辨认的特殊符号，标注为“&#”。
e) 多文种混合处理：若文本块中包含藏文与其他文种（如汉文、英文）混合内容，应按原文逐字标注。对于无法辨认的混合文种字符，标注为“&#”。
f) 标注一致性：同一文本块中，标注内容应保持一致，不应出现部分标注、部分忽略的情况。若文本块中存在多种情况（如部分可辨认、部分不可辨认），按整体情况处理。

5.7 标注格式示例


```
[
  {
    "ocr": "\/data\/upload\/1\/3d04cf38-tibetan.jpg",
    "id": 1,
    "bbox": [
      {
        "x": 0.31169602216505043,
        "y": 16.93548387096774,
        "w": 97.76865228577083,
        "h": 63.70967741935483,
        "rotation": 0,
        "original_w": 326,
        "original_h": 42
      }
    ],
    "label": [
      {
        "x": 0.31169602216505043,
        "y": 16.93548387096774,
        "w": 97.76865228577083,
        "h": 63.70967741935483,
        "rotation": 0,
        "labels": [
          "乌金体"
        ],
        "original_w": 326,
        "original_h": 42
      }
```

```
        ],
        "transcription": "བཀྲ་ཤིས་བདེ་ལེགས་ཕྱན་ཆུབ་ཚོགས། ",
        "annotator": 1,
        "annotation_id": 1,
        "created_at": "2025-03-11T07:24:52.669271Z",
        "updated_at": "2025-03-11T07:24:52.669271Z",
        "lead_time": 165.66099999999997
      }
]
```


5.8 标注需求变更流程

数据需求方提出需求变更时，应按以下流程。

a) 需求变更通知：数据需求方及时将变更需求以书面形式通知标注管理方，包括变更的具体内容和原因。

b) 评审变更：

1) 变更评审：标注管理方对变更需求进行详细评审，评估变更对标注任务的影响，包括对标注规则、任务可行性和资源需求的影响；

2) 可行性分析：确认变更后的需求是否可行，并评估所需的调整和额外工作量。

c) 更新标注任务：

1) 任务调整：根据评审结果更新标注任务，包括修改标注规则、调整时间安排和资源配置；

2) 文档更新：更新相关文档和标注规范（含本文件规定的标注规则及数据需求方的具体要求），检查确认所有参与者掌握最新的标注要求和流程。

d) 重新进入评估阶段：

1) 预标注测试：在更新后的标注任务上进行小规模的预标注测试，验证变更是否有效且操作性强；

2) 结果验收：将测试结果提交给数据需求方进行验收，根据反馈进行必要的调整。

e) 正式执行：在数据需求方确认更新后的标注任务无误后，正式启动变更后的标注任务。

5.9 标注环境要求

5.9.1 选择标注工具

选择适宜的标注工具（以SunOCR标注工具为例），完成数据标注任务。

5.9.2 选择标注场地

数据标注方应根据标注任务中必要的数据安全要求和项目需求，选择或搭建适宜的数据标注场地。

6 标注任务执行过程的要求

6.1 标注任务创建要求

6.1.1 任务包创建要求

任务包创建过程的要求如下：

a) 数据分组：数据需求方应将数据按类型、内容或标注难度合理分组，检查每个任务包中的数据是否具有一致性，以提高标注的准确性和效率；
b) 质量控制：任务包创建时应考虑标注质量的监控措施，设置质量检查点和反馈机制，以检验数据标注的准确性和一致性；
c) 任务说明：为每个任务包提供明确的标注说明和标注规范（含本文件规定的标注规则及数据需求方的具体要求），以指导标注人员进行正确的操作，减少标注错误。

6.1.2 任务账户创建要求

任务账户创建过程的要求如下：

a) 账户配置：根据标注人员的角色和职责创建账户，包括标注人员、审核人员和管理员，检查每个账户是否具备适当的权限和功能；
b) 权限管理：为不同角色分配相应的权限，包括数据访问、标注编辑和质量审核权限，以保护数据安全；
c) 账户规则：制定账户使用规则，包括登录安全、操作规范和数据保护要求。

6.2 标注任务分发要求

6.2.1 标注管理方应根据标注任务的具体要求、环境和工具选择合适的任务分发方式。
常见的方式包括：

a) 主动领取：标注人员自选任务；
b) 系统自动分发：系统根据任务需求和人员状态自动分配。

6.2.2 标注管理方应制定优化策略，平衡标注质量和速度。

示例：针对任务复杂度设置合理的优先级。

6.2.3 标注任务分发应采取以下保障措施：

a) 安全管理制度：制定并实施标注任务分发的安全管理制度，检查分发过程的安全性和合规性，包括数据传输安全、访问控制和权限管理等方面；
b) 日志记录与监控：
   1) 明确任务分发日志的内容，包括分发时间、任务分配情况和操作记录；
   2) 实施实时监控，以便及时发现并解决可能的安全问题或异常情况。

附 录 A
（资料性）
标注人员管理要求

A.1 标注人员范围

数据标注方应根据标注任务内容中规定的标注人员资质和专业需求，选拔符合要求的人员进入标注人员培训环节。标注人员应具备必要的背景知识和技能，能够胜任指定的标注任务。

A.2 标注人员培训

数据标注方应制定并实施针对性的岗前培训计划，根据标注任务说明对标注人员进行系统培训。培训内容应包括标注标准、工具使用方法、数据处理流程等。培训后应进行标注能力考试，只有通过考试的人员方可参与实际标注任务。

A.3 标注人员能力档案

数据标注方应建立详细的标注人员能力档案，记录每位标注人员的培训经历、考试成绩、任务参与情况以及标注质量评价。承担任务结束后，应及时更新标注人员的能力档案，用于进行标注人员能力评估与标注质量追踪，并作为后续培训和任务分配的依据。

参 考 文 献

[1] GB/T 42755—2023 人工智能 面向机器学习的数据标注规程