# **Project Leaders**

Jiefeng Xie

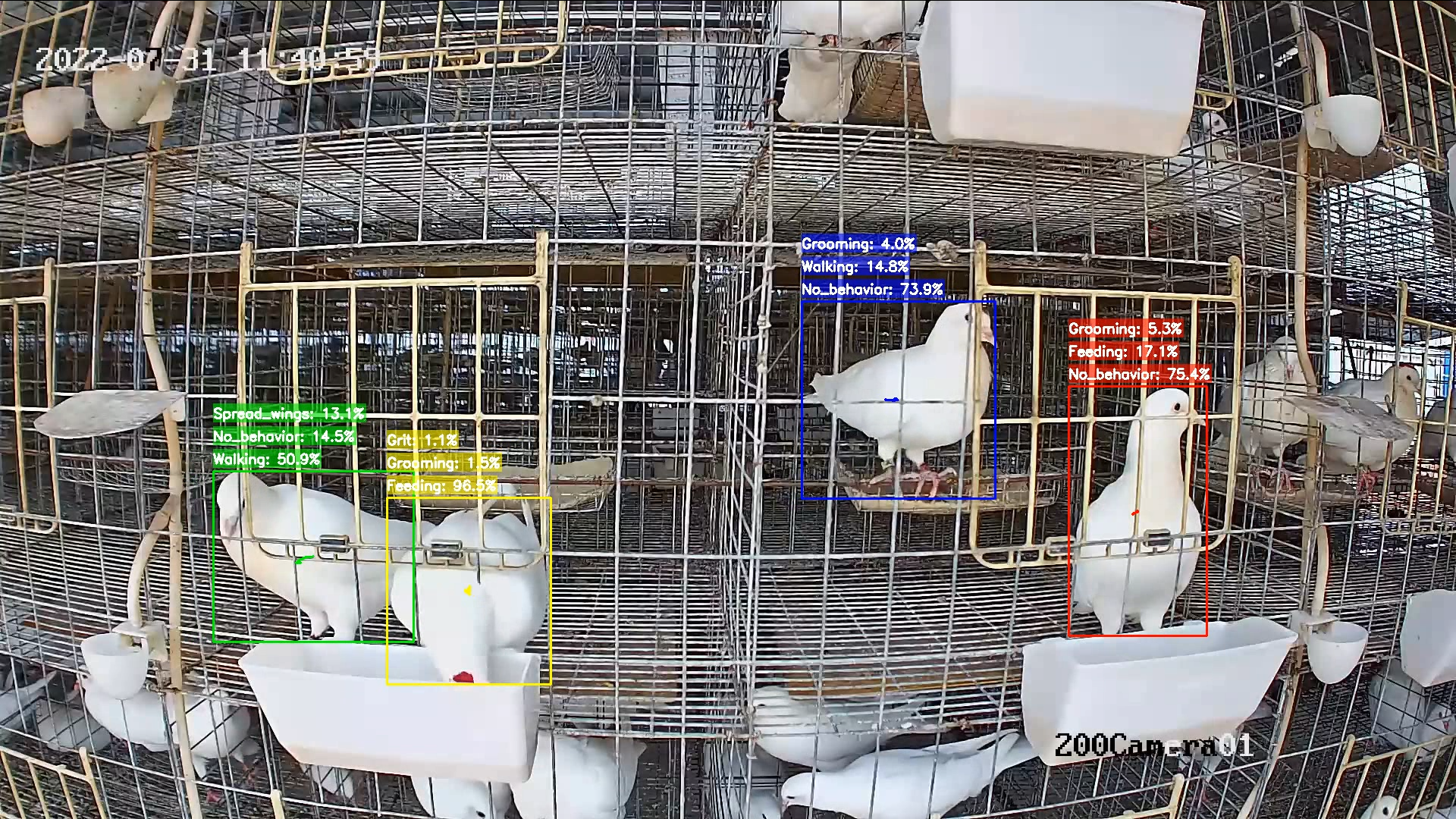
Nuoer Long

Chengpeng Xiong

# Pigeon Farming Behavior Monitoring and Analysis

Traditional pigeon farming heavily relies on manual observation and experience-based judgment, making it difficult to monitor the health status of the flock and detect abnormal behaviors in real-time. This leads to delays in disease prevention, feed waste, and animal welfare concerns. With the development of large-scale farming, the precise identification of fine-grained pigeon behaviors, such as feeding, fighting, and grooming, along with the analysis of health risks, has become a core demand to improve farming efficiency and sustainability. However, existing methods are unable to perform both behavior recognition and tracking based on video simultaneously, making it impossible to measure individual activity levels, analyze social relationships, and accurately capture the behavior of each pigeon.

To address this issue, the project employs an integrated multi-object tracking and behavior recognition algorithm, enabling the real-time tracking and fine-grained behavior identification of pigeons in video streams. In addition, we have constructed a dedicated pigeon dataset to support the training and validation of the algorithm. This technology effectively reduces costs, increases efficiency, improves pigeon welfare, enhances reproductive rates and market quality, and promotes the realization of precision farming.



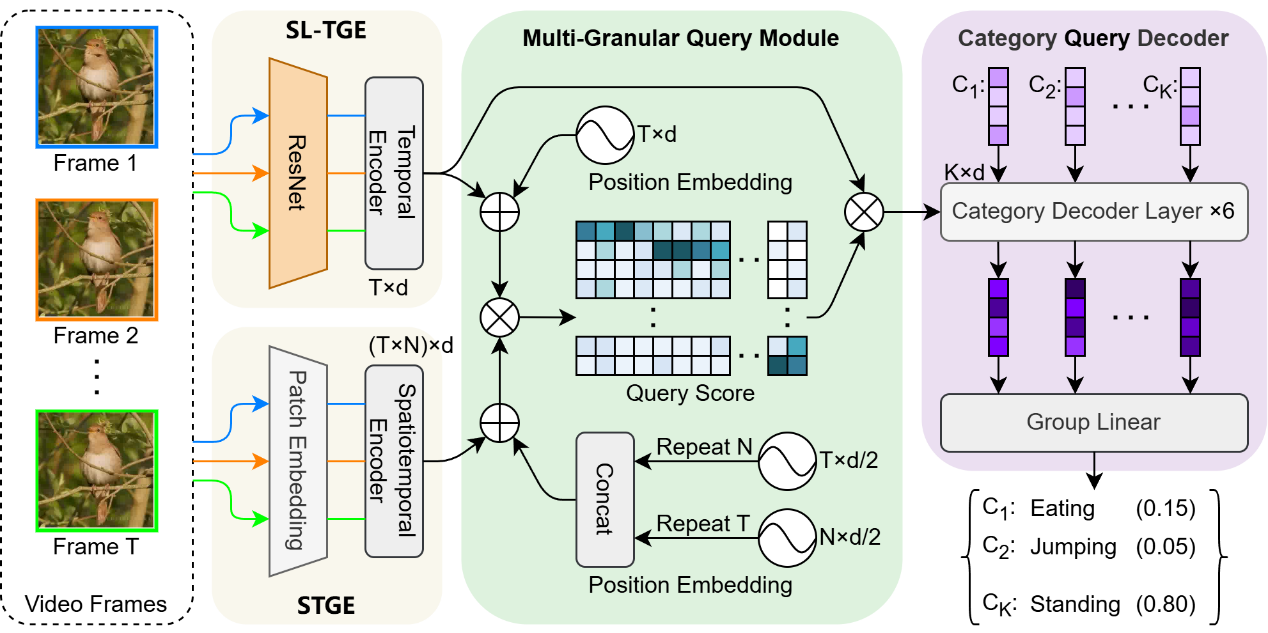
Pigeon breeding behavior monitoring

# Action Recognition

Behavior recognition is an important and challenging task, particularly for networks aiming to provide a unified solution for various actors, including humans and animals. Such data contains complex temporal relationships and biological categories with different morphologies. Most current studies usually focus on a single biological category. This makes most behavior recognition networks unable to meet species classification requirements.

To overcome these limitations, we construct a query-based multi-granularity behavior recognition network. In the spatial dimension, we construct two different granularity features: fine-grained features that focus on the local morphology of the organism and coarse-grained features that focus on the overall appearance of the organism. Central to our approach is a Multi-Granularity Query Module, which processes these features similarly in the temporal dimension. This ensures that the two granularity features have a natural alignment relationship during the query stage. Additionally, we construct a set of learnable embedding vectors, called category query features, where each corresponds to a potential action category.

Through multiple interactions with multi-granularity features, the behavior features in the video are mapped to the corresponding category query features. We achieved a new state-of-the-art (SOTA) performance on the behavior recognition task of the Animal Kingdom dataset and demonstrated good performance on the Charades dataset. Our experiments prove that our method is not only suitable for uniformly solving behavior recognition tasks for various actors but also scalable for behavior recognition tasks of specific biological categories.



Query-based multi-granularity behavior recognition network architecture

# Video Grounding

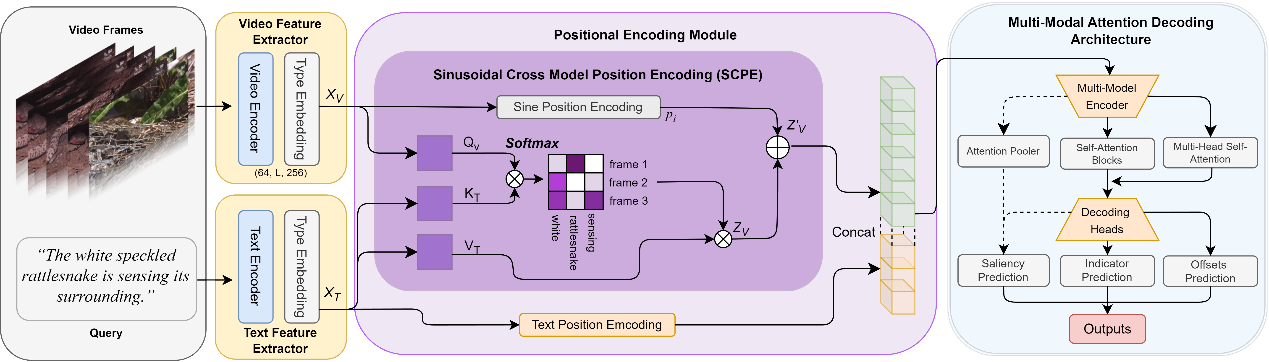
Video grounding is a multi-modal task in computer vision aimed at identifying segments of a video that corresponds to a given textual description. With the explosive growth of videos on the Internet, video grounding is attracting more and more interest and plays an important role in applications such as video retrieval, content analysis and surveillance in people’s daily lives.

A screenshot of a video game

AI-generated content may be incorrect.

Illustration of Video Grounding Task: Localizing activities via language queries

Transformers are widely used in video localization tasks. However, most position encoding methods used in Transformers treat all blocks equally, thus masking key blocks that are crucial for accurate localization. To address this challenge, we propose a new network. By combining sinusoidal cross-modal position encoding (SCPE) and a dual-path modal fusion mechanism, it aims to improve the alignment and temporal understanding between video and text descriptions. The SCPE method dynamically adjusts attention to focus on video frames related to text queries, thereby improving the accuracy of video localization, especially when dealing with complex natural environments and animal behaviors.



SCPE network architecture

# Award

**ICME Grand Challenge 2024 First Place**

A certificate of achievement with text and images

AI-generated content may be incorrect.A certificate of achievement with text and images

AI-generated content may be incorrect.