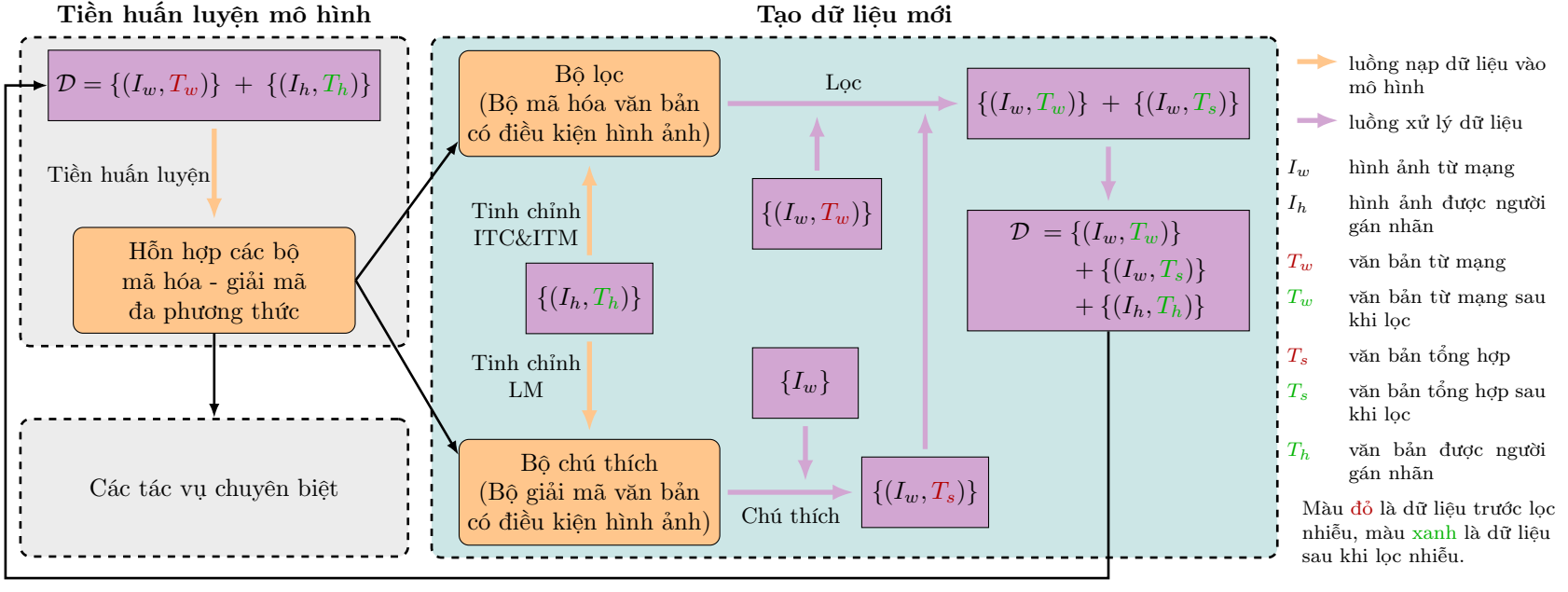
Tiền huấn luyện mô hình
$\mathcal{D} = \{(I_w, T_w)\} + \{(I_h, T_h)\}$
Tiền huấn luyện
Hỗn hợp các bộ mã hóa - giải mã đa phương thức
Các tác vụ chuyên biệt
Tạo dữ liệu mới
Bộ lọc (Bộ mã hóa văn bản có điều kiện hình ảnh)
Tinh chỉnh ITC&ITM
$\{(I_h, T_h)\}$
Tinh chỉnh LM
Bộ chú thích (Bộ giải mã văn bản có điều kiện hình ảnh)
Lọc
$\{(I_w, T_w)\}$
$\{I_w\}$
Chú thích
$\{(I_w, T_s)\}$
$\{(I_w, T_w)\} + \{(I_w, T_s)\}$
$\mathcal{D} = \{(I_w, T_w)\} + \{(I_w, T_s)\} + \{(I_h, T_h)\}$
luồng nạp dữ liệu vào mô hình
luồng xử lý dữ liệu
I_w hình ảnh từ mạng
I_h hình ảnh được người gán nhãn
T_w văn bản từ mạng
T_w văn bản từ mạng sau khi lọc
T_s văn bản tổng hợp
T_s văn bản tổng hợp sau khi lọc
T_h văn bản được người gán nhãn
Màu đỏ là dữ liệu trước lọc nhiễu, màu xanh là dữ liệu sau khi lọc nhiễu.