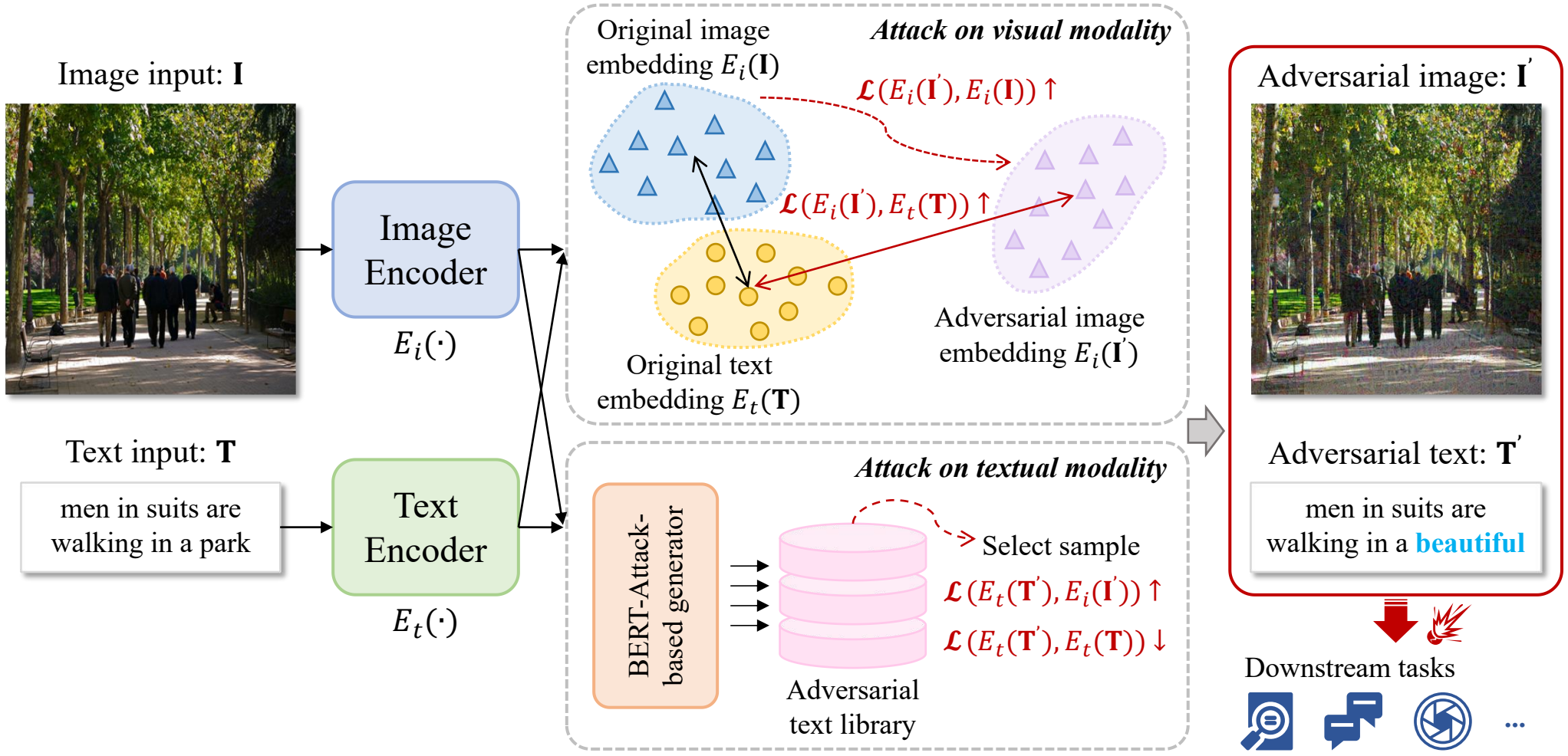
Image input: **I**
Image Encoder
$E_i(\cdot)$
Text input: **T**
men in suits are walking in a park
Text Encoder
$E_t(\cdot)$
Original image embedding $E_i(\mathbf{I})$
Attack on visual modality
$\mathcal{L}(E_i(\mathbf{I}'), E_i(\mathbf{I})) \uparrow$
$\mathcal{L}(E_i(\mathbf{I}'), E_t(\mathbf{T})) \uparrow$
Adversarial image embedding $E_i(\mathbf{I}')$
Original text embedding $E_t(\mathbf{T})$
Attack on textual modality
BERT-Attack-based generator
Select sample
$\mathcal{L}(E_t(\mathbf{T}'), E_i(\mathbf{I}')) \uparrow$
$\mathcal{L}(E_t(\mathbf{T}'), E_t(\mathbf{T})) \downarrow$
Adversarial text library
Adversarial image: **I**′
Adversarial text: **T**′
men in suits are walking in a **beautiful**
Downstream tasks
...