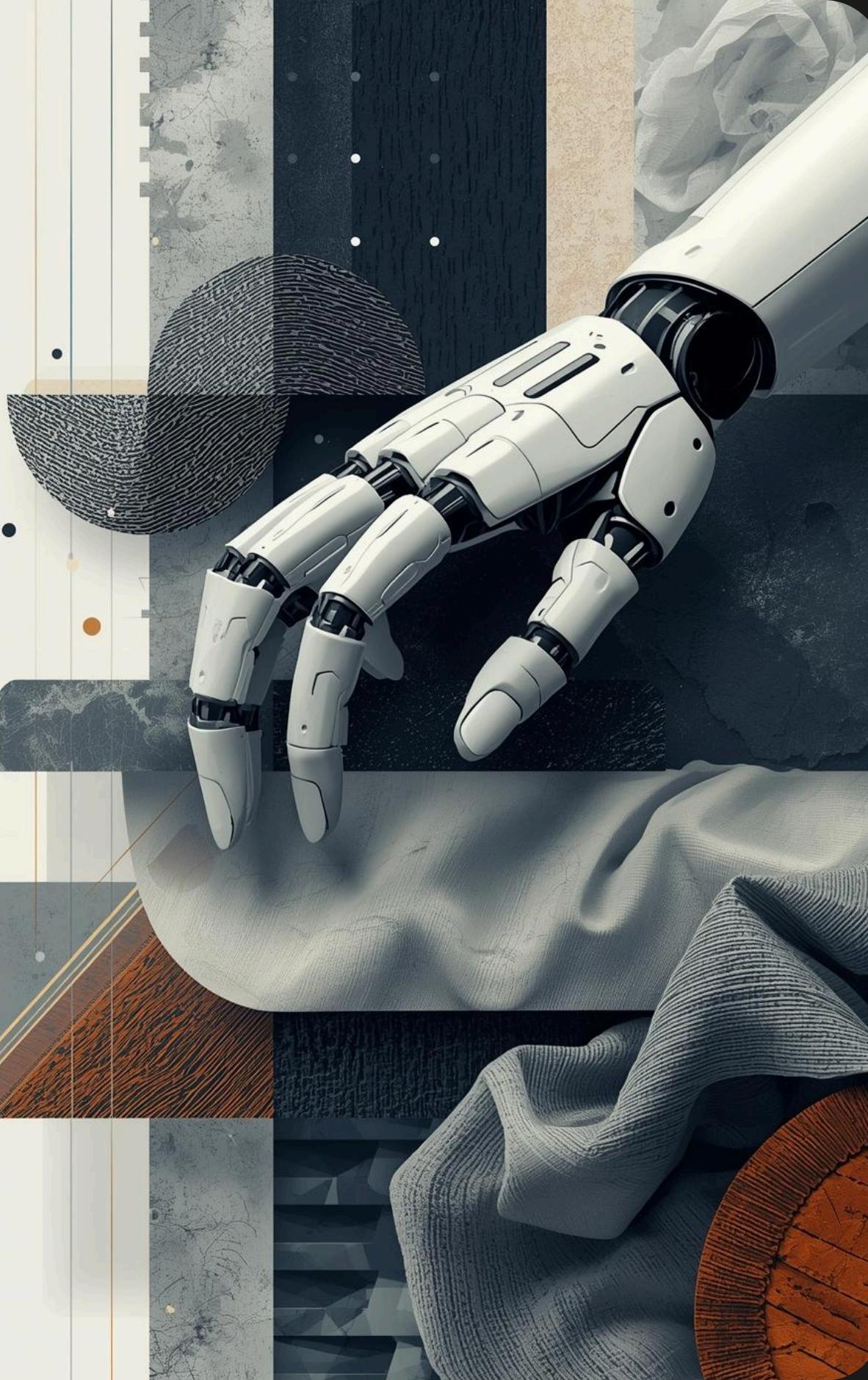


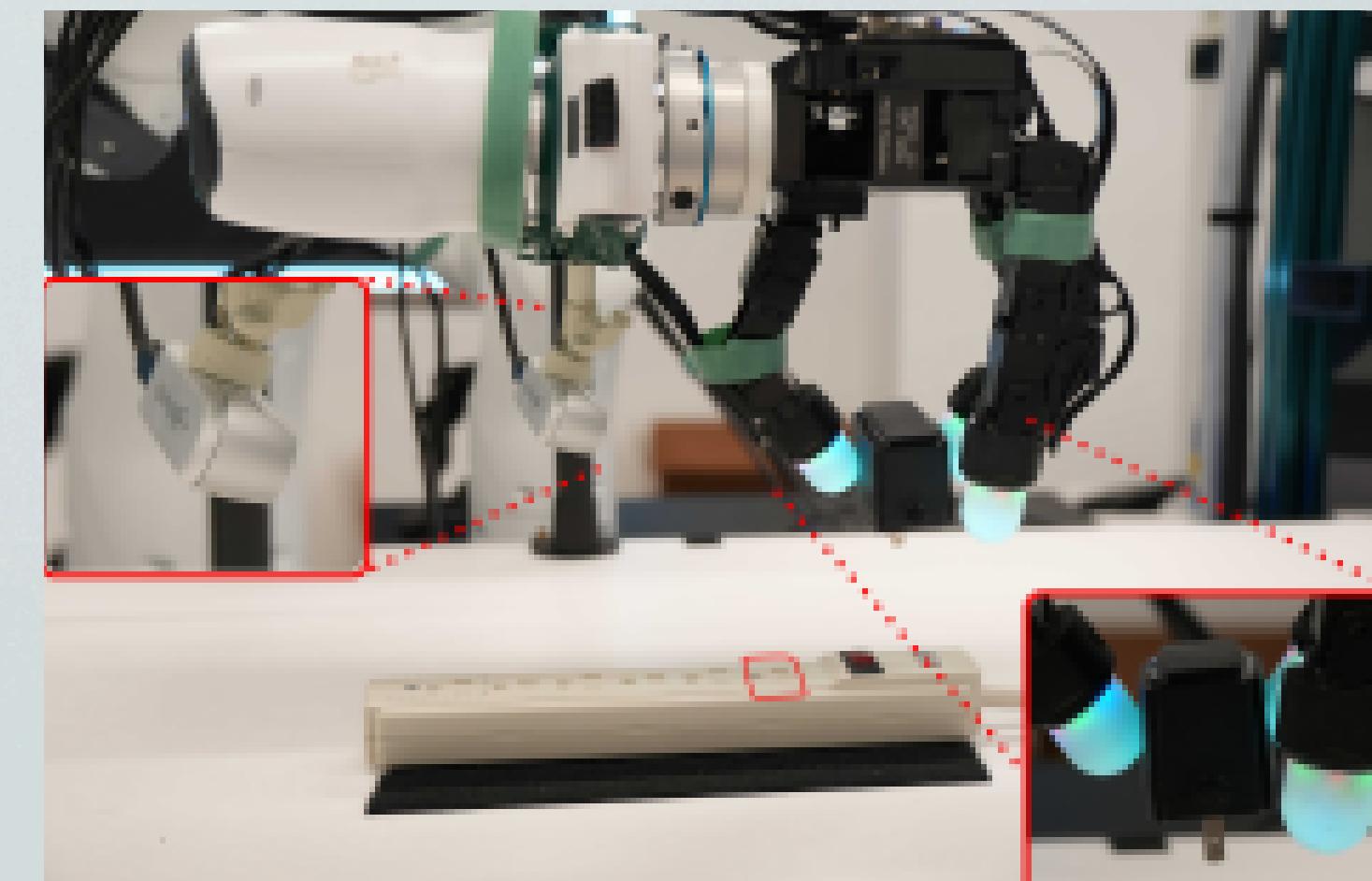
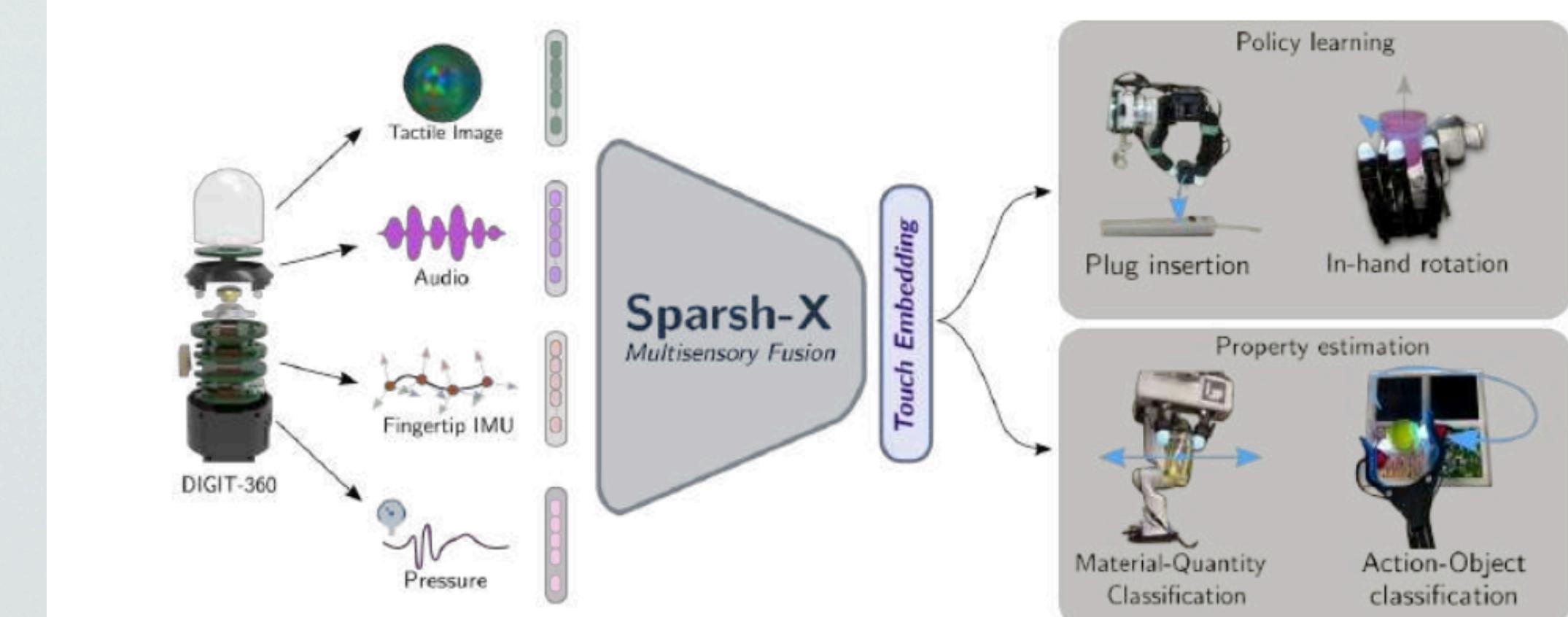
Tactile Beyond Pixels

Multisensory Touch Representations for Robots

Kamil Eray Gündüz
Cem Keleş

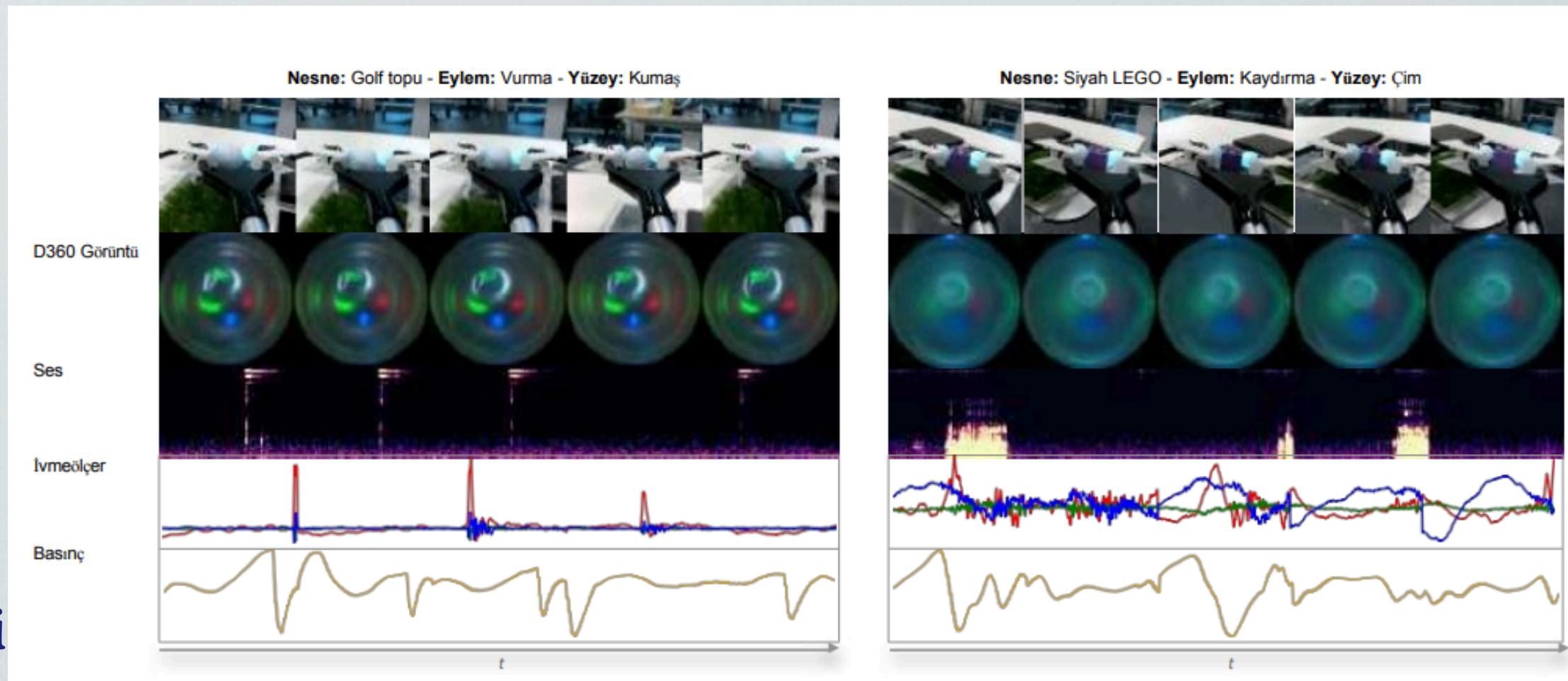


- Dokunmanın Doğası: İnsanlar dokunmayı çok modlu (multimodal) algılar: Titreşim, basınç, hareket, sıcaklık.
- Mevcut Durum: Robotikte dokunma genellikle sadece "görsel" (Vision-based tactile sensors, örn: GelSight) olarak ele alınıyor.
- Eksiklik: Sadece görsel veri, sürtünme, sertlik veya kayma gibi dinamik özellikleri tam olarak yakalayamaz

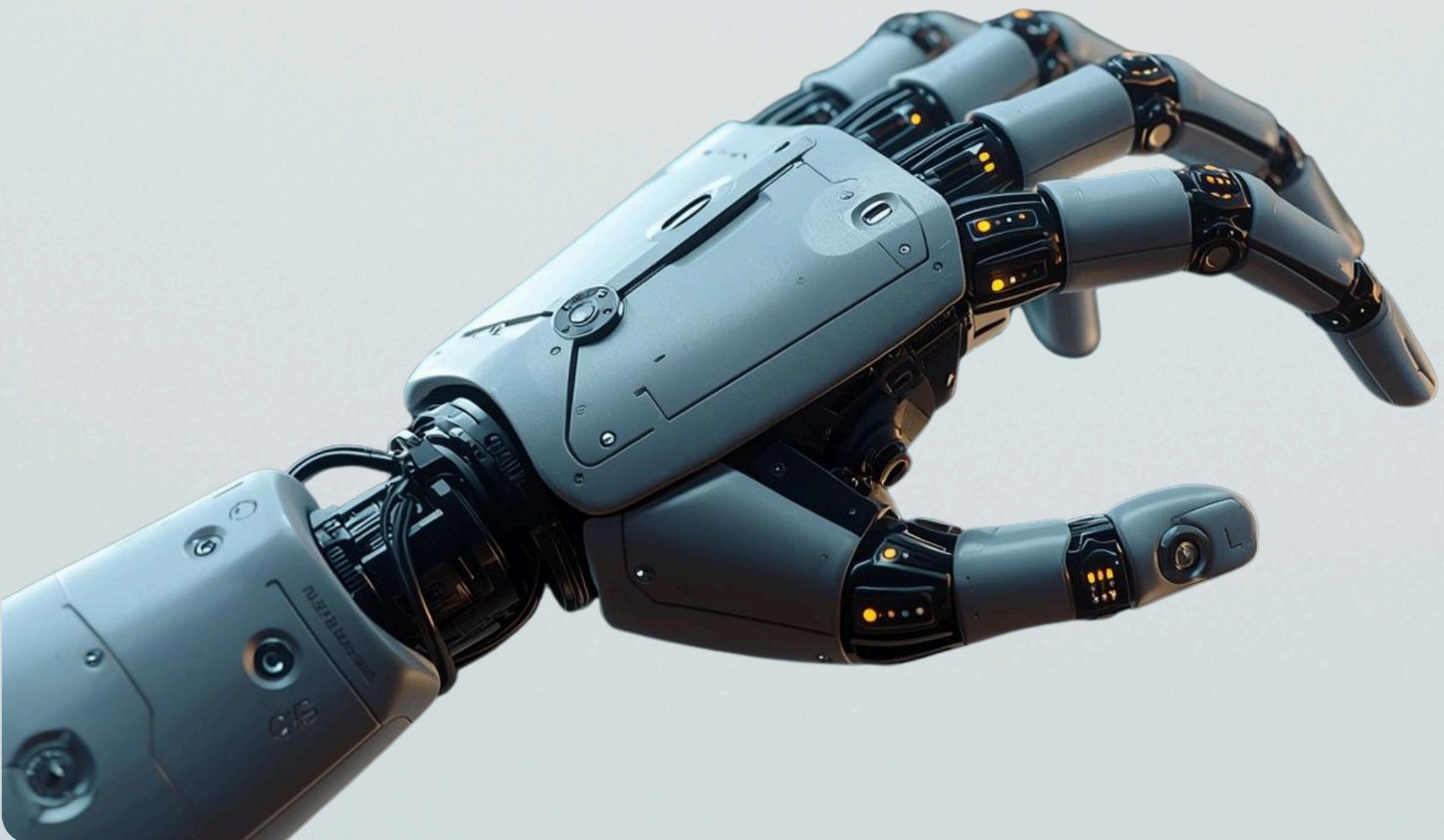


Donanım ve Veri Yapısı (Digit-360)

- Sensör: Digit-360 (Parmak ucu formunda).
- Girdiler (Inputs):
- Görüntü: 30 fps, elastomer deformasyonu.
- Ses (Audio): 48 kHz, temas mikrofonları (yüksek frekanslı titreşimler).
- IMU (Fingertip IMU): 400 Hz, 3 eksenli ivmeölçer (hareket ve sarsıntı).
- Basınç: 200 Hz, statik kuvvet değişimleri .

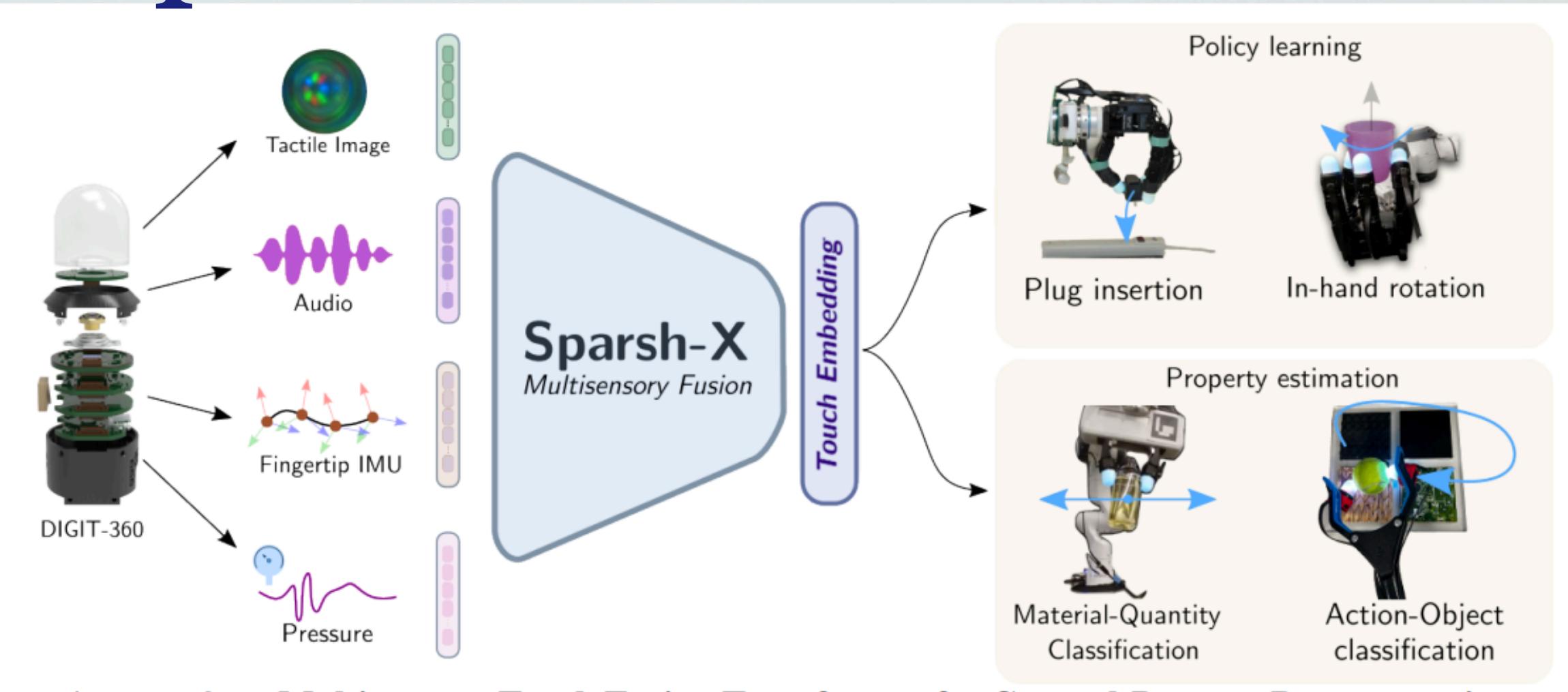


Çözüm: Sparsh-X Nedir?

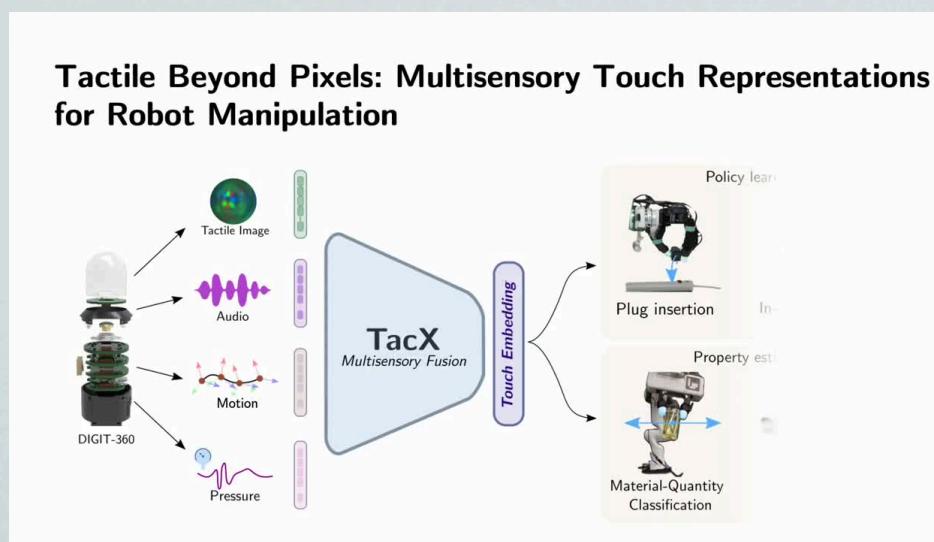


- Görüntü, Ses, Hareket (IMU) ve Basıncı birleştiren ilk genel amaçlı dokunma omurgası (backbone).
- Sensör Teknolojisi: Digit 360 sensörü kullanılarak geliştirilmiştir.
- Temel Yenilik: Farklı duyusal sinyalleri ortak bir gizli uzayda (latent space) birleştirir.

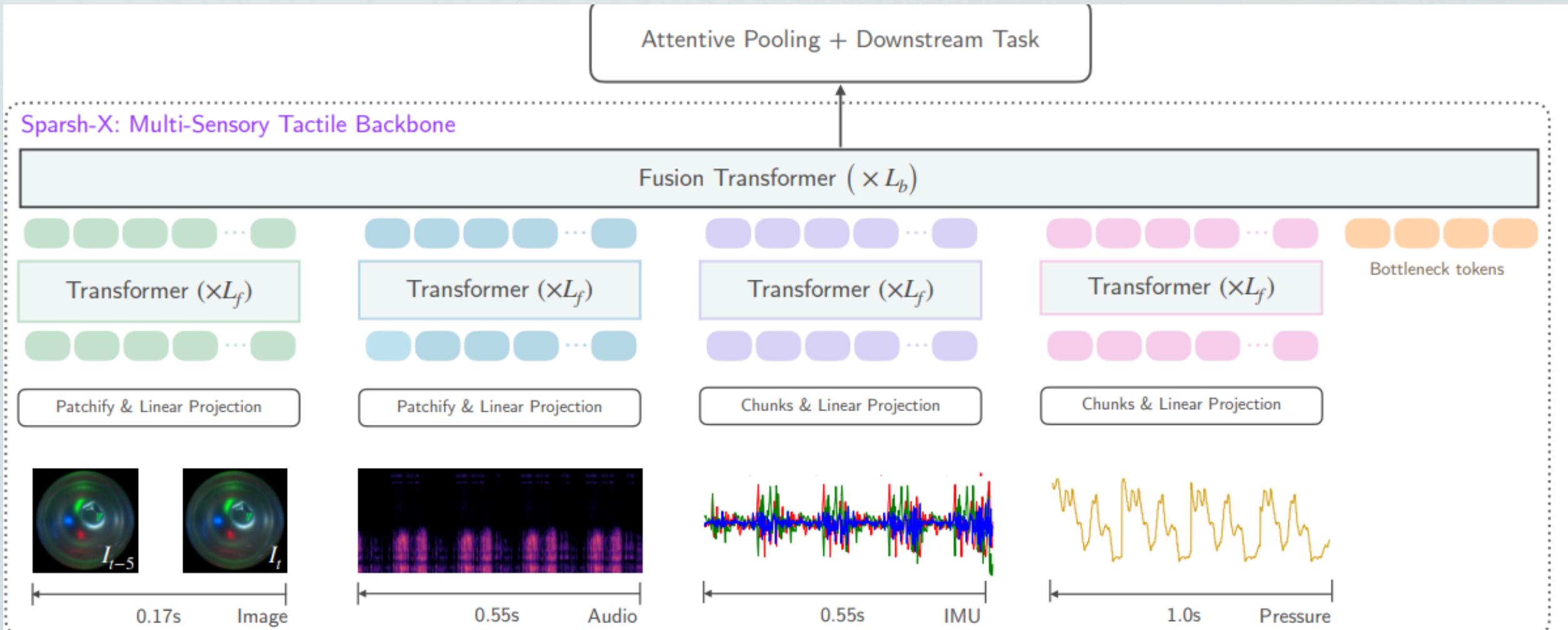
Çözüm: Sparsh-X Model Mimarisi



- Taban Model: Vision Transformer (ViT)
- Yapı: $L = 12$ katman (Layer).
- $L_f = 8$ Katman: Tek modlu (Uni-modal) bağımsız işleme.
- $L_b = 4$ Katman: Çok modlu (Multi-modal) füzyon.
- Tokenizasyon: Her veri tipi (yama) birer 'token' (vektör) haline getirilir.

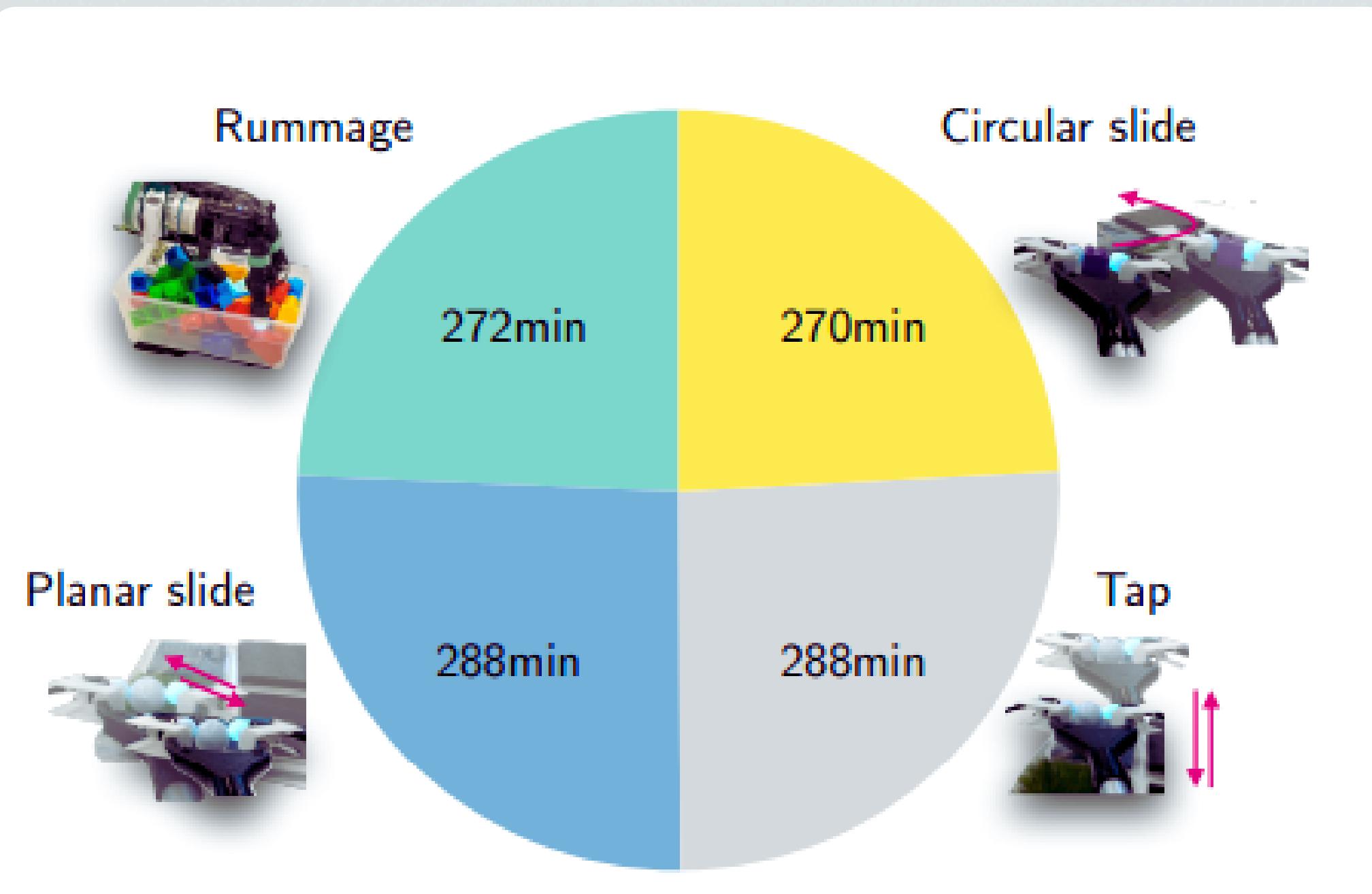


Bottleneck Attention ve Füzyon Mekanizması



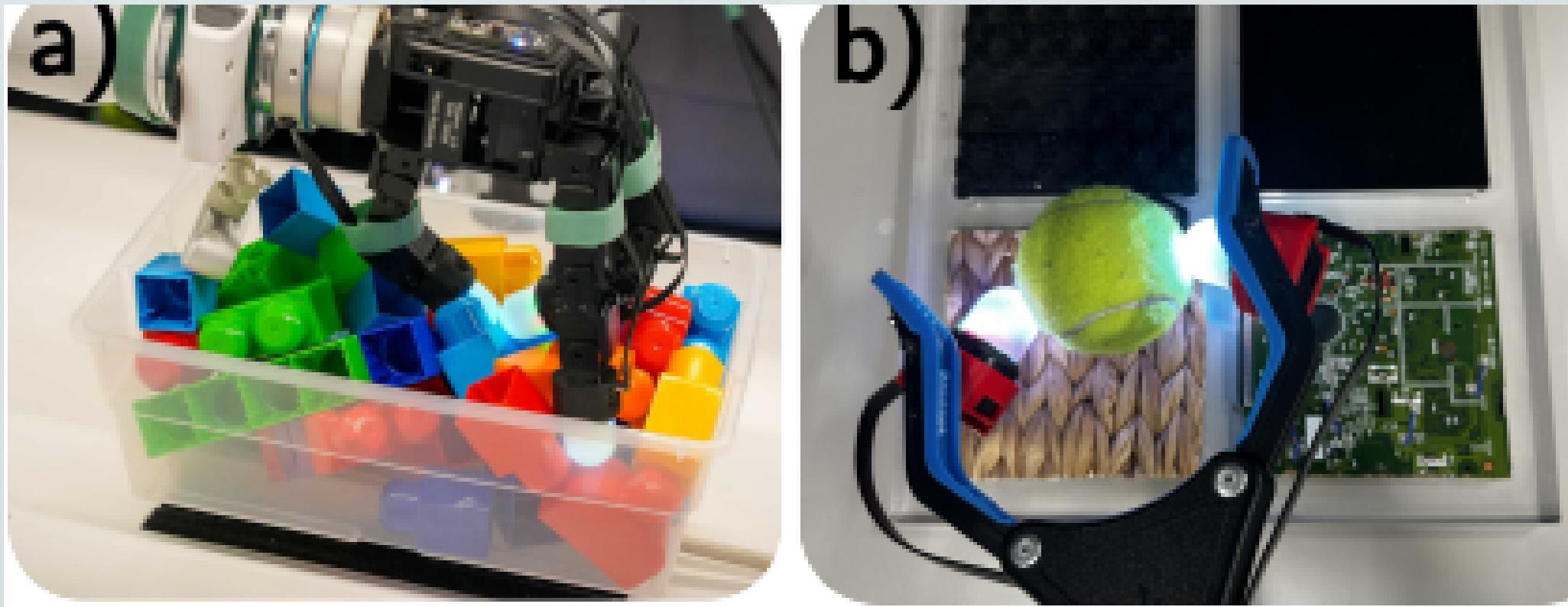
- Neden? Tam dikkat (Full Self-Attention) hesaplama maliyeti $O(N^2)$ 'dir (çok pahalı).
- Çözüm: Bottleneck Attention¹².
- Mekanizma:
- Her modaliteye $B=4$ adet "Darboğaz Tokeni" eklenir¹³.
- Bilgi akışı sadece bu tokenler üzerinden gerçekleşir.
- Füzyon katmanlarında bu tokenlerin ortalaması alınarak paylaşılır¹⁴

Kendi Kendine Denetimli Öğrenme (SSL)



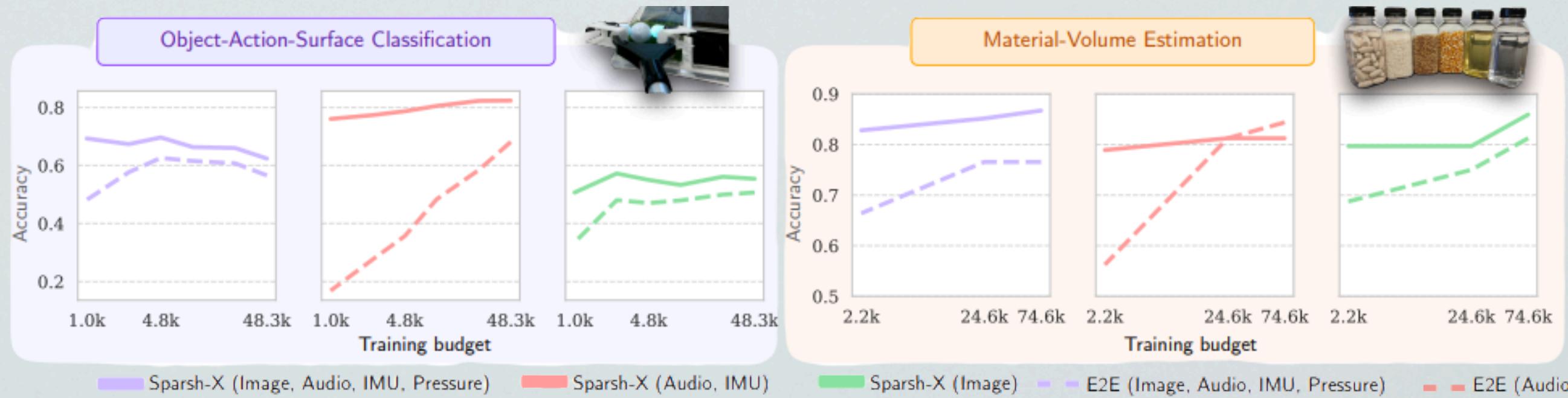
- Veri Seti: ~1 Milyon kare, etiketlenmemiş veri (Unlabeled).
- Yöntem: Masked Image Modeling (MIM) + Teacher-Student (DINOv2 benzeri).
Eğitim:
- Öğrenci (Student): Maskelenmiş (%50-90 kayıp) veri görür.
- Öğretmen (Teacher): Tam veriyi görür.
- Hedef: Öğrenci, öğretmenin çıkardığı özellikleri tahmin etmeye çalışır (Cross-Entropy Loss).

Veri Toplama Platformları



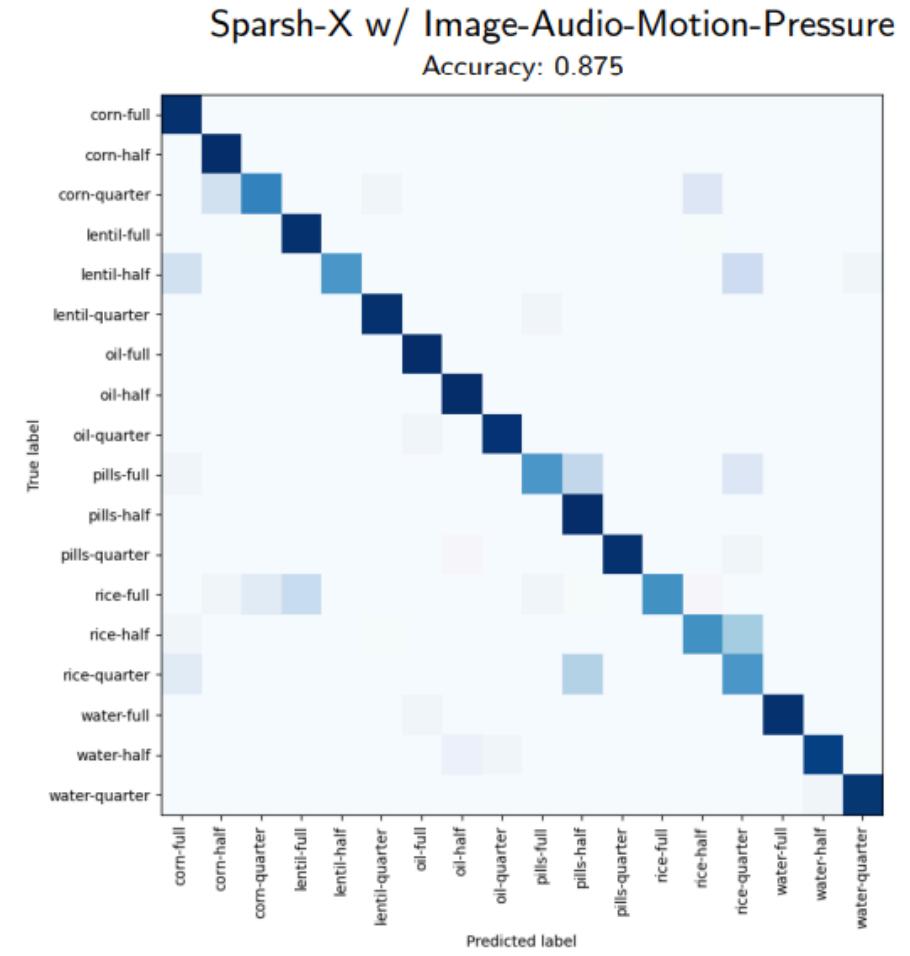
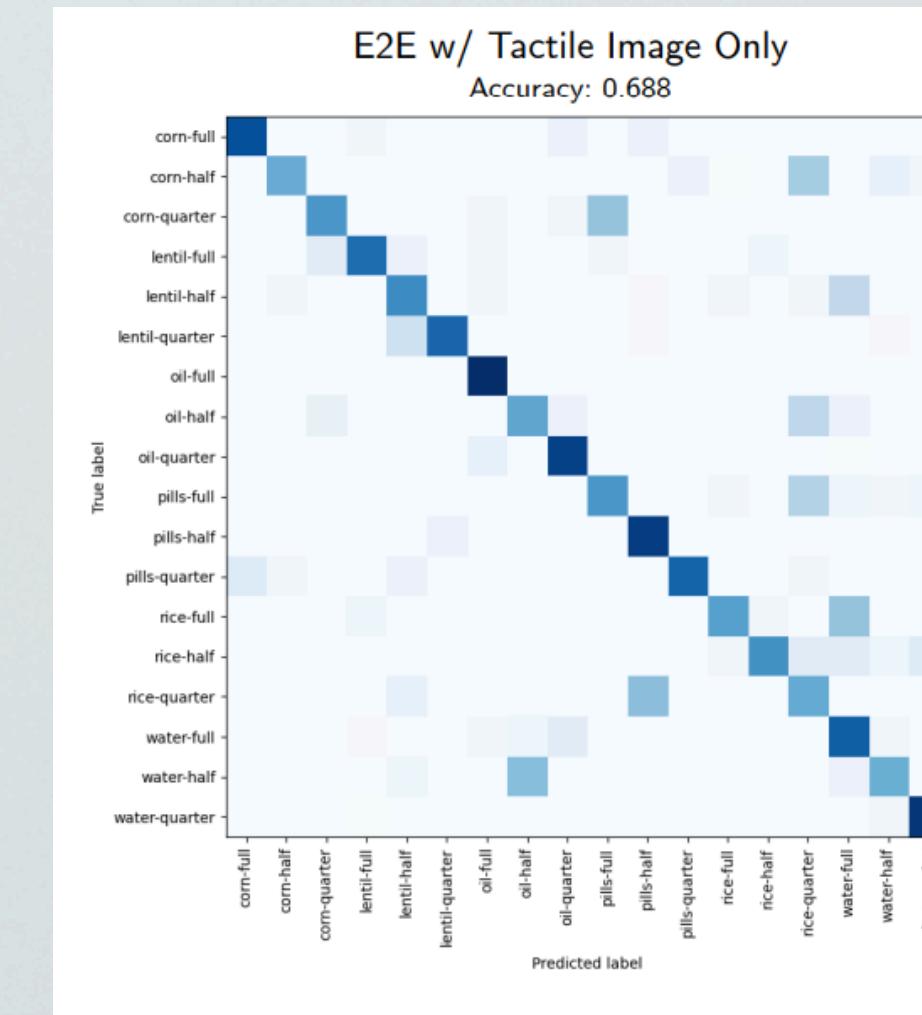
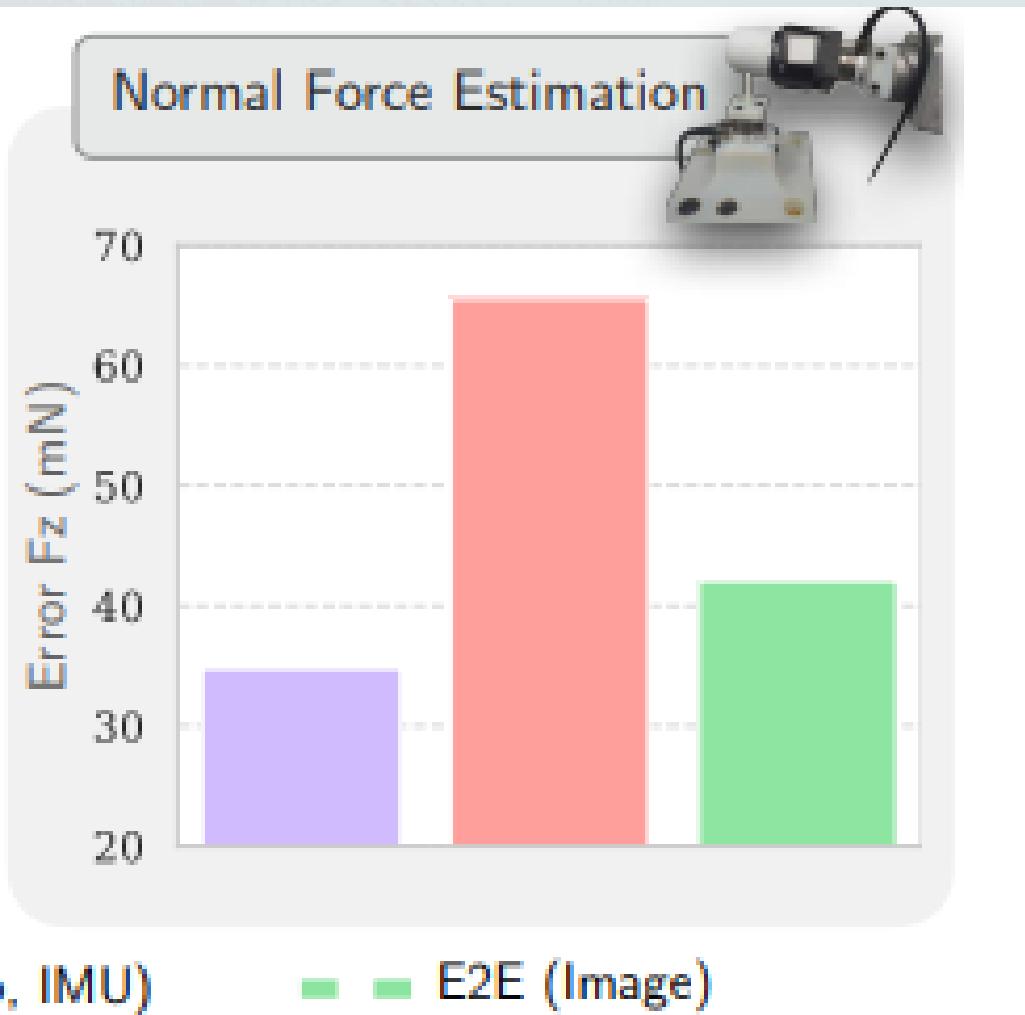
- Platform 1: Allegro Hand (Çok parmaklı el) – Karmaşık manipülasyon.
- Platform 2: Mobil Toplayıcı (Mobile Picker) – Farklı yüzeylerde sürünme/doku verisi.
- Çeşitlilik: Rummage (karıştırma), kaydırma, dokunma eylemleri.

Fiziksel Özelliklerin Anlaşılması



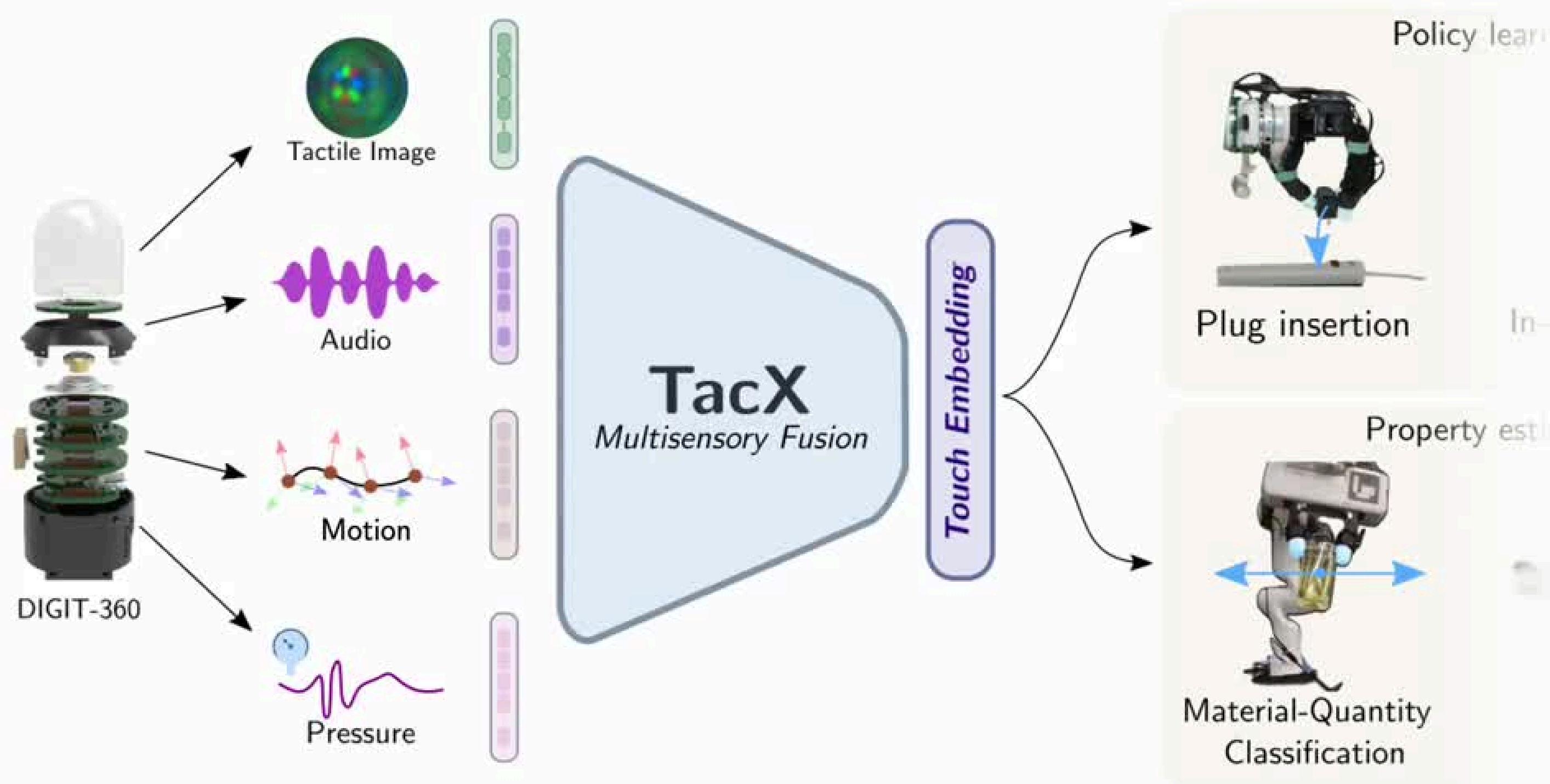
- Sparsh-X (Dondurulmuş/Frozen) kullanılarak yapılan testler.
- Görev 1: Nesne - Eylem - Yüzey Sınıflandırması.
- Sonuç: Çoklu modalite (Ses+IMU), sadece görüntüye göre %13 artış sağlar.
- Neden? Görüntü sürütmeyi göremez, ama mikrofon sürüünme sesini duyar.

Malzeme Miktarı ve Kuvvet Tahmini

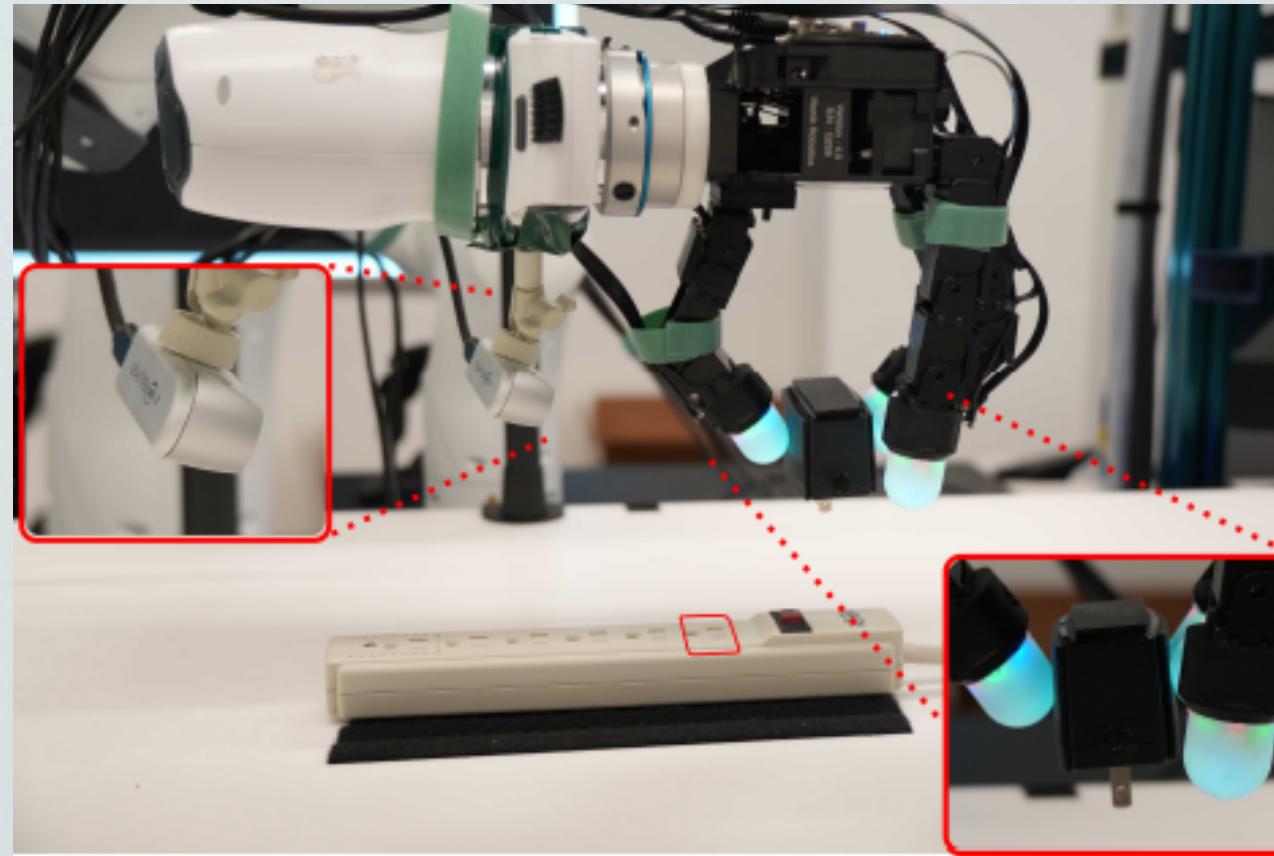


- Görev: Şişe sallayarak içindeki maddeyi (Pirinç, Su, Yağ) ve miktarını anlama
- Başarı: Sparsh-X, uçtan uca (E2E) modellere göre %20.5 daha başarılı.
- Kuvvet Tahmini: Normal kuvvet hatası 35 mN'a kadar düşmüştür.

Tactile Beyond Pixels: Multisensory Touch Representations for Robot Manipulation

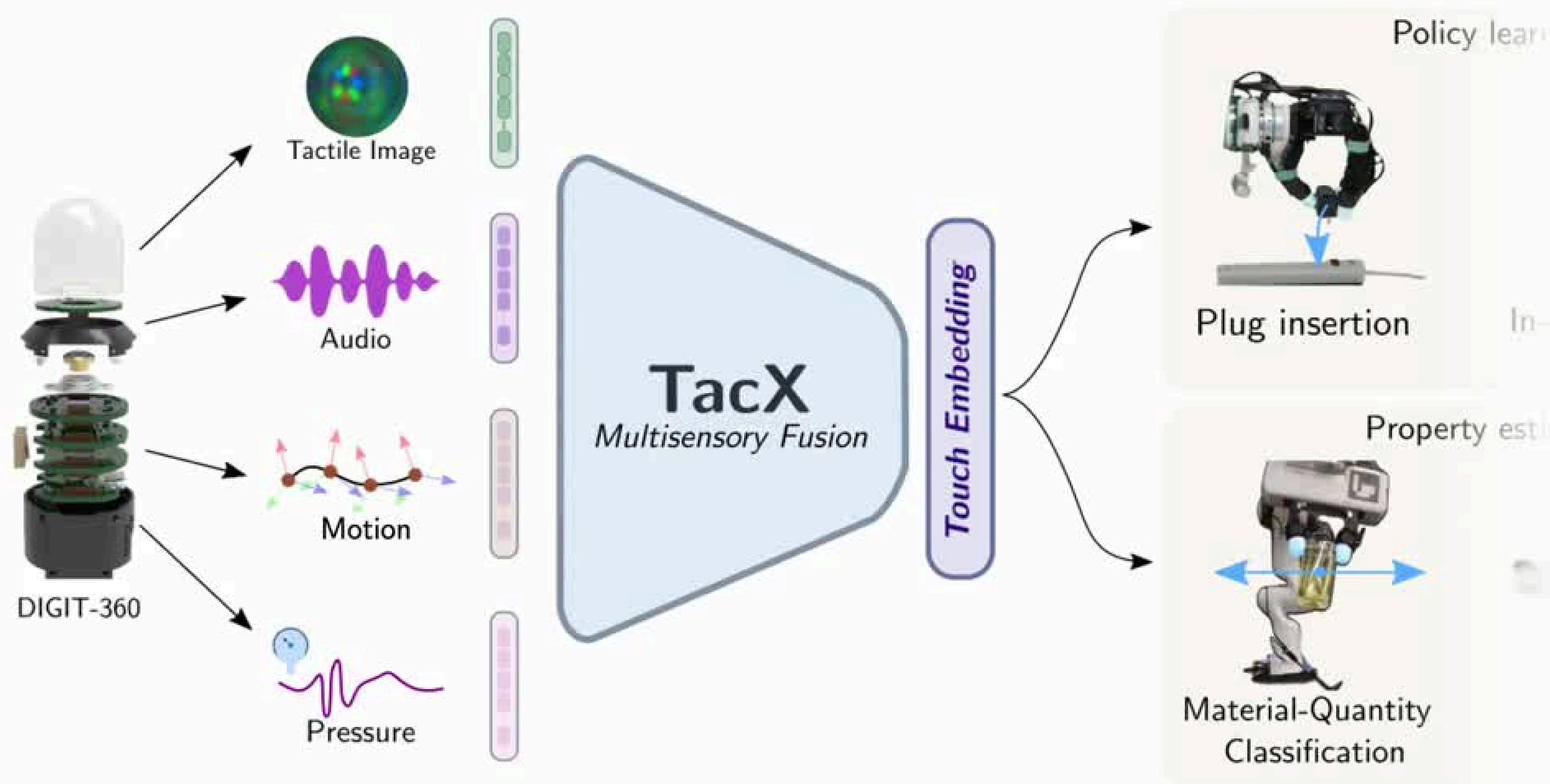


Robot Politika Öğrenimi 1: Fiş Takma (Plug Insertion)

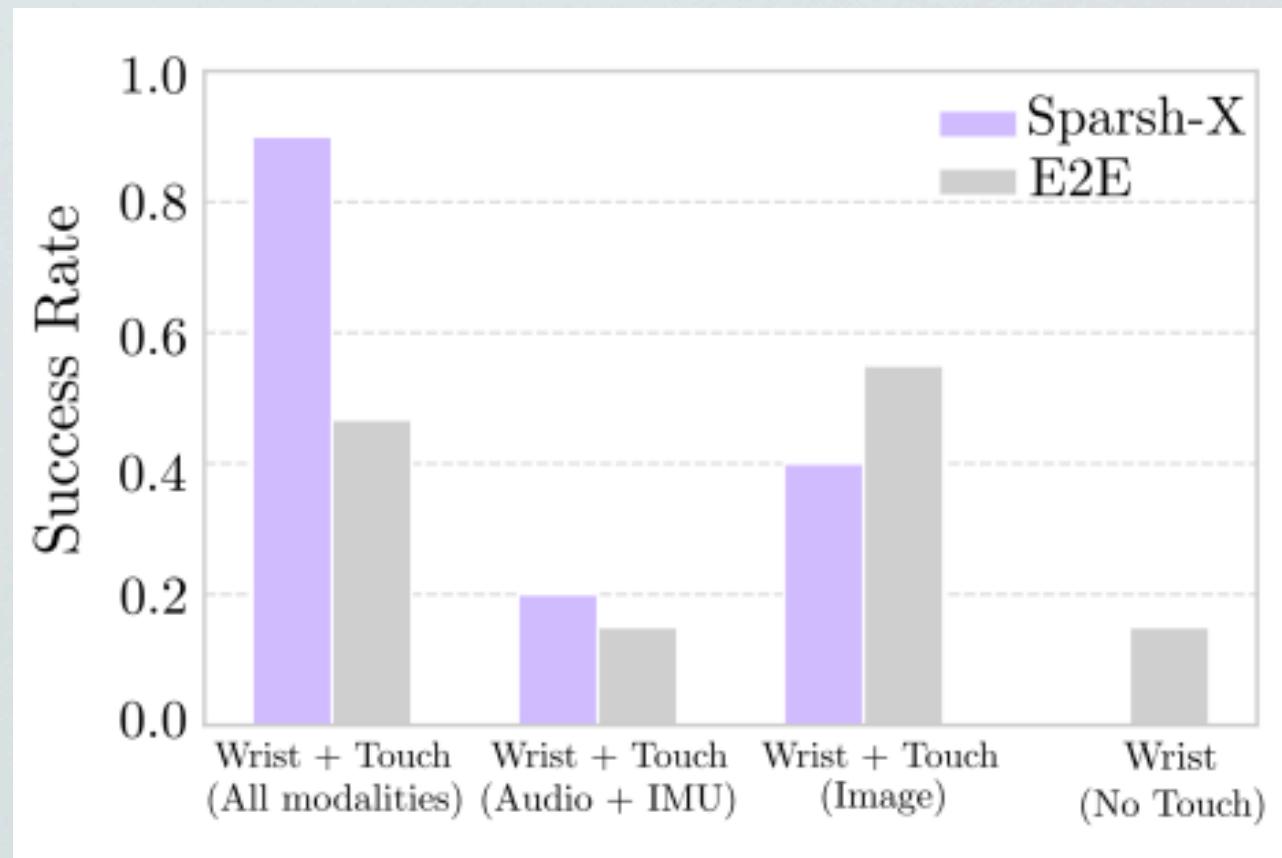


- Görev: Hassas toleranslı USB/Fiş takma işlemi.
- Yöntem: Taklit Öğrenme (Imitation Learning - ACT mimarisi).
- Girdi: Bilek Kamerası + Sparsh-X (3 parmak ucu).
- Zorluk: "Visual Aliasing" (Kamera derinliği algılayamaz ve delikleri yanlış hizalar).

Tactile Beyond Pixels: Multisensory Touch Representations for Robot Manipulation

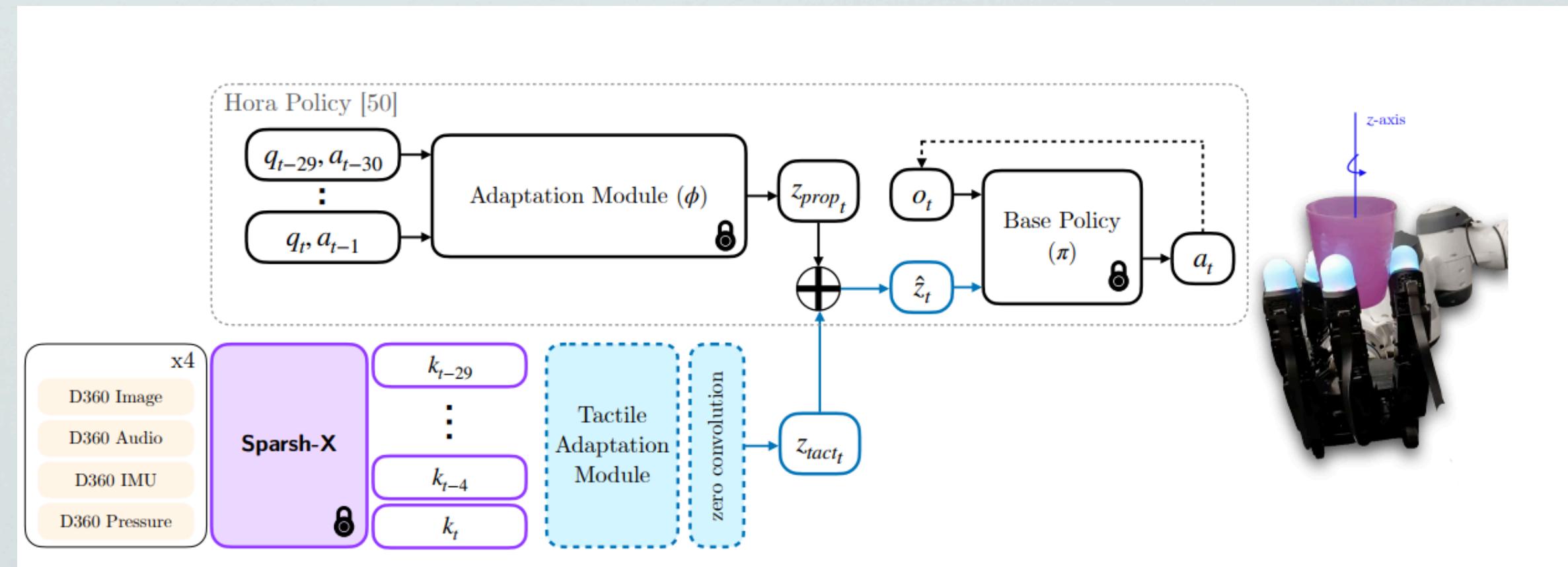


Fış Takma Sonuçları



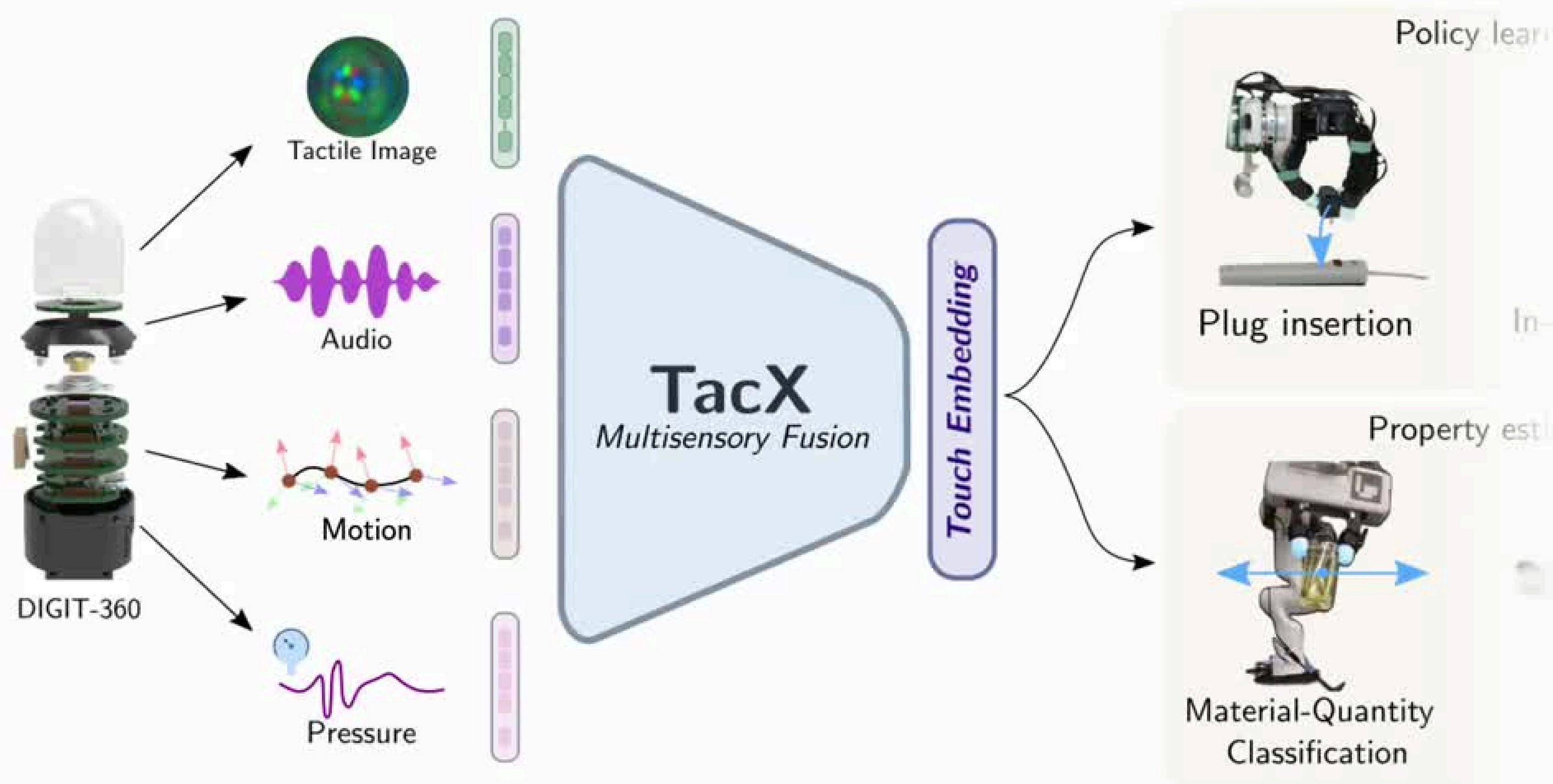
- Sadece Kamera: %20 başarı (Çok düşük).
- Kamera + Dokunsal Görüntü: Orta seviye başarı.
- Sparsh-X (Tüm Modaliteler): %90 Başarı.
- Analiz: Ses (temas 'klik' sesi) ve Basınç (hızalama kuvveti) kritik rol oynar.

Politika Öğrenimi 2 - El İçi Döndürme (In-Hand Rotation)

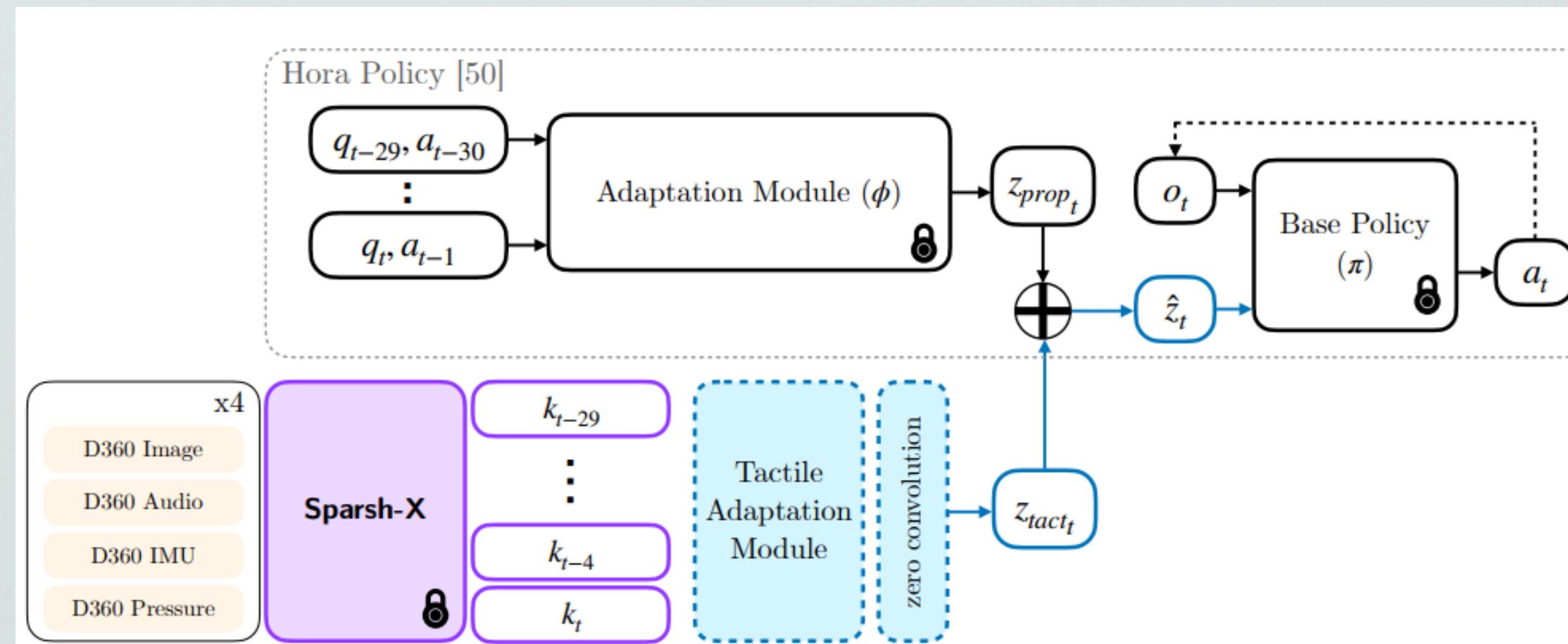


- Görev: Nesneyi düşürmeden Z-ekseninde döndürmek.
- Temel Politika (Base Policy): Hora (Simülasyonda eğitilmiş, sadece eklem açılarını kullanır).
- Problem: Sim-to-Real Gap. Gerçek dünyadaki sürtünme ve ağırlık simülasyonla uyuşmaz.
- Amaç: Temel politikayı bozmadan dokunsal geri bildirim eklemek.

Tactile Beyond Pixels: Multisensory Touch Representations for Robot Manipulation

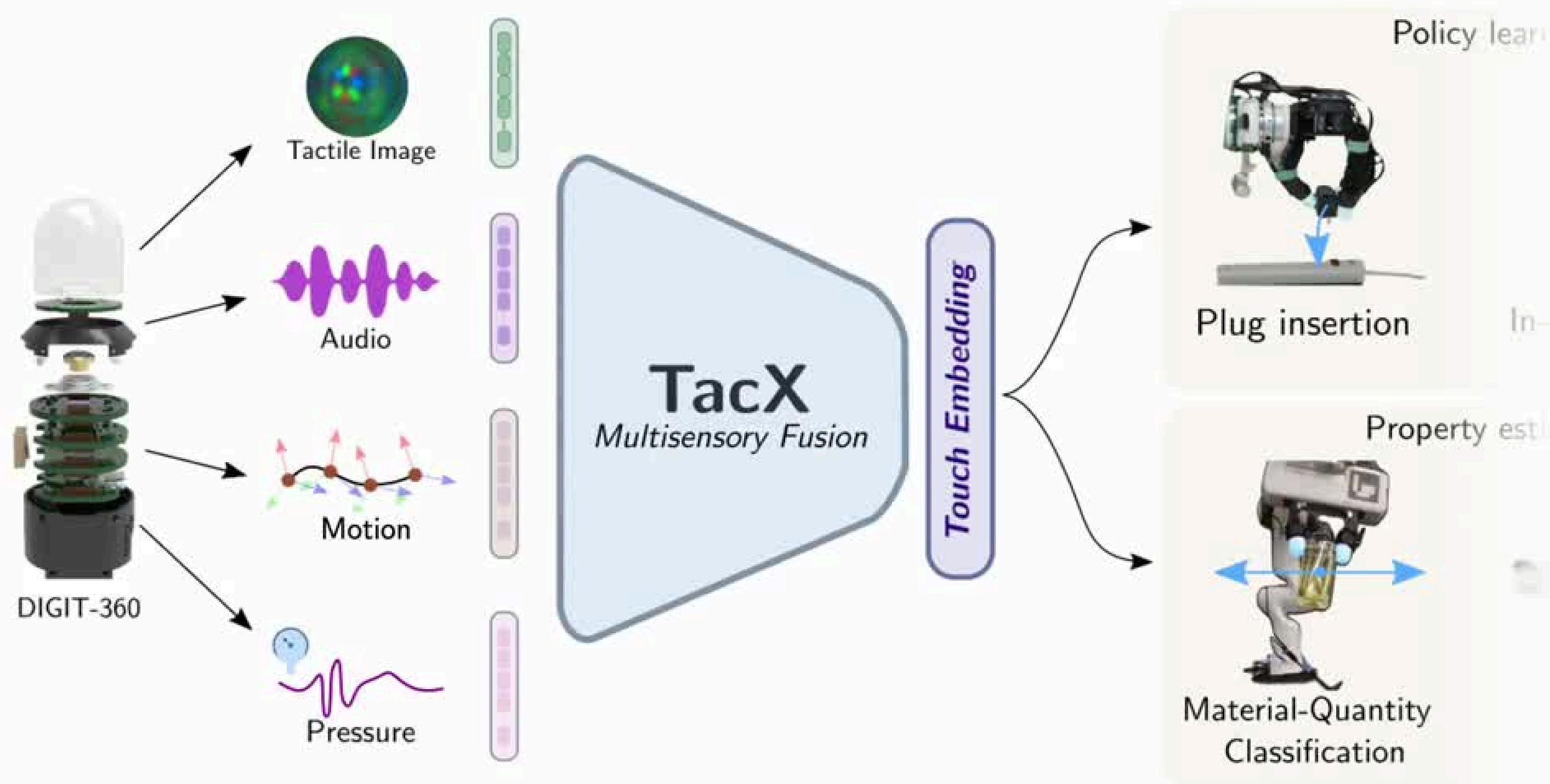


Adaptasyon Yöntemi: ControlNet ve Zero-Convolution

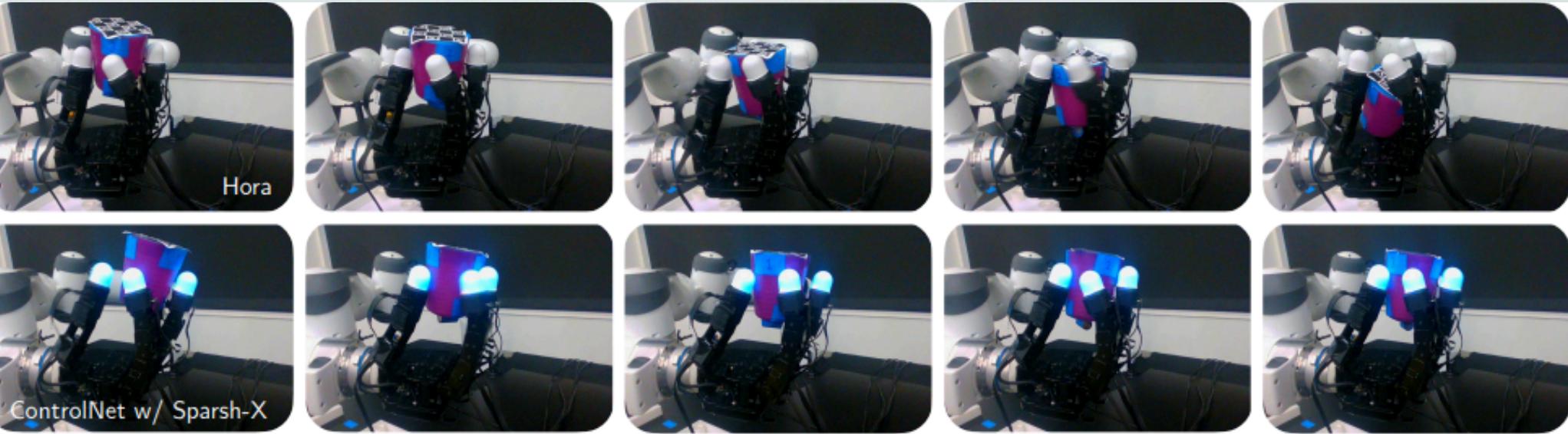


- Teknik: ControlNet mimarisi.
- İşleyiş:
- Temel politika (Hora) dondurulur (Locked).
- Sparsh-X temsilleri, eğitilebilir bir yan ağa verilir.
- Zero-Convolution: Başlangıçta 0 çıktısı verir, eğitim ilerledikçe dokunsal bilgiyi yavaş yavaş sisteme enjekte eder.

Tactile Beyond Pixels: Multisensory Touch Representations for Robot Manipulation

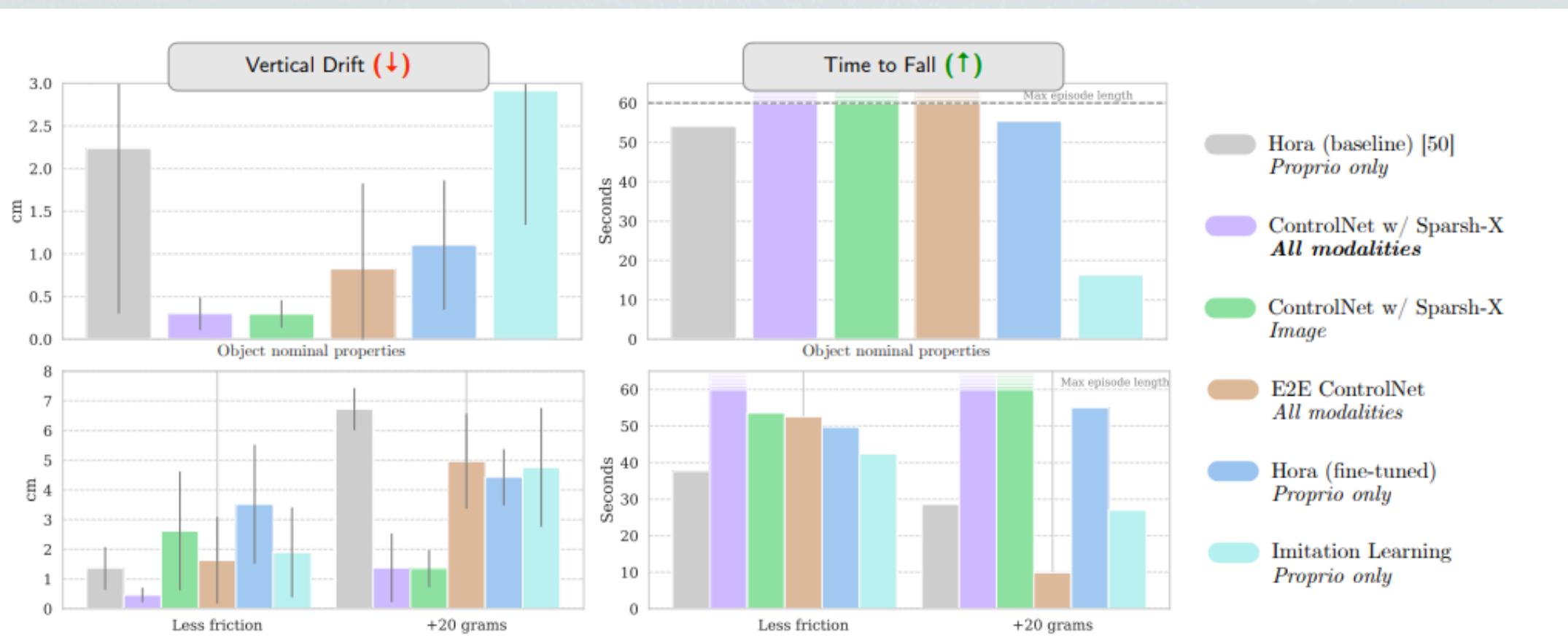


El İçi Döndürme Sonuçları



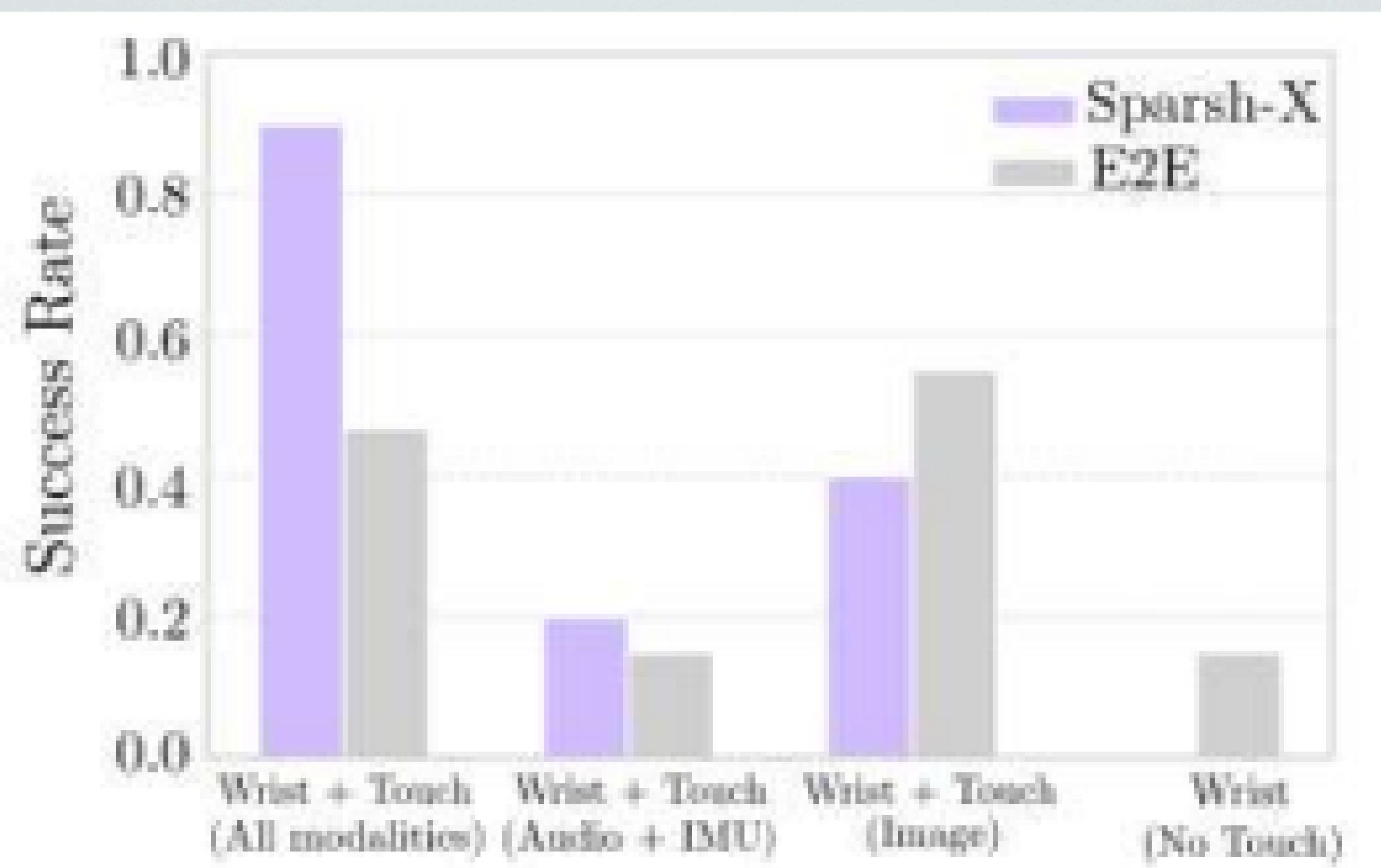
- Metrik: Dikey sürüklene (Vertical Drift) ve Düşme Süresi.
- Sonuç: Sparsh-X adaptasyonu, dikey kaymayı %90 azaltmıştır.
- Sağlamlık: Nesne ağırlaştırıldığında (+30g) veya sürtünme azaltıldığında bile stabilité korunmuştur

Genel Değerlendirme ve Sınırlamalar



- Genel Başarı: Sadece dokunsal görüntü kullanan (E2E) modellere kıyasla ortalama %48-%63 performans artışı
- Sınırlamalar:
 - Optik Artefaktlar: Digit-360'in iç yansımaları genellemeye zorlaştırabilir.
 - Hesaplama Maliyeti: 4 parmak için ses spektrogramı oluşturmak gerçek zamanlı (50Hz) çalışmayı zorlayabilir.
 - Kesme Kuvveti: Yatay kuvvetler (Shear force) henüz modellenmemiştir.

Sonuç



- Sparsh-X, robotik için ilk çok modlu (Görüntü+Ses+IMU+Basınç) dokunma omurgasıdır.
- SSL sayesinde etiketsiz büyük veriden öğrenir.
- ControlNet ile var olan politikalara dokunma duyusu kazandırılabilir.
- Gelecek: Daha çeşitli sensörler ve kesme kuvveti tahmini.
- Dinlediğiniz için teşekkürler.